

PREDICTING THE PHYSICOCHEMICAL PROPERTIES OF
AMORPHOUS POLYMER MIXTURES WITH ATOMISTIC
MOLECULAR SIMULATION AND DATA-DRIVEN MODELING

By Ziqi GAO,

*A Thesis Submitted to the School of Graduate Studies in the Partial Fulfillment
of the Requirements for the Degree Master of Applied Science in Chemical
Engineering*

McMaster University © Copyright by Ziqi GAO November 16, 2023

McMaster University

Master of Applied Science in Chemical Engineering (2023)

Hamilton, Ontario (Department of Chemical Engineering)

TITLE: Predicting the Physicochemical Properties of Amorphous Polymer Mixtures
with Atomistic Molecular Simulation and Data-driven Modeling

AUTHOR: Ziqi GAO (McMaster University)

SUPERVISOR: Dr. Li Xi

NUMBER OF PAGES: xii, 129

Abstract

Molecular dynamics (MD) simulations play a pivotal role in understanding the behavior of complex molecular systems, offering insights into the behavior of molecules at the atomic level, while their accuracy heavily depends on the force field parameters used. In this study, we present an investigation focusing on two distinct aspects: the validation of MD simulations for plasticizers, and the development of a quantitative structure property relationship (QSPR) model to fit data derived from these simulations. Our goal is to provide researchers with valuable insights into the choice of force fields to improve the accuracy of simulations in various scientific domains and the modeling of prediction of properties of plasticizers. In the first part, We explore various aspects of validation, including force field accuracy, equilibration protocols, and comparison of simulation results of plasticizers with experimental data. We begin by validating popular force fields: PCFF, SciPCFF and COMPASS. By examining the behavior of small molecules, we aim to ensure the reliability of force fields for these compounds with specific desired functional groups. Density, heat of vaporization and shear viscosity results are used for the validation of force fields. We compare various equilibration methods and their impact on simulation outcomes to address issues related to system stability and convergence, for enhancing the efficiency and accuracy of simulations. The second part of our research shifts focus to the prediction modeling of plasticizers, a class of chemical additives commonly used in the polymer industry to enhance the flexibility of plastic materials. We attempt to predict the solubility parameters of plasticizers by QSPR. Simple counts, Wiener Indices and Randic Branching Indices are used as descriptors in the QSPR. Our prediction model results show the dependence of plasticizers on the descriptors while the QSPR equation obtained from our current data-set with five descriptors has the $R^2 = 0.73$. In conclusion, this comprehensive study bridges the gap between force field validation and equilibration for plasticizers. Moreover, the integration of QSPR models offers insights to a robust approach for predicting molecular behaviors.

Acknowledgements

I would like to express my sincere gratitude to the following individuals and organizations for their invaluable support and contributions to the completion of this Master's thesis.

First of all, I owe an immense debt of gratitude to my advisor, Dr. Li Xi, for his support, guidance and patience throughout this research journey of both my undergraduate and graduate studies career. I am thankful for the opportunity to learn from him, whose mentorship has been instrumental in my intellectual development.

I appreciate the discussions, collaborations, and intellectual exchanges with my colleagues in the Xi Research Group. Their willingness to share their experiences and insights was crucial to the success of this thesis. I would also like to acknowledge Digital Research Alliance of Canada (CCDB, previously known as Compute Canada) for the facilities and computing recourse used to run my computational tasks.

I want to express my deepest thanks to my family for their unwavering support throughout this challenging journey. Their constant understanding and encouragement kept me motivated and focused. During my toughest moments of this academic journey, they came to Canada and their care were the pillars of my success, providing me with the stability and encouragement I needed.

Finally, I want to express my gratitude to anyone else who played a role, no matter how small, in this journey toward completing my Master's thesis. Your collective support has made this achievement possible, and I am deeply thankful for your presence in my academic journey.

Contents

Abstract	iii
Acknowledgements	iv
Declaration of Authorship	xii
1 Overview: the scope of study	1
I Validation and Exploration of Molecular Dynamics Simulation Procedure	7
2 Introduction to the molecular simulation part	8
2.1 Force Fields for molecular dynamics simulation	8
2.2 Motivation of comparing three force fields	13
2.3 Motivation of comparing equilibration protocols	15
3 Methodology: Simulation	18
3.1 Full-atom simulation	18
3.2 Procedure for results calculation	21
3.2.1 Procedure for shear viscosity calculation	21
3.2.2 Procedure for solubility parameters calculation	24
3.2.3 Procedure for heat of vaporization calculation	26

4	Results and Discussion	28
4.1	Force Field Comparison for Small Molecules	28
4.1.1	Force Field Validation for Small Molecules	28
4.1.2	Force Field Parameters Modification for Methyl Acetate	37
4.2	Equilibration protocol comparison	39
4.3	Force Field validation for plasticizers	45
4.4	Discussion of general observations	49
II	Exploration of QSPR Modeling for Predictions of Properties	52
5	Introduction: Review of QSPR	53
5.1	Introduction of QSPR and Descriptors	53
5.2	Motivation of investigating QSPR	55
5.3	Motivation of investigating descriptors	56
6	Methodology: Modeling Preparation	59
6.1	QSPR Data Selection	59
6.2	QSPR Molecular Modeling	64
7	Results and Discussion: Modeling	65
7.1	QSPR Results	65
7.1.1	QSPR Results for commercial plasticizers	65
7.1.2	QSPR Results for imaginary plasticizers with Wiener Index and Radic Branching Index	72
7.1.3	QSPR Results conclusion and discussion	76
7.2	Future plans	77
7.2.1	Extension of data base	77
7.2.2	New descriptors and encoder system	78
7.2.3	Numerical analysis method for prediction	80

A Appendix A: Chemical structures of plasticizers	81
B Appendix B: Raw Data and Raw Plots	89
C Appendix C: LAMMPS (31 Mar 2017) Input Scripts for Molecular Dynamics Simulation	100
D Appendix D: Chemical structures of imaginary plasticizers	118
Bibliography	123

List of Figures

1.1	Structure of DEHP and phthalates	2
3.1	Temperature profile during the amorphous simulation cell generation process for the plasticizers	20
3.2	Example results of shear viscosity calculation using the Green-Kubo relation	22
3.3	Energy calculation at separation	25
4.1	Comparison between the simulation results of density and heat of vaporization and experimental results of small molecules	29
4.2	Shear viscosity calculation details	32
4.3	Comparison between the simulation results of shear viscosity and experimental results of small molecules	33
4.4	RDF of acetone	35
4.5	Density of acetone and methyl acetate vs. temperature	36
4.6	Density and heat of vaporization comparison of methyl acetate with modified PCFF and COMPASS parameters	38
4.7	Radial distribution function of C-O in Methyl acetate	39
4.8	Alternatives in equilibration step	40
4.9	Temperature and density profiles for 3 plasticizers	42
4.10	Density and solubility parameters of DEHS with three equilibration routine	43
4.11	Solubility parameters of plasticizers along cycles	44

4.12	Comparison between the simulation results of density and solubility parameters and reference results of plasticizers	45
4.13	Radial distribution function of DEHP	48
4.14	Radial distribution function of C(=O)-O(=C) DEHP	49
5.1	QSPR process	54
6.1	Structure of DEHP and DOTP	61
6.2	Chemical structure of an example of "imaginary plasticizers": a5-b1	62
7.1	Solubility parameters of plasticizers vs. NCL	66
7.2	Solubility parameters of plasticizers vs. NCT and NCA	67
7.3	Plot of δ (predicted from the QSPR model) vs. δ (calculated from the MD simulation)	69
7.4	Plot of δ (predicted from the QSPR model) vs. δ (calculated from the MD simulation)	71
7.5	Plot of δ (predicted from the QSPR model) vs. δ (calculated from the MD simulation)	74
7.6	Solubility parameters of plasticizers vs. Wiener Index and Randic Branching Index	75
A2.1	Shear viscosity calculation details of ethanol with SciPCFF in Appendix .	90
A2.2	Shear viscosity calculation details of methyl acetate with COMPASS in Appendix	91
A2.3	Shear viscosity calculation details of methyl acetate with PCFF in Appendix	92
A2.4	Shear viscosity calculation details of methyl acetate with SciPCFF in Appendix	93
A2.5	Shear viscosity calculation details of acetone with COMPASS in Appendix	94
A2.6	Shear viscosity calculation details of acetone with PCFF in Appendix . .	95
A2.7	Shear viscosity calculation details of acetone with SciPCFF in Appendix .	96

A2.8 Shear viscosity calculation details of diethyl ether with COMPASS in Appendix	97
A2.9 Shear viscosity calculation details of diethyl ether with PCFF in Appendix	98
A2.10 Shear viscosity calculation details of diethyl ether with SciPCFF in Ap- pendix	99

List of Tables

2.1	Pair coefficients of PCFF, SciPCFF and COMPASS	11
4.1	Density and heat of vaporization comparison of small molecules with PCFF, SciPCFF and COMPASS	30
4.2	Shear viscosity comparison of small molecules with PCFF, SciPCFF and COMPASS	33
4.3	Density and solubility parameter comparison of plasticizers with PCFF, SciPCFF and COMPASS	46
6.1	Descriptors and solubility parameters of plasticizers	60
6.2	Descriptors and solubility parameters of imaginary plasticizers	63
A1.1	Chemical structures of common plasticizers	81
A4.1	Chemical structures of imaginary plasticizers	118

Declaration of Authorship

I, Ziqi GAO, declare that this thesis titled, “Predicting the Physicochemical Properties of Amorphous Polymer Mixtures with Atomistic Molecular Simulation and Data-driven Modeling” and the work presented in it are my own. I confirm that:

- List each chapter
- and what you have done for it

Chapter 1

Overview: the scope of study

Plasticizers are usually incorporated into amorphous polymers to alter the interactions within the polymer system so that these additives could adjust properties of materials[1]. By the definition given by IUPAC (International Union of Pure and Applied Chemistry), plasticizers are described as: substances incorporated in a material (usually a plastic or elastomer) to increase its flexibility, workability, or distensibility. Usually the addition of plasticizers would essentially soften the system and reduce the stiffness. Plasticizers are widely used as additions to commercial polymers such as Poly(vinyl chloride) (PVC). As a popular used material, pure PVC is stiff and brittle. Since plasticizers would make polymers flexible and durable, the addition of these additives to PVC would be welcomed. According to existing data, approximately 500 different plasticizers are commercially available in polymer industry nowadays and 80% of plasticizer consumption directs to PVC[2][3]. Since plasticizers play an important role in the polymer industry, the studies about them are of interest to polymer researchers.

One of the most used plasticizers is DEHP (Bis(2-ethylhexyl) phthalate), which is one of phthalates[3]. The structure of DEHP is shown in Figure 1.1. DEHP can be separated into two parts, the torso part (with an aromatic ring) and the leg part (alkane chains) and these two parts are connected by carboxyl group. Nearly all other available

plasticizers have similar chemical structures. There are other types of plasticizers. Except for phthalates, citrates, adipates, epoxides, trimellitates and phosphates are also available for commercial application[4]. Phthalate based plasticizers are very commonly used in the plasticizer market, in which they occupy more than 80% of the industrial consumption[5].

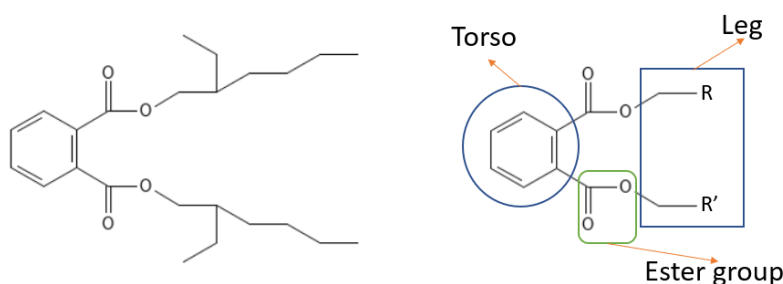


FIGURE 1.1: Chemical structure of DEHP (left) and typical structure of phthalates with torso and legs labeled (right)

However, phthalates could migrate from PVC to the surrounding environments, which would cause potential hazard to human health[2]. This issue becomes more serious and the usage of phthalates and other kinds of plasticizers are limited in the food and car industries for the environment and human health protection purposes[6]. The study of new green plasticizers is encouraged to solve the issue. The development of polymer design might solve the problem[7]. New environmental friendly plasticizers could be predicted and produced. Then these new plasticizers may replace the phthalates currently being used. However, there are unanswered questions of plasticizer design. More rationales of interactions between additive plasticizers and polymers are studied by many researchers. Based on their research in the area of effects of plasticizers in plasticized polymer systems, more practical methods for plasticizer design could be established.

Recently Li et al. presented a comprehensive study on the performance of plasticized PVC[8]. Multiple kinds of plasticizers were tested, including phthalates, trimellitates,

citrates and etc.. Seven molecular design parameters (MDPs) were tested and the effects of these MDPs on plasticizer performances were discussed. The compatibility between the plasticizer and PVC, plasticization efficiency, and plasticizer mobility were the criteria in the comparison. The MDPs include leg length, number of legs, position of legs and etc.. It is clear that the variation of the MDPs would cause changes in the performances of plasticizers. The molecular mechanisms behind the observed relationships between molecular structure and plasticizer performance are then discussed in that study. Based on his research, the importance of effects of structures to performances is shown for the purpose of providing a guideline to polymer design. In this kinds of research, computer simulation is a commonly used tool.

Using computer simulation, the analysis in the atom level could be done quickly. Molecular Dynamics (MD) is a type of simulation which solves for the motion of particles by solving the Newton's equation of motion[9]. Thermodynamics will be captured by this process. The most important component of MD is to accurately calculate the potential of the particle systems. This is achieved by force field and equilibration protocols. Force field would describe the interactions among atoms, which could evaluate the bond (e.g., covalent bonds) and non-bonded (e.g., van der Waals forces and electrostatic interactions) energy of the systems.

MD simulation is widely used for the prediction of properties by researchers in the area of pure polymers and polymer blends[10][11]. A disadvantage of using computer simulation is that it is even slower than experiments. While other researchers are exploring the rationale behind the phenomena of effects of plasticizers with molecular simulation, I would like to explore a faster tool for prediction based on their understanding of effects of legs, torsos and connecting of each parts in plasticizers. The long time waiting of atomistic molecular simulation could be saved.

The topic I am interested in for my project is the prediction of properties of amorphous polymer mixtures with data-driven modeling. This will be achieved from following aspects. The final target is the prediction of the properties through a mathematical model without running simulations or experiments. The model starts from describing the structure of polymers by some of the chemical-structure-based descriptors. Those descriptors will work as the input to the model and the output of the model will be the properties that we want to predict. To train this model, a valid training data-set would be required. The training data-set would be obtained from MD simulation. In the practice of using MD simulation to generate data, I found that the accuracy of MD needs to be improved. The improvement of its accuracy could be done through (1) force fields and (2) equilibration protocols. The solubility parameters are used as the criteria for testing the accuracy of MD simulation since understanding the solubility parameter can help predict the compatibility of polymers with other materials, such as additives.

The Hildebrand solubility parameter is a way of quantify solubility parameters, which is derived from cohesive energy density and this quantity also relates with enthalpy of vaporization. The solubility parameter is a quantitatively value to reflect the relative solvency behavior between materials including polymers[12]. For polymer-solvent miscibility, when the solubility parameters of two substances are similar, the dissolution would more likely happen. This property is relatively easy to be calculated from simulation results. Meanwhile, solubility parameters could also represent the miscibility of particles, which might be useful to compare with one of the important performance of plasticizers, compatibility.

The data-driven model would be developed after adequate data are obtained. With the development of computational resources, complex regression models are available. A multi-linear regression method called QSPR (quantitative structure property relationship) will be introduced in this project and be applied to develop a model from plasticizer

data[13].

In general, our end goal is to extract a procedure of computer simulations for polymer design of plasticizers or other molecules those we are interested in. We would like to find good plasticizers and replace the current using ones with them. This requires us to test and confirm the plasticizers from our prediction model. We choose QSPR as a tool to for prediction.

To train this QSPR model, a data-set is required. In this data-set, plasticizers with their target properties are listed. This data-set also contains the values of plasticizers' descriptors. These descriptors do not require further simulation or experiments to be generated. They are usually calculated from structure information of molecules.

To get the information of properties in the data-set we needed, we choose to use molecular dynamics simulation. With the help of high speed and quality computational resources, numbers of required information and be generated. For the purpose of making chemically specific prediction and comparison, all-atom simulation was applied.

To make the prediction model more accurate, two more aspects are considered. The force field used in the simulation and the equilibration protocol. The force field should accurately capture the interactions between plasticizer particles. The equilibration protocol should make the output of MD simulation polymer system reliable. The comparison of force fields and equilibration protocols will be based on the heat of vaporization for small molecules and solubility parameters for plasticizers. The effects of parameters of force fields would be tested to extend our study of plasticizers. We want to find the best way to estimate the behavior of molecules including plasticizers using QSPR model base on the training data from MD simulation, which requires us to explore the correlations among these topics.

Generally, the goal of this exploration road contains making accurate MD simulation

for our data preparation of QSPR modeling. This thesis only addresses a few challenges in the road map. We explored improving the accuracy of MD simulation and descriptor validation and comparison for QSPR modeling. We did not solve big problems of making accurate MD and QSPR. Only some specific contributions were made in these areas. In pursuit of our significantly larger objectives, there are a lot of challenges and opportunities in the future. Our achievements are not the end of the journey but rather a step along the way.

Part I

Validation and Exploration of Molecular Dynamics Simulation Procedure

Chapter 2

Introduction to the molecular simulation part

2.1 Force Fields for molecular dynamics simulation

Polymer is one of the most important materials to current industry. With the development of computer power, simulation techniques are applied to polymer fields. Molecular dynamics (MD) simulation is a primary technique widely used in computational studies in material science. It can provide useful predictions of the system of polymers without time-consuming experiments. Using full-atom molecular simulations, certain properties such as densities or viscosities would be predicted. Force fields (FF) based simulations are used in this procedure instead of ab initio methods to save the expensive time cost for simulation[14]. For high accuracy of simulation, simplified models with proper force fields should be validated. Comparison between experimental and simulation results would show the behavior of the simulation model. This section is a brief introduction of how important of force fields are to molecular dynamics simulation and what are these force fields.

MD simulations use the classical equations of motion and integrated these equations to molecular level systems with certain amount of particles which could represent atoms, molecules or coarse-grained beads. The simulation starts by establishing an initial arrangement of positions and velocities for every atom within the system. Then the simulation proceeds by integrating the equations of motion over small time steps. During each increment, the positions and velocities of the atoms are adjusted according to the forces acting on them. With the development of computer science, the first exploration of Monte Carlo and Molecular Dynamics simulations of models was performed by Metropolis and Alder[15][16]. This method was then widely used in many scientific areas such as chemical industry. There is a lot of good software to perform MD simulations. Largescale Atomic/Molecular Massively Parallel Simulator (LAMMPS) is one of them[17]. With the help of computers and MD, complex systems containing many particles could be simulated and users could get their acquired information with only minimum and easy training. Meanwhile, to make sure the results simulated are accurate and reliable, good models which could capture the behaviors of particles properly are required. The accuracy and reliability of molecular dynamics simulation depends on the methods that are modeling the intramolecular and intermolecular interactions. The intramolecular interactions include bond stretching, angle bending, and dihedral and improper torsions. The procedure of simulating the motion of each particle should capture the interactions. The expression to compute these interactions in molecular dynamic simulation is called force field.

A force field is a mathematical expression describing the dependence of the energy of a system on the coordinates of its particles[18]. It includes a set of parameters for analytically modeling the potential energy, which could simplify the true potential during simulation. It makes molecular dynamics simulation focus at the atomic level. The settings of parameters containing interatomic information. Usually the parameters are derived from experimental and quantum-mechanics simulation data by proper fitting

methods. There are a number of different kinds of force fields available nowadays.

The first force fields were used to study small organic molecules in the 1960's[19]. Then the force fields were extended to study hydrocarbons by Allinger et al. with the development of molecular mechanics[20]. With the study of much more complex systems, more research and developments were made in force fields. The first-generation force fields or the so-called Class I force fields uses an expression with intramolecular contributions to the total energy, Van der Waals interactions and the Coulombic interactions. While Class I force fields are using simple function expressions, second-generation force fields or Class II force fields contain some other additional cross-coupling terms to describe contributions to the total energy. These terms are describing polarization effects and coupling between bonds. Some popular Class II force fields are CFF (consistent force field), UFF, MMFF and COMPASS[14][21][22]. The more flexibility of the expressions of Class II force fields would be parameterized more accurately.

A force field named polymer consistent force field (PCFF)[23], which was developed from CFF91 and CFF93[24], is widely used for organic materials. The force field parameters are adjusted to match quantum mechanical data, which contains total energies, first and second derivatives of these energies, and electrostatic potentials. The reason why PCFF was commonly used was because its development in the parameterization of functional groups of most common organic materials by Hagler and co-workers[25].

SciPCFF (Scienomics Polymer Consistent Force Field) was extended from the PCFF by Scienomics Inc. and the parameters of SciPCFF are available in their company's software, MAPS[26][27]. The major difference between PCFF and SciPCFF is in the update of several non bonded and bonded parameters.

Another interesting Class II force field is COMPASS (condensed-phase optimized molecular potentials for atomistic simulation studies). Sun published the works of the

TABLE 2.1: Pair coefficients of PCFF, SciPCFF and COMPASS

Atom Type	Parameter values					
	PCFF		SciPCFF		COMPASS	
	ϵ	σ	ϵ	σ	ϵ	σ
C(=O)	0.12	3.81	0.12	3.81	0.064	3.9
C(aromatic)	0.064	4.01	0.071	3.922	0.068	3.915
O(=C)	0.267	3.3	0.27	3.3	0.192	3.43
O(-C)	0.24	3.42	0.24	3.3	0.096	3.3
C(sp3)	0.054	4.01	0.062	3.854	0.062	3.854
H	0.02	2.995	0.023	2.878	0.023	2.878

*parameters were from [21][27][26]

COMPASS force field starting from 1997[21]. COMPASS was also generated based on the PCFF force field with a hybrid approach including ab initio and empirical methods. Parameterization and validation were introduced in those works. The most important differences between the COMPASS and PCFF parameters are in the nonbond vdW LJ-9-6 parameters, ϵ (the well depths) and σ (size of the particle). Usually the values of ϵ are larger and σ are smaller in COMPASS. COMPASS force field is parametrized relatively accurately in predicting both intramolecular properties and intermolecular properties for molecules. Table 2.1 shows the details of the pairwise force field coefficients for each atom type. Note that in the atom type column, "C(=O)" indicates that the carbon is the carboxyl carbon connected to an oxygen atom through a double bond.

For Class II force fields like COMPASS, the nonbond interactions are represented by the 9-6 Lennard-Jones potential function:

$$E_{vdW} = \begin{cases} \epsilon[2(\frac{\sigma}{r})^9 - 3(\frac{\sigma}{r})^6] & r < r_c \\ 0 & r \geq r_c \end{cases} \quad (2.1)$$

where E_{vdW} is the vdW nonbond interactions, r is the distance between atoms and r_c is the cut-off distance. The ϵ is the amount of energy at the minimum energy stage when

$r=\sigma$ [28]. This equation is for the same atom types. In the case of pairs of atoms with different atom types, a 6th order combination law is applied to capture the parameters ϵ_{ij} and σ_{ij} [29]. The Lennard-Jones potential offers a simplified representation that captures the fundamental characteristics of repulsive and attractive interactions between atoms and molecules. Two atoms engaged in interaction exhibit a repulsive force when in close distance, an attractive force at a moderate separation, and remain non-interacting at an infinite separation.

To enhance the prediction of properties of polymers in material science, the parameterization and validation of the common functional groups of common organic and inorganic polymers is studied and presented. PCFF is a widely used force field for organic polymer simulation and SciPCFF and COMPASS are developed from PCFF. SciPCFF is new and less tested, and COMPASS is good but only limitedly available. These three force fields are of our interested for comparison in this project.

Meanwhile, since the plasticizers are of most interest, some small common organic molecules are studied as well. They are acetates, acetone, ether, alcohol and benzene. These small molecules contain the functional groups which also in the conformation of plasticizers. Acetates and acetone have ester groups or carbonyl group (which could be considered as a part of ester group) in their formulas, where ester groups connect the torso and leg part of plasticizers. Ether and alcohol are included for the purpose of exploring the effects of single oxygen atoms. Benzene and toluene contain phenol group, which is the torso part for phthalates. The study of these small common molecules would be a benchmark to the analysis of force fields since plasticizers are complex molecules which are usually formed by aromatic rings, ester groups and alkane chains. Difference between force fields in plasticizers is hard to pinpoint because it could come from errors of multiple functional groups. The simple molecules allow us to study the accuracy of those functional groups separately. The descriptors we want to use should describe the

formational information of the torso and leg parts. Therefore, the comparison of force fields for small molecules with aromatic rings, ester groups and alkane chains would be done before the comparison for plasticizers.

Density and heat of vaporization are selected as the criteria for validating force fields for small molecules. However, while density and heat of vaporization are used in the parameterization of COMPASS, viscosity is a dynamic quantity that is not used in the parameterization of any force fields. Therefore, calculating shear viscosity is essential for this project to provide a much more stringent test.

2.2 Motivation of comparing three force fields

Since there are a large number of force fields, it is difficult to decide whether a force field is suitable for a certain particular system. This leads to the area of comparing the performance of these force fields. However, there is none prior work studied the suitability of force fields that could accurately capture the dynamics of plasticizer systems.

Chen et al. studied on validating the force fields COMPASS and PCFF in predicting the physical and thermophysical properties of polyaniline[30]. Both density and solubility parameters are accurately generated by the MD and COMPASS yield more accurate predictions.

Recently, Nikzad et al. have optimized the force fields of SciPCFF and PCFF for liquid crystalline elastomers (LCEs)[31]. In their work, some properties such as steady-state density and transition temperature were simulated by these force fields and the accuracy of the results was compared to achieve a proper selection of force fields. The results showed that the SciPCFF is the most appropriate force field to study the LCEs with better agreement to the experimental data.

In our project, due to our goal which is to build prediction models for certain properties such as density and solubility parameters of pure polymers and mixtures with additives, the comparison among some different force fields is focused on these specific properties of selected molecules.

As mentioned in the previous section, PCFF is commonly used in simulation of polymers. However, it was proved that some parameters from CFF91 do not meet the application of molecular dynamics simulations at finite temperature. Since many of its non-bonded interaction parameters were taken from CFF91 directly, PCFF would have the some issue as CFF91 and the results of density from simulation would be too low. Therefore, an improvement could be made in this area.

The study of Sun’s group makes a contribution to the description of atomistic simulation, but generally COMPASS force field is not available due to the proprietary. In our project, the initial conditions of polymers are prepared in MAPS, a modeling platform for model building of any types of materials. PCFF and SciPCFF are build-in force fields but COMPASS force field is not available in MAPS.

Both PCFF and SciPCFF are reachable from MAPS. However, COMPASS is not available to all users due to the proprietary while COMPASS is more accurate in certain simulation results. This issue leads to our goal, to make use of publicly available force fields and develop alternatives to predict polymer systems with limited availability of force fields. For people with only PCFF or other kinds of Class II force fields, an alternative to improve PCFF and/or SciPCFF so that their accuracy can be closer to COMPASS might work. The idea is to find a method by modifying the PCFF with some known and public information of COMPASS to enhance the results from this new modified PCFF. This is achieved by adjusting the force field parameters ϵ and σ to make the simulation results converge to real substance properties. While limited information of COMPASS force fields are available from published papers, some modification based

on PCFF and COMPASS might be applied to PCFF to enhance the simulation results in certain areas.

Our end goal is to help researchers derive key performances of polymers such as dynamical or mechanical properties from the chemical structure of the polymers. With the use of molecular simulation, the polymer systems could be predicted and Visualized. With the use of modified force fields, more accurate models of polymer systems could enhance the understanding of polymer properties. This study might serve as a guidance in the development of force fields parameterization and validation for other researchers.

2.3 Motivation of comparing equilibration protocols

In addition to accurate force field, the equilibration protocol – i.e., the procedure to prepare a molecular model that represents the real material – is also critical to the accurate calculation of materials property data. After initial conditions of particles were generated as input for equilibration, this simulation protocol needs to provide a valid procedure for preparing the initial conditions for production runs which will generate the property data. In molecular dynamics simulations, in order to guarantee that the system reaches a stable and representative state prior to any subsequent analysis or production runs, the equilibration protocol is crucial. For amorphous polymer simulation, the proper equilibration protocol should help the relaxation of system configuration, which leads to a steady-state with removed artificial imbalances or inconsistencies from the initial configurations. Speeding up the simulations while maintain the simulation time under proper relaxation time is desired by polymer researchers[32][33].

Our initial attempt of setting up the simulation protocol was modified based on the equilibration method for plasticized PVC generated from Li et al.[8]. While exploring the effects of molecule design parameters to plasticizers, the study used molecular dynamics simulation to investigate the behaviors of polymer mixtures, where long PVC chains

are packed with plasticizers by PACKMOL[34]. For our pure plasticizer systems which have smaller molecules, shorter equilibration may potentially be desired. Although Li et al.'s protocol was good, the expectation that plasticizers may take shorter runs gives us room to further reduce the computational cost, which is important for the reason that we need to run many MD simulations since we are using MD as our data generator. With the objective of trying to find out the minimal amount of computation for equilibrating such systems for reliable data generation, we would like to study the differences between different equilibration protocols for plasticizers to explore the potential improvement in this project based on the comparison and understanding to save computational cost of simulation.

Density and solubility parameters are chosen as the criteria for the validation. For establishing the simulation protocol, 3 plasticizers, TOTM, DITP and DEHS (chemical structures could be found in Appendix A) were chosen to test the equilibration protocol. These 3 plasticizers were selected based on the idea of representing typical chemical structures of phthalate or aliphatic dicarboxylate plasticizers while their molecular sizes are relatively large. Protocols suitable for these larger molecules should also work for molecules with similar structures and smaller size. Comparison among the validation results of all three plasticizers would allow discussion about effects of different equilibration protocol.

In this project, the performances of three different protocols (a prototype protocol which were introduced in section 3.1 and two modified protocols which were introduced in section 4.2) were compared by the observations of simulation results. Originally we have repeated heating-cooling cycles as a common approach for molecular cell equilibration. After the conformations of molecules were generated, an energy minimization step and a relaxation step would follow the initial configuration. For the density to converge, a number of heating-cooling cycles will be used to prepare the model for production runs.

The details of this protocol will be introduced in the next methodology section. While considering the results of the distribution of free volume in the comparison of well equilibrated polymer cells with insufficiently equilibrated polymer cells, the heating-cooling cycles were replaced by 2 more modifications. These modifications of the equilibration protocol will be introduced in the section of comparison of different protocols.

Chapter 3

Methodology: Simulation

3.1 Full-atom simulation

Full-atom molecular models were used in this project. Both the small molecules and plasticizer molecules were constructed in MAPS, a modeling platform for model building of any types of materials. Amorphous cells which would represent as the ensemble of realizations of amorphous polymer structures were generated in MAPS. The initial configuration of these amorphous cells was built at a density of $0.5-0.8 \text{ g/cm}^3$. The simulation box contains 100 plasticizer molecules or 1000 small molecules. The amorphous cells further undergoes an extensive multistep equilibration protocol[8].

After the initial configurations were generated from MAPS, an energy minimization step and further equilibration steps were applied. The potential energy is calculated with three force fields for atomistic simulation studies, including the polymer consistent force field (PCFF), Scientific polymer consistent force field (SciPCFF) and

condensed-phase optimized molecular potentials (COMPASS). Molecular dynamics simulation is implemented with the Largescale Atomic/Molecular Massively Parallel Simulator (LAMMPS)[17]. The cutoff distance for pairwise van der Waals (vdW) and electrostatic interactions is established as 15 Å. The contribution of long-range vdW interaction is estimated using a tail correction, while the long-range electrostatic interaction is calculated using the conventional Ewald summation method[35]. To carry out time integration, the standard velocity Verlet algorithm is employed with a time step of 1 femtosecond (1 fs), and energy minimization is accomplished using conjugate gradient algorithms[36]. When utilized, the thermostats and barostats are implemented using Nose Hoover chains[37]. Periodic boundary conditions were applied in all directions.

The structure needs to be further equilibrated before production runs. The multistep equilibration protocol was different from plasticizers to smaller molecules. For plasticizers, Li et al.’s protocol was used as a benchmark, against which our new protocols will be compared, since these plasticizer molecules were also analyzed in that study[8]. The following multistep amorphous cell building procedure is found to be robust for plasticized PVC, in which should also be robust for pure plasticizers. By this procedure, the solubility parameter is found to be independent of the initial configuration.

Here is the details of the multistep equilibration protocol for plasticizer molecules:

Step 1. Molecular construction, force-field assignment, and plasticizer packing (with MAPS). The density of the initial cell is within 0.5 – 0.8 g/cm³, which is close to the reference experimental density of those molecules.

Step 2. Energy minimization to remove atom overlaps and energy singularities.

Step 3. Keep the density constant and the molecule configuration frozen and run extended (≈ 5 ns) NVT simulation at a set temperature to quickly relax and redistribute the molecules. The set temperature for plasticizers is 600 K. And for small molecules,

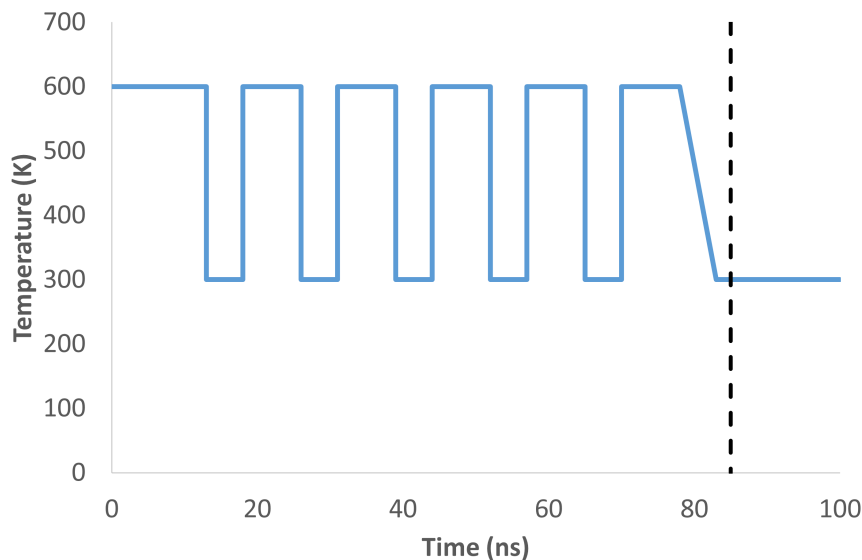


FIGURE 3.1: Temperature profile during the amorphous simulation cell generation process for the plasticizers.

the set temperature is set to be the same as the temperature where reference density is compared.

Step 4. For plasticizer molecules, the systems were equilibrated by repeated (5 – 7) heating-cooling cycles, Each cycle was with an 8 ns run at 600 K followed by a 5 ns run at 300 K, both NPT at 1 atm. For small molecules, the systems were equilibrated for 12 ns in the isothermal-isobaric (NPT) ensemble at set temperature and 1 atm.

Step 5. For plasticizer molecules, the last cooling cycle was running at 1 atm by a ramp cooling from 600 K to 300 K followed with a 2 ns run at 300K.

After the multistep equilibration protocol, the production runs for density and solubility parameters were 2 ns in the isothermal-isobaric (NPT) ensemble at set temperature and 1 atm. The density and energy terms were saved every 10 ps.

For our purpose of investigating effects of different protocols, the equalibration protocol of the heating-cooling cycles in step 4 is modified for DEHS, DITP and TOTM for

discovering improvements to the results. The details would be presented in section 4.2.

3.2 Procedure for results calculation

3.2.1 Procedure for shear viscosity calculation

There were a number of simulation methods for calculation of the shear viscosity of liquids. The Green-Kubo approach (3.1) is widely used for the calculation of shear viscosity due to its simplicity[38].

In the Green-Kubo approach, the shear viscosity is calculated based on the molecular dynamics simulation at zero shear rate. The integration of the autocorrelation function of the stress tensor is expected to converge to a constant value after a specific point. This behavior arises from the theoretical decay of the autocorrelation function to zero as time approaches infinity[36]. The Green-Kubo relation is shown below:

$$\eta = \frac{V}{k_B T} \int_0^{\infty} \langle P_{\alpha\beta}(t) \cdot P_{\alpha\beta}(0) \rangle dt \quad (3.1)$$

V is the system volume, k_B is the Boltzmann constant, T is the temperature and $P_{\alpha\beta}$ records the non-diagonal elements of the stress tensor. The angle bracket denotes the ensemble average, which indicates an average over different time origins.

The Figure 3.2 below presents 3 trials of Green-Kubo relation, which used as an example that shows how the trajectories generated by the running integral of the Green-Kubo equation look like. Three trajectories were evaluated from three independent simulation runs which were 10 ns long for ethanol at 298 K. These 3 curves seem to be converging before 6 ns and they start to deviate from one another after 6 ns. In practice, the shear viscosity should be estimated from a plateau region of the trajectory.

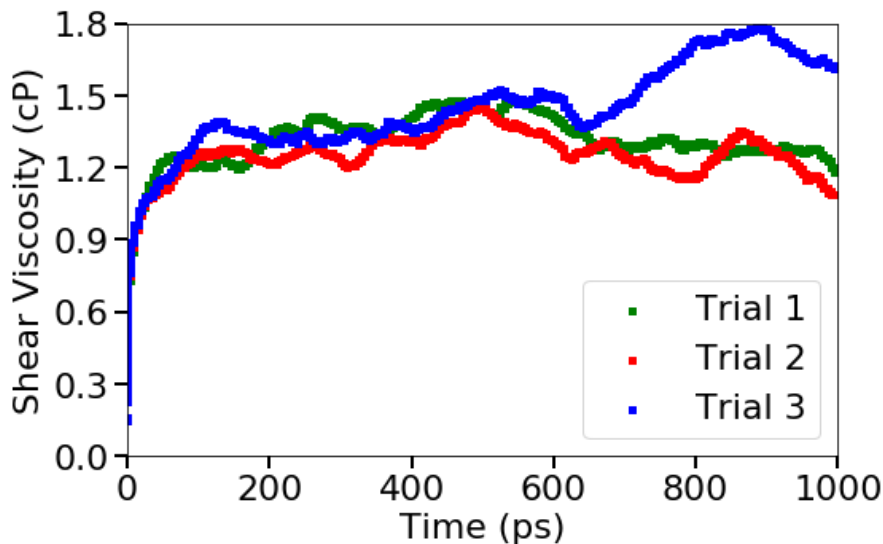


FIGURE 3.2: Example results of shear viscosity calculation using the Green-Kubo relation. Three trials were calculated from three independent simulation runs for ethanol at 298 K.

However, there are some disadvantages for the original Green-Kubo relation method. Firstly, this approach is computationally expensive. It usually requires a long time simulation trajectory. Meanwhile, the plateau region of the integral is difficult to identify in practice. In the example of Figure 3.2, even for the period between 1 ns to 6 ns, the plateau regions for the 3 trials were hard to identify. After 6 ns, due to the noises in the simulation, the integration could not converge to a constant value for each trial. In most cases, the convergence of the trials is worse than this example. This adds a large amount of variability to the calculation. Human judgment and intervention were impractical in such studies.

In this project, the calculation procedure was based on the research of Zhang and Maginn for reliable viscosity calculation, which is a more systematic way to calculate the shear viscosity[39]. This method has been proved to be reliable in the calculation of shear viscosity of liquids with differing viscosities. The shear viscosity was calculated by following procedure:

Step 1. Generate N independent trajectories from the same initial configuration at the corresponding temperature. The initial configuration was generated from the procedure introduced above. Then the independent trajectories were generated by assigning different random seeds to the initial velocity. Each trajectory would run for 1 ns in an NPT ensemble which is aimed to be equilibration. The production runs for shear viscosity were in canonical (NVT) ensemble at the set temperature for 10 ns. The stress tensor components were saved every 5 fs during the simulation.

Step 2. For each trajectory, the shear viscosity η was calculated based on the Green-Kubo equation (3.1). The stress tensor components were calculated from the 1 ns production runs in step 1.

Step 3. Averaging the running integrals over N trajectories and the standard deviation:

$$\sigma(t) = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (\eta(t)_i - \langle \eta(t) \rangle)^2} \quad (3.2)$$

Step 4. Fit the standard deviation to a power law function:

$$\sigma(t) = At^b \quad (3.3)$$

Step 5. Fit the averaged running integral by the double-decay function developed from Rey-Castro and Vega. The A is from Equation 3.3. And α , τ_1 and τ_2 are the fitting parameters.

$$\eta(t) = A\alpha\tau_1(1 - e^{-t/\tau_1}) - A(1 - \alpha)\tau_2(1 - e^{-t/\tau_2}) \quad (3.4)$$

From Rey-Castro and Vega's note[40], the weight factor was set to $1/t_2$ while Marginn's group chooses $1/t_b$ since they found that the $1/t_2$ weighting factor would make the model

decay too fast, where b is fitted from Equation 3.3. We adopted suggestions from Zhang and Marginn. The cutoff time was chosen based on the proper plateau region. The final calculated shear viscosity was derived from the infinite long-time limit of the double-decay function.

3.2.2 Procedure for solubility parameters calculation

The Hildebrand Solubility parameter δ is defined as the square root of cohesive energy density (equation 3.5).

$$\delta = \sqrt{\frac{E_{coh}}{V}} \quad (3.5)$$

E_{coh} is the cohesive energy and V is the specific volume. Cohesive energy basically means the energy required to overcome the intermolecular interactions of a system and pull the individual molecules apart from the condensed phase to infinite separation. The definition of cohesive energy is defined in equation 3.6.

$$E_{coh} = E_{sep} - E_{bulk} \quad (3.6)$$

E_{bulk} is the specific potential energy in the condensed phase. E_{bulk} is calculated from the equilibrated simulation cell by equation 3.7:

$$E_{bulk} = \frac{E^t}{m^t} \quad (3.7)$$

where E^t is the total potential energy of the simulation cell and m^t is total mass of the molecules in the cell. And E_{sep} is the the specific potential energy of molecules at infinite separation, in which would behave like an ideal gas. In practice, E_{sep} is calculated

from the averaged individual molecule chains from the equilibrated simulation cell by equation 3.8,

$$E_{sep} = \frac{\langle E_{ind}^t \rangle}{m_{ind}} \quad (3.8)$$

where E_{ind}^t is the averaged total energy in vacuum and m_{ind} is the mass of the molecule. Each molecule would be picked out separately and then moved to an empty cell while no further simulation steps would be applied. The potential energy of each isolated individual molecule was then computed from the frozen separated conformation in these cells, which were considered to be infinitely apart from the original cell where the individual molecules came from.

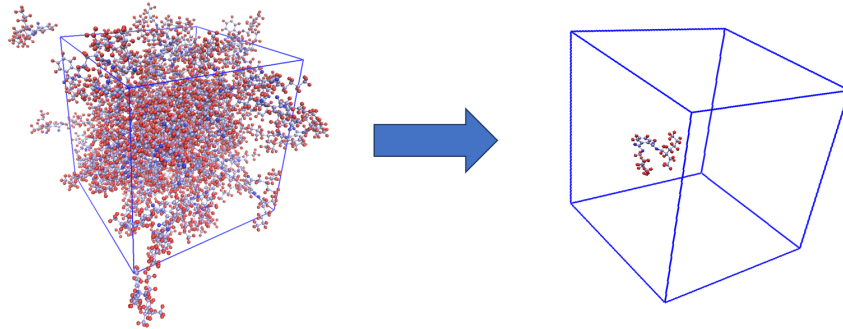


FIGURE 3.3: Simulation cells for each molecule in the procedure of calculation of E_{coh} . Left simulation cell shows 100 DEHP molecules. Right simulation cell shows 1 DEHP molecule while the other 99 molecules are deleted. No simulation steps are applied in this deletion. E_{bulk} is calculated based on the left simulation cell and E_{sep} is calculated base on the average of the right simulation cells for 100 times in the 100 DEHP simulation situation.

Figure 3.3 shows the simulation cell for the calculation procedure of cohesive energy. E_{bulk} was directly computed from the condensed phase of molecules (simulation box on the left) by Equation 3.7. E_{ind}^t of each individual molecule was computed by the deletion of the rest of the molecules in the simulation box while maintaining the velocity profile

of each atoms unchanged. By this step, the simulation box is treated as a vacuum cell in which the target molecule cannot interact with other molecules. The deletion step should be repeated by the number of molecules in the simulation box to compute the averaged E_{ind}^t . Then the E_{bulk} was calculated by the equation 3.8. The final result of E_{coh} was then computed by the equation 3.6.

3.2.3 Procedure for heat of vaporization calculation

For small molecules in this project, instead of comparing the solubility parameters, heat of vaporization was used for the comparison as the validation criteria. This is due to the availability of data source for small molecules. The experimental heat of vaporization is available for small molecules.

The first few steps of the procedure of calculating heat of vaporization is the same as the solubility parameters until the calculation of cohesive energy. After the cohesive energy was calculated with the method in section 3.2.2, the heat of vaporization would be computed by the equation 3.9,

$$\Delta H_{vap} = E_{coh} + RT \quad (3.9)$$

where R is the ideal gas constant and T is the temperature.

The equation 3.9 comes from the concept of cohesive energy while ideal gas law is assumed. Based on this assumption, cohesive energy should be the energy of vaporization. By the definition of enthalpy,

$$H = U + PV \quad (3.10)$$

For vaporization, $V_{vapor} \gg V_{liquid}$,

$$E_{coh} = \Delta E_{vap} = \Delta H_{vap} - P(V_{vapor} - V_{liquid}) = \Delta H_{vap} - RT \quad (3.11)$$

Chapter 4

Results and Discussion

Results of properties of small molecules and plasticizers were calculated and discussed in this section. Section 4.1 focused on the comparison and validation of three different force fields for selected small molecules. Modification test of force field non-bonded parameters was also discussed in this section. Section 4.2 focused on the comparison of different routines of equilibration protocols. We want to find a decent and easy routine for the simulation of plasticizers. Section 4.3 focused on the comparison of plasticizers with different force fields. We compared the densities and solubility parameters of some commonly used phthalate plasticizers to validate our decision on the selection of force fields. The observed results of these sections would provide insights on our goal of setting up the valid prediction procedure while some information were missing.

4.1 Force Field Comparison for Small Molecules

4.1.1 Force Field Validation for Small Molecules

Small molecules including acetates, acetone, ether, alcohol and benzene, were simulated with the three force fields, PCFF, SciPCFF and COMPASS. The reason we chose these molecules was explained in section 2.1. These molecules share common functional groups, including ester groups and aromatic carbons, with plasticizers. Density and heat of

vaporization were calculated through the 2 ns production run after equilibration. Density was averaged over the 2 ns simulation. Heat of vaporization was calculated through the method shown in section 3.2.2.

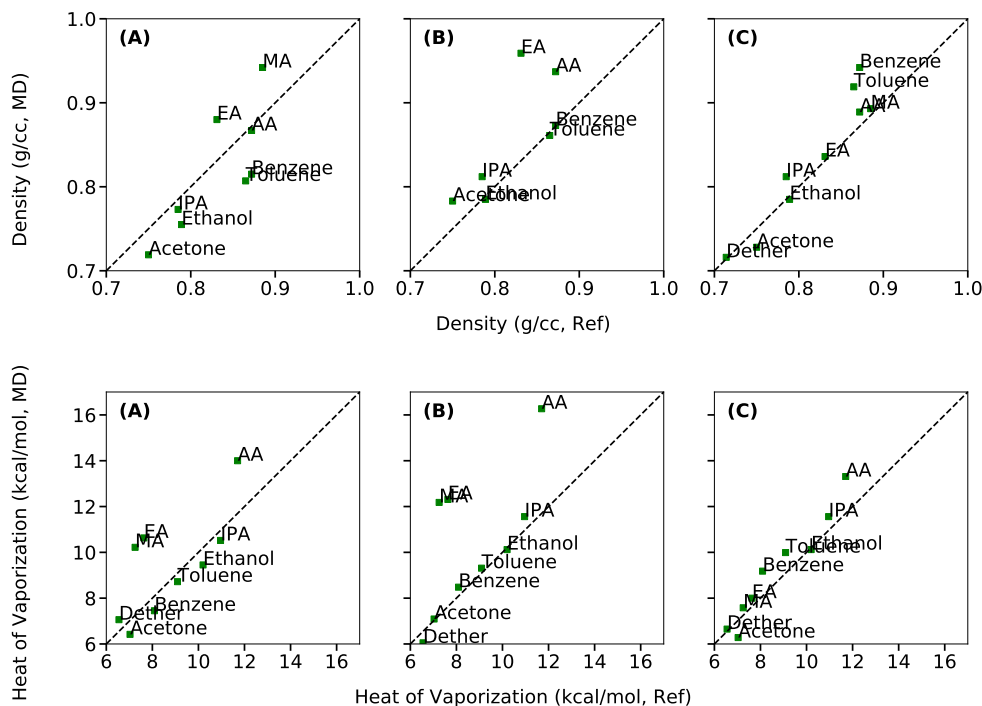


FIGURE 4.1: Comparison between the simulation results of density (above) and heat of vaporization (below) and experimental results of small molecules with PCFF (A), SciPCFF (B) and COMPASS (C). All plotted data were from Table 1. The diagonal line shows the limit of perfect agreement.

Comparison of simulation results with three force fields were shown in Figure 4.1. Both density and heat of vaporization data were compared with their experimental values corresponding to their reference experimental temperature. For density, SciPCFF seems to overestimate density of acetates more than the other 2 force fields. For the other molecules, all three force fields perform well. For heat of vaporization, COMPASS also provides the most accurate results. For molecules except acetates, all three force fields were relatively accurate. Both PCFF and SciPCFF overestimate heat of vaporization of acetates by 40% to 70%, while the difference between COMPASS and reference values

TABLE 4.1: Density and heat of vaporization comparison of small molecules with PCFF, SciPCFF and COMPASS

Molecule	Temperature (K)	Density (g/cc)			
		PCFF	SciPCFF	COMPASS	Reference
Methyl acetate	330.3	0.942	1.032	0.885	0.885
Ethyl acetate	350.3	0.880	0.959	0.836	0.831
Amyl acetate	298.0	0.867	0.937	0.889	0.872
Acetone	329.3	0.719	0.783	0.728	0.750
Diethyl ether	300.0	0.682	0.689	0.716	0.714
Ethanol	293.2	0.755	0.785	0.785	0.789
Isopropyl alcohol	293.2	0.773	0.812	0.812	0.885
Toluene	300.0	0.807	0.861	0.919	0.865
Benzene	300.0	0.815	0.873	0.942	0.872

*Reference values were experimental data from [21]

Molecule	Temperature (K)	Heat of Vaporization (kcal/mol)			
		PCFF	SciPCFF	COMPASS	Reference
Methyl acetate	330.3	10.22	12.18	7.62	7.25
Ethyl acetate	350.3	10.63	12.31	8.00	7.63
Amyl acetate	298.0	14.00	16.27	13.31	11.70
Acetone	329.3	6.42	7.09	6.28	7.03
Diethyl ether	300.0	7.06	6.05	6.65	6.55
Ethanol	293.2	9.45	10.12	10.12	10.20
Isopropyl alcohol	293.2	10.52	11.56	11.56	10.96
Toluene	300.0	8.72	9.31	9.99	9.09
Benzene	300.0	7.45	8.48	9.18	8.09

*Reference values were experimental data from [21]

is about 13%.

In conclusion, COMPASS provides more accurate results for both density and heat of vaporization compared with PCFF and SciPCFF, especially for acetates with ester groups, COMPASS performs better. This result is consistent with the fitting method of COMPASS, which means this agreement add no new information to the validation of COMPASS. Some other properties which are not used for parameterization should be tested in this case. As mentioned in section 2.1, viscosity was not used for parameterization of the three force fields, which means viscosity serves a more rigorous validation compared with density and heat of vaporization. To validate the behavior of COMPASS for other properties, the simulation of shear viscosity of some of the small molecules is added. Since the most difference among the force fields were the results of acetates, the molecules which share the substructure of ester groups, acetone, diethyl ether, ethanol and methyl acetate were compared in the next section.

Figure 4.2 shows the detailed calculation of shear viscosity for ethanol from simulation with PCFF. Here the results of ethanol were shown as a representative to show the shear viscosity calculation procedure. Other detailed figures for other molecules and other force fields were shown in the Appendix B. Figure 4.2(A) shows the direct calculations using Green-Kubo relation based on different numbers of trajectories with 10 ns NVT runs at 293.2 K. The green curve represented the averaged shear viscosity of 5 trajectories. And the yellow curve represented the averaged shear viscosity of 11 trajectories, where there were 11 trajectories in total for ethanol with PCFF. The values vs. number of trajectories were shown in case there were fitting errors when the number of trajectories were small. For the results from PCFF, the plateau could be observed around 400 ps. In this case, the cutoff time is set to be 400 ps. The standard deviation was shown in Figure 4.2(B). These values were used for fitting Equation 3.3. The fitted results were shown in Figure 4.2(C). Figure 4.2(D) incorporated two curves: one representing

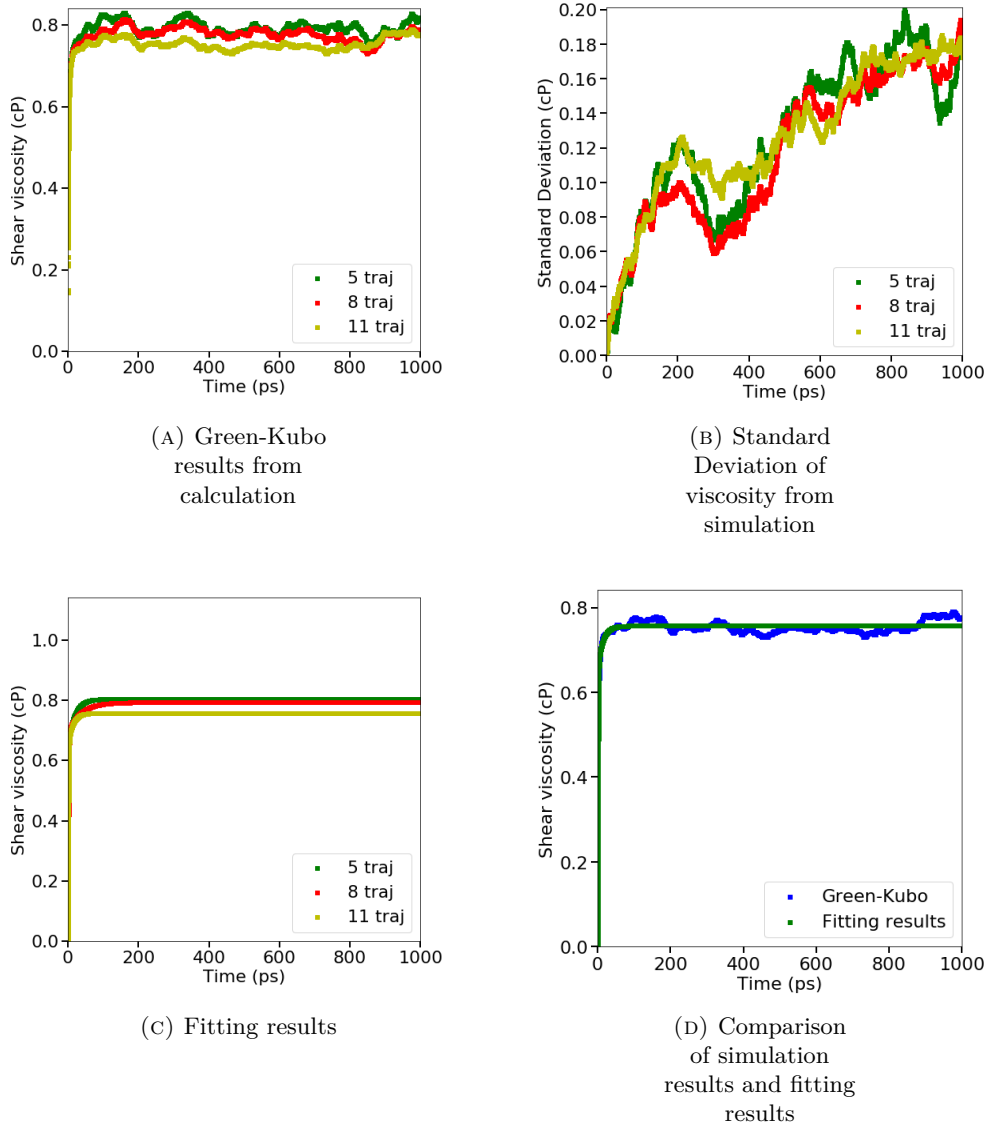


FIGURE 4.2: Detailed calculation process for shear viscosity of ethanol with PCFF as an example. Number of trajectories represented the count of trials employed in the computation of the averaged viscosity results.

TABLE 4.2: Shear viscosity comparison of small molecules with PCFF, SciPCFF and COMPASS

Molecule	Temperature (K)	Viscosity (mPa*s)			
		PCFF	SciPCFF	COMPASS	Reference
Methyl acetate	330.3	0.677	1.143	0.591	0.267
Acetone	329.3	0.269	0.398	0.302	0.236
Diethyl ether	300.0	0.453	0.356	0.383	0.220
Ethanol	293.2	0.655	1.216	1.216	1.194

*Reference values were from [41][42][43]

the simulation results of the averaged values of all trails, while the other illustrates the fitting results of previous values. Both curves are presented together in this figure to provide a comprehensive view. The shear viscosity results reported in Figure 4.3 and Table 4.2 were based on the same simulation and fitting procedure.

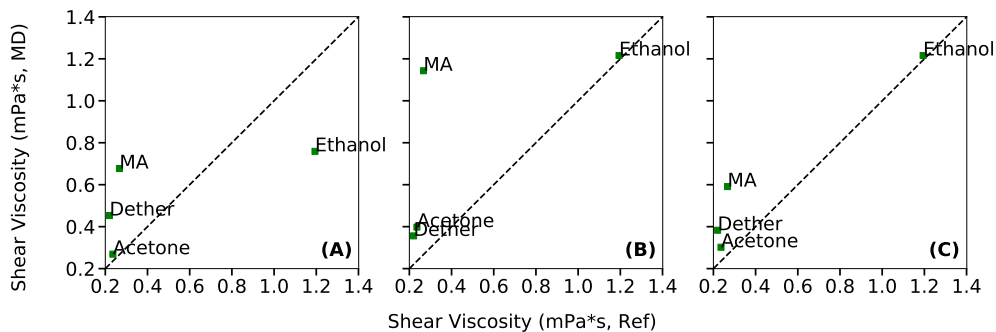


FIGURE 4.3: Comparison between the simulation results of shear viscosity and experimental results of small molecules with PCFF (A), SciPCFF (B) and COMPASS (C). All plotted data were from Table 2. The diagonal line shows the limit of perfect agreement.

Comparison of simulation results with three force fields were shown in Figure 4.3. the viscosities were compared with their experimental values corresponding to their reference experimental temperature. For acetone, all 3 force fields perform well. For ethanol, both SciPCFF and COMPASS estimated accurately while PCFF shows a large difference from the reference value, where the estimated value from PCFF is half of the reference value. For diethyl ether, the differences between the simulation results from three force fields and the reference values were around 50%. For methyl acetate, the differences were

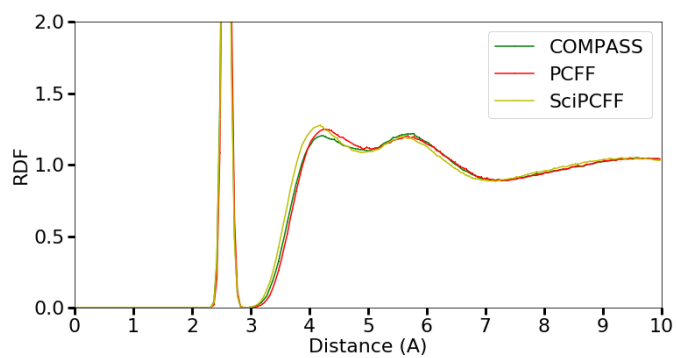
significantly strong. The result from SciPCFF is 5 times larger than the reference value and the results from PCFF and COMPASS were about 2 times larger than the reference values.

In conclusion, COMPASS also provides more accurate results for shear viscosity compared with PCFF with respect to ethanol. COMPASS and SciPCFF provide more accurate results for shear viscosity compared with PCFF with respect to ethanol and diethyl ether. However, the advantage of this accuracy is not significant.

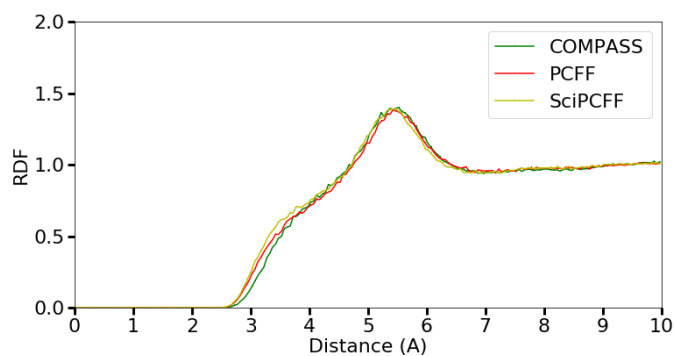
The contributions of the interactions of atoms could be further investigated by Radial distribution function (RDF). Radial distribution function is calculated to capture the neighbors of atoms as a function of distance. By definition, RDF is the number density of type 1 atoms found at distance r from a reference type 2 atom. This value is normalized. The RDF depends on the relative density of atoms which can illustrate the space structure of a system. Systems with higher RDF plots indicates more attraction behavior between molecules.

To better understand the differences in structural dimensions, the C(-H3)-C(-H3), O(=C)-O(=C) and C(=O)-O(=C) radial distribution functions were considered. The results of RDF were shown in Figure 4.4.

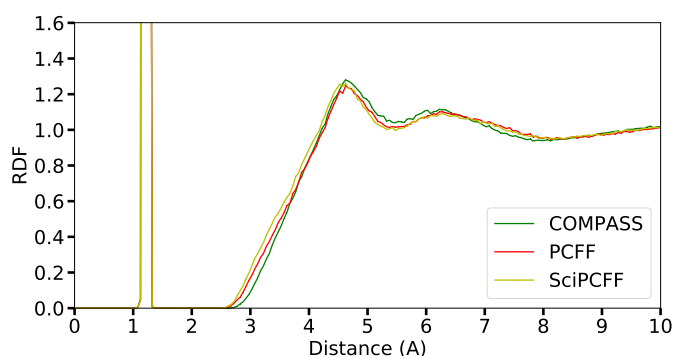
For the C-C radial distribution functions, both carbons were the sp^3 carbon from the methyl group. The first peak at 2.6 Å represents the intramolecular neighbors, and the other peaks were at 4.1 Å and 5.8 Å. There were no great differences between the overall observations of the three force fields. For the O-O radial distribution functions, the peaks of all three force fields were at 5.5 Å. No other peaks were observed. The position of the first coordination shell of acetone molecules around the reference molecule was dominated by oxygen atoms. For the C-O radial distribution functions, the carbon is the sp^2 carbon from the carbonyl group. The first peak represents the intramolecular



(A) RDF of C(-H3)-C(-H3)



(B) RDF of O(=C)-O(=C)



(c) RDF of C(=O)-O(=C)

FIGURE 4.4: Radial distribution function (RDF) of acetone with PCFF, SciPCFF and COMPASS.

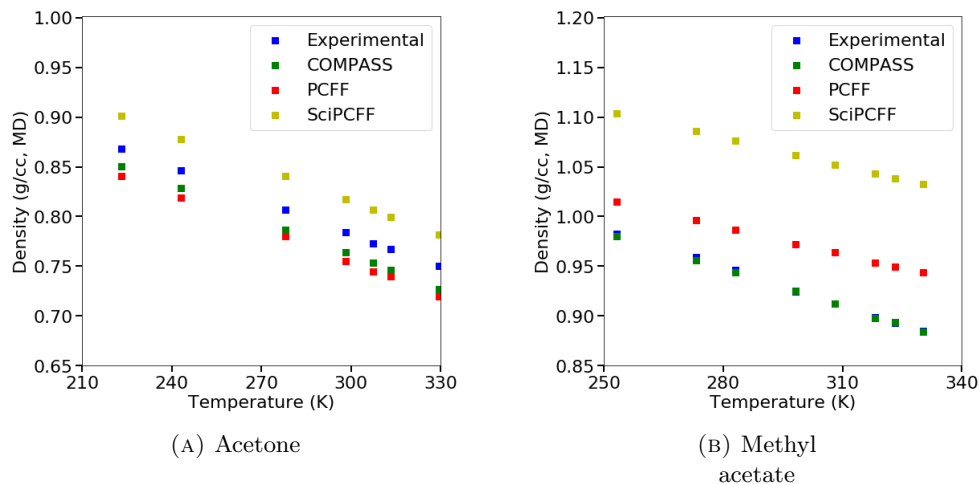


FIGURE 4.5: Density of acetone (A) and methyl acetate (B) vs. T (temperature) obtained from force fields PCFF (red), SciPCFF (yellow) and COMPASS (green). Blue dots were the reference experimental density as a function of temperature.[44][45][46][47]

bonding at 1.2 Å. The second peak which represents the intermolecular interactions was at about 4.6 Å. Then there was a second peak at around 6.2 Å. These 2 peaks suggested different orientations of C=O bonds. After 8 Å, the RDF converged to 1. In all RDFs, the order of attaining maximum values corresponds to SciPCFF, PCFF, and COMPASS. The overall RDF results showed that the estimations of three force fields were similar, while based on the density comparison results, SciPCFF overestimated the attractions and PCFF and COMPASS underestimated the attractions.

The densities of acetone and methyl acetate as functions of temperature were reported in Figure 4.5. The blue dots represented the reference experimental values at each temperature. Among the temperature range of 210 K to 350 K, the trend is smooth with no unusual transitions. SciPCFF overestimate the density consistently and COMPASS and PCFF were in close agreement with each other and underestimate the density of acetone. All three force fields provide density estimates with a precision of under 0.3 g/cc. Since there is no phase transition in this range and all force filed provide a linear correlation between density and temperature without crossover, all three force fields have

similar contributions to the thermo expansion of acetone. A comparable conclusion could be derived for methyl acetate. PCFF and SciPCFF would overestimate density of methyl acetate while COMPASS could accurately predict it. There is no phase transition or differences in the thermal expansion coefficient in this temperature range.

4.1.2 Force Field Parameters Modification for Methyl Acetate

We now try to find out whether the gaps between these force fields can be shortened by adjust a few potential parameters in selected atoms. Such information will help us improve the accuracy of openly available force fields such as PCFF when simulating more complex materials. These modified force fields were based on the PCFF. Instead of changing all non-bonded parameters from PCFF to COMPASS, only one of the parameter or combinations of parameters were modified from PCFF to COMPASS. The modification is based on the atom type. For each case where one or more atom types were selected, the values of ϵ and σ would be changed from their PCFF values to COMPASS values. Case 1 changed the non-bonded parameters of all kinds of oxygen. Case 2 changed carbonyl carbon. Case 3 changed carbonyl oxygen. Case 4 changed single bonded oxygen. Case 5 changed sp³ carbons. Case 6 changes both carbonyl carbon and oxygen. Case 7 changed both carbonyl carbon and single bonded oxygen. Case 8 changed carbonyl carbon and all oxygen. The density and heat of vaporization results were shown in Figure 4.6.

Since all the modified cases were having some of the parameters converted to COMPASS, the difference among these modified cases could show which atom's non-bonded vdW parameter would contribute more to the errors in the density and heat of vaporization among these force fields. In Figure 4.6, for density, case 1, case 2, case 3, case 4 and case 6 showed more COMPASS behavior than other cases. Meanwhile, case 5 provided overestimated density results. Case 7 and case 8 would underestimate the density. For heat of vaporization, case 1, case 7 and case 8 showed more COMPASS behavior than

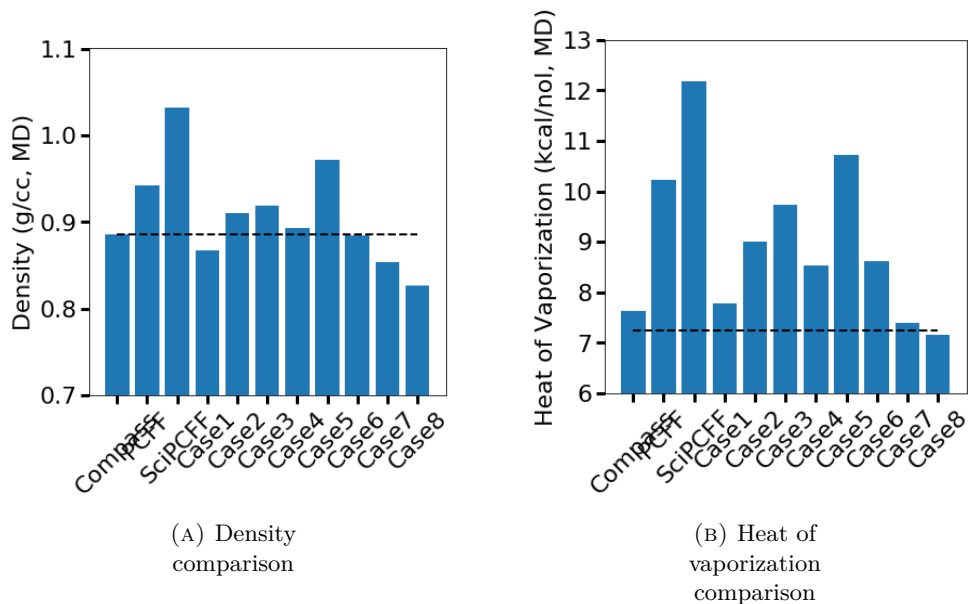


FIGURE 4.6: Density and heat of vaporization comparison of methyl acetate with modified PCFF and COMPASS parameters. Black dashed lines represent the reference values.

other cases. Case 5 provided very similar results as PCFF. Different pairs of modification would give different results.

For density, sp³ carbons provided a balancing contribution to the density compared with other cases. The modifications to oxygen and carbonyl carbons and to sp³ carbons would result in composite behaviors. Only changing parameters of oxygen and carbonyl carbons would underestimate the density. The sp³ carbon's parameter played a role to cancel this underestimation. For heat of vaporization, the single bond oxygen and carbonyl carbon provided dominant contributions.

In conclusion, upon considering all observations, it is evident that Case 1, which involved modifying the non-bonded parameters of all oxygen types, yields the most favorable overall performance. To enhance accuracy, it is advisable to concentrate on the parameters associated with oxygen atoms when adjusting PCFF parameters in future studies.

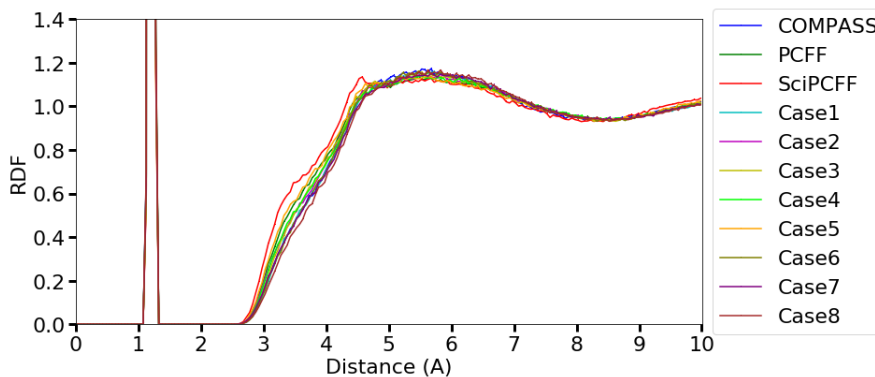


FIGURE 4.7: Radial distribution function (RDF) of C-O in Methyl acetate with with modified PCFF and COMPASS parameters.

Figure 4.7 shows the C-O RDF of methyl acetate with three force fields and the 8 cases modified force fields. The first peaks were intramolecular interactions, and the second wave showed the closest neighbors of atoms. While other force fields could provide similar behavior, SciPCFF clearly underestimate the repulsion in the LJ model at the distance region of 3 Å to 4.5 Å. From 4.5 Å to 6 Å, the RDF of the other cases except SciPCFF were converging to the RDF of SciPCFF. After 6 Å, all force fields produce the same behavior.

4.2 Equilibration protocol comparison

The equilibration protocol is also important to MD simulations to get steady and representative results. Due to our simulation of distinct systems compared to those previously described by Li et al., we have the potential to realize computational time savings. The reduction of computational time is critical for efficient data collection. In accordance with the prototype protocol, two additional modified protocols are introduced in this section for the purpose of comparative analysis. The original protocol was introduced in the methodology section 3.1. The first few steps remain the same as used in the small molecule protocols. The step 4 of this heating-cooling protocol in the methodology section was modified in these cases. In original step 4, repeated heating-cooling cycles were

applied to the system. There were 2 more alternatives for step 4. One is using pure heating cycles to replace the heating and cooling cycles. The other is to insert compression procedure during cooling cycles to increase the density from the high temperature memory. Figure 4.8 provided a straightforward procedure based on the timeline.

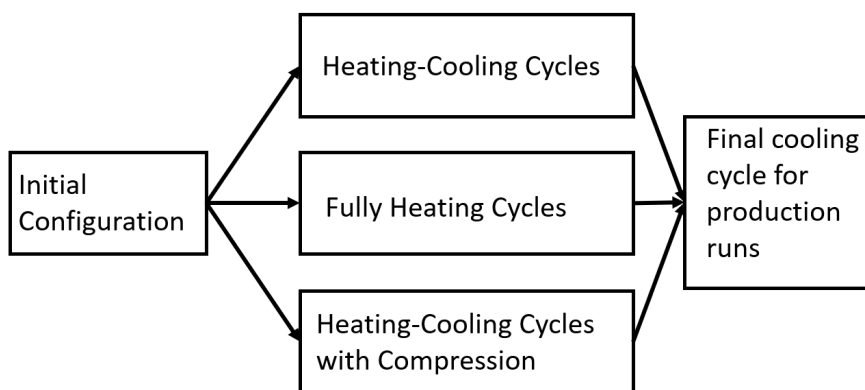


FIGURE 4.8: Alternatives for the heating-cooling cycles in the equilibration step.

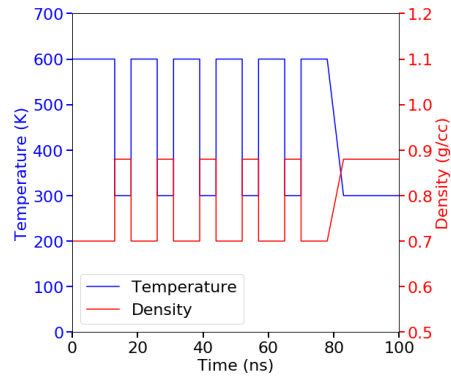
The rationale behind replacing heating-cooling cycles by pure heating procedure was based on a hypothesis. Assuming the cooling procedure during the repeated heating-cooling cycles does not contribute as the heating procedure, there is no need to keep the cooling procedure in the protocol. The purpose the repeated heating and cooling cycles were to let the polymers in the cells move around freely to their well equilibrated position until the distribution of molecules reaches a uniform level. In this procedure, the crowded molecules may move from their dense area to sparse area. In the final well-equilibrated cells, the distribution of free volume of individual molecules will be the same among all areas in the cell. At high temperature, the velocities of molecules are higher, which gives molecules more kinetic energy to move around. The time required for the molecules moving to a equilibrated state may be shorter. In other words, with the same amount of simulation time, high temperature condition may provide better equilibrated cells compared with low temperature condition. In the comparison section,

if the pure heating cycles give an early convergence of density and solubility parameters, our hypothesis would be validated.

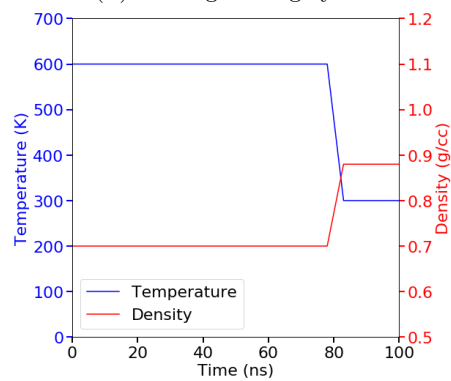
The rationale behind inserting a compression procedure during the cooling cycle was also a consideration about molecular movement and distribution. With a compression procedure, the density will be higher. The external stronger force would push the chain to move. Thus the high density condition would help the movement of molecules. During expansion, the external force would be reduced and the polymer chains would be pushed out from a high density condition back to their original density. However, there is also a side effect for this procedure. The external force would also reduce the free space for the movements of molecules. The effects of compression will be compared in the following section.

DEHS has only alkane chains in the torso. (The chemical structure of DEHS was listed in Appendix A) The branching configuration of DEHS is the 2-ethylhexyl group, which represents the most common substitution in our interesting plasticizer pool. For the comparison of equilibration types, DEHS was used. Figure 4.9 provided the temperature and density profiles based on the timeline.

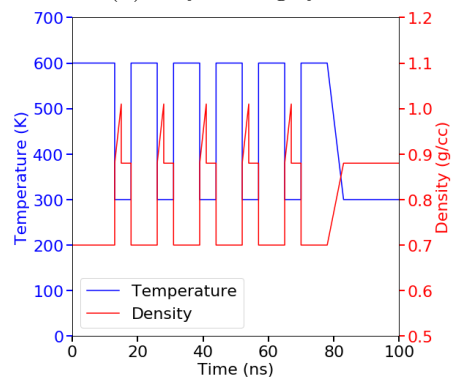
Besides the different protocols, the number of repeated cycles were also compared in the following sections. Originally in Li et al's research[8], 5 to 7 cycles were used for plasticized PVC[8]. In the validation of force fields of plasticizers, not as many cycles may be needed. For the purpose of saving computational resources, results from different number of cycles were compared. Meanwhile, the total number of cycles increased to up to 11 to check the convergence of results in the long time relaxation regime. For the comparison of number of cycles, DITP and TOTM were used. (The chemical structures of DITP and TOTM were listed in Appendix A) DITP features isoalkyl group branches in its structure, while TOTM exhibits an unusual three-legged structure attaching to the aromatic groups on its torso, which contributes to a reduced diffusion rate. The size



(A) heating-cooling cycles



(B) fully heating cycles



(C) heating-cooling cycles with compression

FIGURE 4.9: Temperature and density profiles for the three protocols during the amorphous simulation cell generation process for three testing plasticizers. There were 6 cycles in this example case. The density profiles were based on DEHS for example.

of these plasticizer molecules is sufficiently large so that equilibration protocols effective for them would be applicable to most other smaller plasticizers as well.

In the subsequent sections, solubility parameters were selected for comparison among various equilibration protocols, specifically concerning the outcomes for plasticizers. Solubility parameters serve as indicators of the challenges associated with molecular separation or mixing. Given the objective of establishing a protocol to investigate the behavior of additives, testing solubility parameters is a pertinent choice.

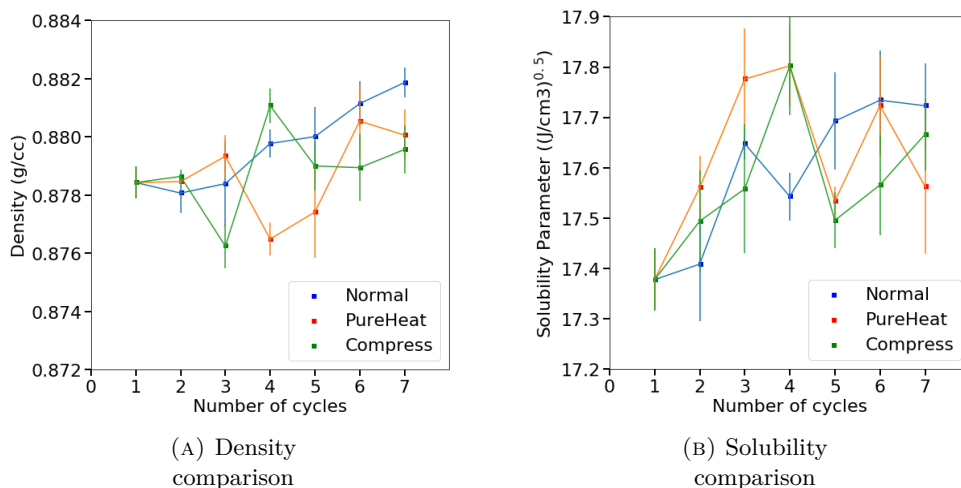


FIGURE 4.10: Density (A) and solubility parameters (B) of DEHS with three equilibration routine. Force field is PCFF. Error bars calculated from standard error.

Density and solubility parameters were calculated through the 2 ns production run. Density was averaged over the 2 ns simulation. The simulation results of DEHS were displayed in Figure 4.10. Blue curve represents the heating and cooling cycle method. Orange curve represents the routine where the cooling cycles were replaced by pure heating cycles. Green curves represents the routine where compression of the simulation system were added to control the density during the cooling cycles. The error bars came from the repeated simulation runs. The large uncertainties were caused by the small simulation box. For the future interest of comparison, larger simulation cells are required.

Based on the blue curve, the density and solubility parameters from the heating and cooling cycles would converge well after 5 cycles. The differences among the results from simulation with 5, 6 and 7 cycles were about 0.2% for density and 2% for solubility parameters. The other 2 curves also indicated convergence after 5 cycles. However, the fluctuation from these 2 routine would be stronger than the heating and cooling cycle method for both density and solubility parameters.

In conclusion, there is no need to convert to other routine. The modification of the equilibration steps would not enhance the simulation behavior. The hypothesis discussed in the previous section might not serve well for the rationale behind the behavior. Heating and cooling cycles in the equilibration protocol are sufficient to be used for simulation.

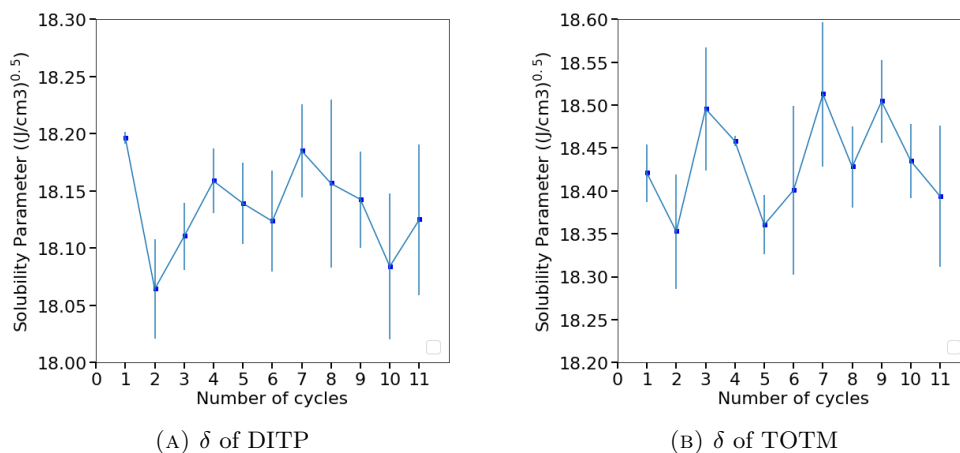


FIGURE 4.11: Solubility parameters of DITP (A) and TOTM (B) along cycles. Force field is PCFF. Error bars calculated from standard error.

After the comparison among the three routine, the traditional heating and cooling cycle routine was chosen to be used in the equilibration protocol. For the number of cycles comparison, heating and cooling cycles would remain unchanged. In this section, 2 more plasticizers were used for the comparison, DITP and TOTM. These 2 plasticizers have more molecular weight. The time for equilibration might be longer for them. The comparison results were shown in Figure 4.11. By increasing the number of heating and

cooling cycles, the convergence did not improve as well significantly. The uncertainty was about 1% in the results. For these 2 large molecules, 6 cycles were sufficient to give an idea of how solubility parameters performed. Our protocol for simulation would remain unchanged in the comparison of plasticizers and future QSPR modeling applications.

4.3 Force Field validation for plasticizers

In this section, all molecular simulation production routines were the same as shown in the methodology section. Including DEHP, some commercially used plasticizers were simulated to validate the accuracy of three force fields. The chemical structures of these plasticizers were listed in Appendix A.

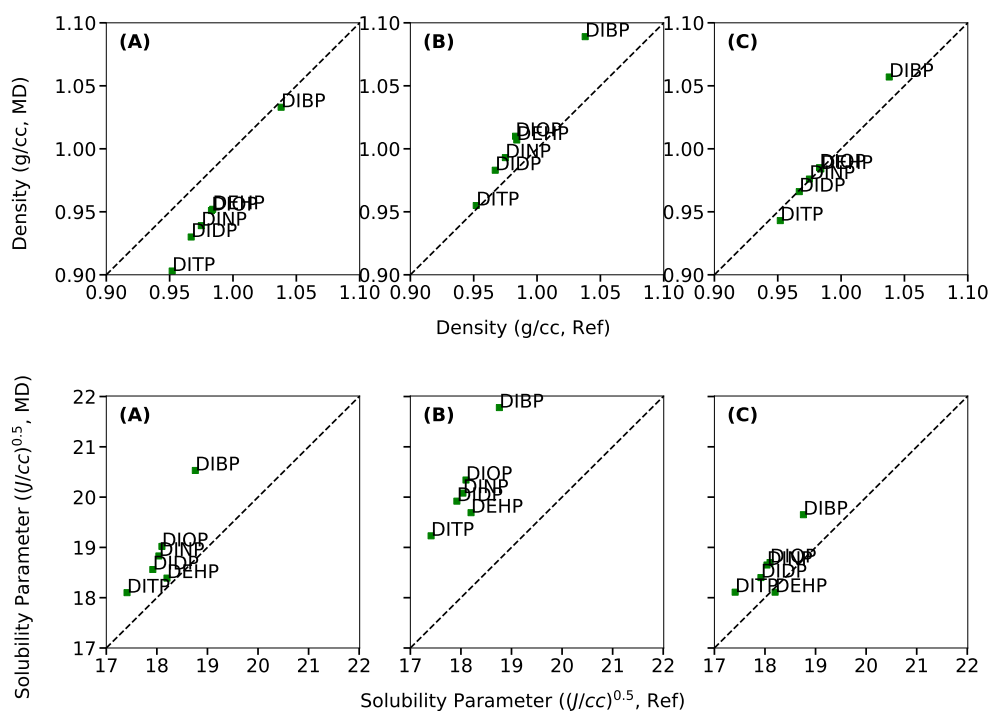


FIGURE 4.12: Comparison between the simulation results of density (top) and solubility parameters (bottom) and reference results (experimental results for density and group contribution method for solubility parameter) of plasticizers with PCFF (A), SciPCFF (B) and COMPASS (C). All plotted data were from Table 3. The diagonal line shows the limit of perfect agreement.

Density and solubility parameters were calculated through the 2 ns production run. Density was averaged over the 2 ns simulation. Solubility parameters were calculated through the method shown above. The results were displayed in Figure 4.12 and Table 4.3.

TABLE 4.3: Density and solubility parameter comparison of plasticizers with PCFF, SciPCFF and COMPASS

Molecule	Density (g/cc)			
	PCFF	SciPCFF	COMPASS	Reference
DIBP	1.089	1.033	1.057	1.038
DIOP	1.010	0.951	0.985	0.983
DINP	0.993	0.939	0.976	0.975
DIDP	0.983	0.930	0.966	0.967
DITP	0.955	0.903	0.943	0.952
DEHP	1.007	0.952	0.984	0.984

*Reference values were experimental data from [48]

Molecule	Solubility Parameter (J/cc) ^{0.5}			
	PCFF	SciPCFF	COMPASS	Reference
DIBP	21.78	20.53	19.65	18.76
DIOP	20.34	19.02	18.70	18.10
DINP	20.08	18.83	18.65	18.04
DIDP	19.92	18.56	18.40	17.92
DITP	19.23	18.10	18.11	17.41
DEHP	19.69	18.39	18.11	18.20

*Reference values were from [49] as GCM results

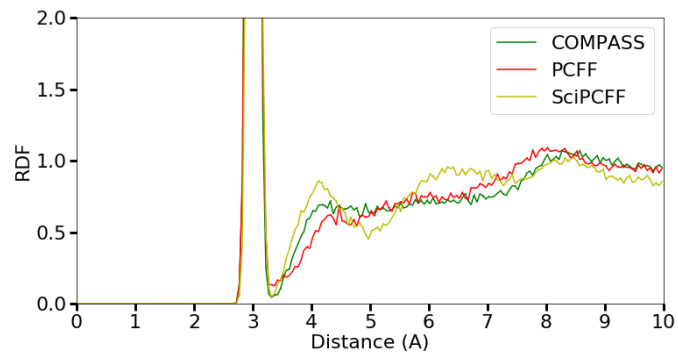
Comparison of simulation results of three force fields were shown in Figure 4.12. Density was compared with their experimental values around 300 K. Solubility parameters were compared with the results from group contribution method. Unlike the comparison conclusions from small molecules, for density, PCFF seems to underestimate density of plasticizers more than the other 2 force fields. SciPCFF would slightly underestimate density and COMPASS provided more accurate results. These results were different from our observations in section 4.1.1, where SciPCFF would overestimate density greater than PCFF.

Meanwhile, for solubility parameters, the reference values were from group contribution method by Small[49][50]. This method contains errors so we focused on the trend. The simulation results from the three force fields were all slightly higher than the reference values while all three force fields captured the similar trends and patterns for solubility parameters as the group contribution method.

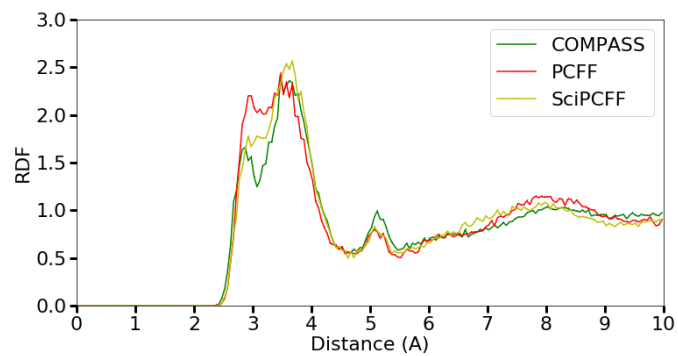
For the purpose of comparing the density of DEHP and the distribution behind this, we also applied RDF analysis for DEHP with three force fields. The RDF results of DEHP were shown in Figure 4.13. All normalized RDF curves converged to 1 after 9 Å. The RDF values were properly normalized.

In Figure 4.13(A), the first peak at 3 Å indicated the intramolecular sp² carbons. The peak at 4 Å contained both intramolecular and intermolecular interactions. In Figure 4.13(B), the first large peak was observed at 3 Å and 3.6 Å, while at 5.1 Å, there was a smaller peak. In Figure 4.13(C), the first peak at 1.1 Å indicated the intramolecular C=O bonds. The peaks for (B) and (C) at 3 Å and 4 Å contained both intramolecular and intermolecular interactions. Due to the double leg structure of DEHP, unlike the previous RDFs, one more plot for only intermolecular interactions was shown in Figure 4.14.

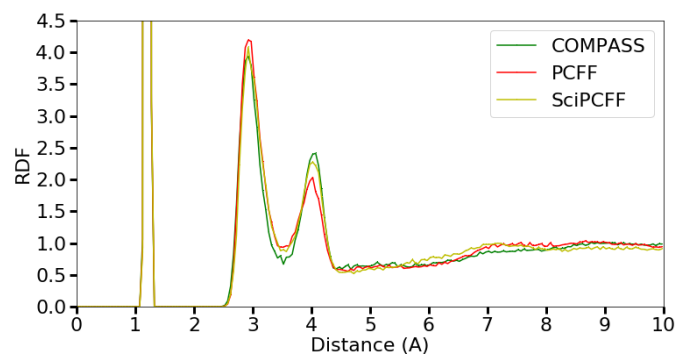
In Figure 4.14, different force fields provided different RDF results. At the peak around 3.4 Å, the peak intensity decreased from SciPCFF, COMPASS, to PCFF. From 4 Å to 5.7 Å, the RDF of SciPCFF increased not as much as the other two force fields. In the range of 6 Å to 8 Å, PCFF and SciPCFF showed very similar RDF while COMPASS provided the lowest RDF. The overall comparison showed different estimations of three force fields. SciPCFF overestimated the attractions in both first and second solvation shell while PCFF only overestimated the attractions in the second solvation shell. The observed trend was the same as we found for small molecules in section 4.1.1.



(A) RDF of C(=O)-C(=O)



(B) RDF of O(=C)-O(=C)



(c) RDF of C(=O)-O(=C)

FIGURE 4.13: Radial distribution function (RDF) of DEHP with PCFF, SciPCFF and COMPASS

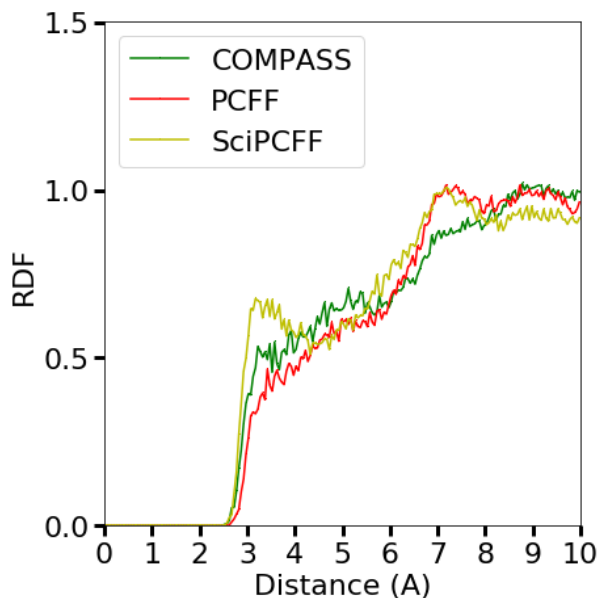


FIGURE 4.14: Radial distribution function (RDF) of intermolecular C(=O)-O(=C) for DEHP with PCFF, SciPCFF and COMPASS

In conclusion, COMPASS provides more accurate results for density compared with PCFF and SciPCFF. Meanwhile, all three force fields predict the solubility parameters with the same trend as the group contribution method.

4.4 Discussion of general observations

Previous sections provided some conclusions from different aspects. Section 4.1 focused on the comparison and validation of three different force fields for small molecules containing alkane chains, carbonyl groups, ethers and alcoholic groups. For these molecules, the non-bond parameters of force fields were tested and compared. COMPASS would achieve the most accurate simulation results and SciPCFF would achieve better results than PCFF. In section 4.2, different routines of equilibration protocols were compared. The attempt to upgrade our simulation protocol did not achieve. We keep our simulation protocol the same for all the simulations in the end. Section 4.3 focused on the plasticizers with three different force fields. We validated the density and solubility parameters

for these plasticizers. COMPASS still showed more decent results than the other two force fields while for plasticizers, PCFF showed better accuracy than SciPCFF.

Clearly there were some differences between the conclusions from our observations of small molecules and plasticizers. SciPCFF provided more accurate results for small molecules than those by PCFF. Meanwhile, PCFF showed higher accuracy for plasticizers. The reasons causing the differences of plasticizers with three force fields might be related to the rationale of the effects of simulation methods.

In section 4.1.2, we showed that changing the force field parameters of oxygen from PCFF to COMPASS would underestimate the density of small molecules, while changing the parameters of sp³ carbons would overestimate the densities. Meanwhile, for heat of vaporization, single bond oxygen and carbonyl carbon would contribute more COMPASS-like results than other parameters. For both density and heat of vaporization, case 1 provided more similar simulation results in all cases. Parameters of oxygen would dominate the differences.

Initially we thought the differences between SciPCFF and PCFF were caused from all their different parameters. However, COMPASS has very different settings for oxygen atoms. We found that PCFF perform similarly to COMPASS instead of SciPCFF. Firstly we focused on the comparison of some small molecules containing oxygen groups. COMPASS would result in more decent heat of vaporization and shear viscosity. Meanwhile, SciPCFF had more similar behavior as COMPASS than PCFF. This might be caused by the differences in parameters. While both PCFF and SciPCFF have smaller ϵ and larger σ , the values of SciPCFF were much similar to COMPASS than PCFF. For some of the parameters, SciPCFF has the same values as COMPASS like sp³ carbons. For plasticizers, COMPASS would also perform the best in these three force fields while SciPCFF would shift away from the PCFF and COMPASS. It seems that compounds like phthalates and adipates may be more properly simulated by COMPASS and PCFF.

The balance of each parameters in SciPCFF works better in small molecules while it works worse for plasticizers.

In our section 4.1 and 4.3, different force fields were compared. We would like to improve the behavior of PCFF since PCFF is usually available for most researchers. The improvement from PCFF to COMPASS is clear. However, we still miss some parameters in COMPASS. For the future research on other kinds of plasticizers such as phosphates and epoxides[4], more types of atoms and structures will be involved. Our suggestion for dealing with this kind of question is using a modified PCFF setting. The bonding parameters will remain the same as PCFF. Most non-bonded vdW parameters will also remain unchanged. Meanwhile, the ϵ (the well depths) and σ (size of the particle) for oxygen will be modified. Instead of using the exact values for oxygen parameters in COMPASS, we could apply an optimization routine to find a proper value for oxygen atoms.

Our completed work was only a few initial steps towards our larger goal. In the progress of improving the accuracy and efficiency of MD simulation, we explored the the force fields and equilibration protocols. For force fields, the validation and comparison of PCFF, SciPCFF and COMPASS start an opening move in the developing and selecting appropriate force fields for our specific research questions. The modification of parameters of force fields is still far away from perfectly capturing the behavior of specific molecules or systems, but it provides an insights on that route. Due to the limited amount of simulation trials and comparisons we have, the equilibration process in MD simulations still required further research for our specific system. For our long term goal, an accurate and efficient data collection progress before data treatment by QSPR or other modeling systems is desired. Therefore, accurate force fields and sufficient equilibration protocols are essential since they are crucial to MD simulations.

Part II

Exploration of QSPR Modeling for Predictions of Properties

Chapter 5

Introduction: Review of QSPR

5.1 Introduction of QSPR and Descriptors

A quantitative structure property relationship (QSPR) is a method based on simple and multiple linear regression. The foundation is principle of poly-linearity, which assumes the linear relationship between the property and other descriptive parameters[51]. As mathematical relationships are established, these studies enable the prediction of molecular behavior for new chemicals or hypothetical molecules. Consequently, the fundamental framework of the QSPR technique can be mathematically expressed as follows:

$$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + \dots + a_nX_n \quad (5.1)$$

where X are the independent variables of chemical structural features or physico-chemical properties in the form of numerical quantities and Y is the model response. a_0 to a_n are the contributions of corresponding X values.

Figure 5.1 shows the general procedure of QSPR. Generated descriptors were used as input to the model. A multi-linear regression model will provide a preliminary result while the quality will be judged by the correlation coefficient and the standard error.

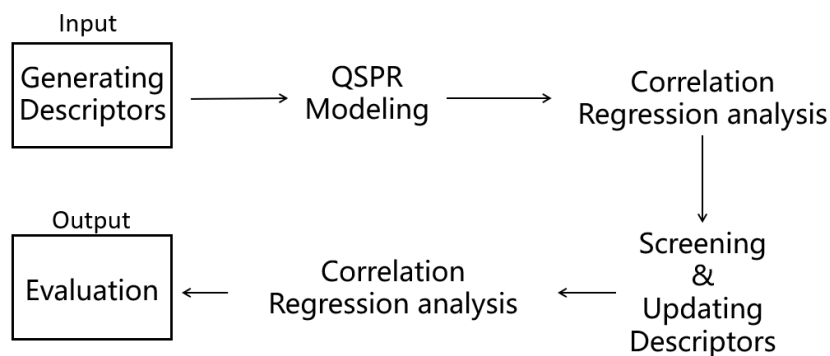


FIGURE 5.1: General process for QSPR method

Then further analysis will be carried to decide if more or fewer descriptors were needed from the current results. With the new settings of descriptors, one more step of regression analysis will occur to generate the final result.

QSPR use molecular descriptors as input variables. These descriptors could represent the molecular structures of molecules since descriptors are usually generated from mathematical transformations of some distinguishable characteristics of molecular structures. Descriptors can describe the properties of molecules from different aspects. they can describe the 2D or 3D structures of a molecule and they can also represent sub-structures of a molecule[52]. The simplest descriptor for a molecule would be simple counts. For example, the number of carbons in an alkane chain can be considered as a descriptor which represents the length of the alkane chain. Descriptors also includes physicochemical properties, topological indices etc. Wiener index and Randic Branching index were used as my initial step since these descriptors do not require simulation to be extracted. These descriptors were calculated from the structures of molecules which could work as index values to represent the molecules in polymer design[53].

After the descriptors were carried out, a correlation relationship could be formed from the regression analysis between the descriptors and the physical properties data. This

QSPR model would be trained and selected based on the correlation coefficient and the standard error to make sure the QSPR model is reflecting the relevance of descriptors and target properties.

Chemical structures of molecules would affect the chemical or mechanical properties of molecules. Chemical structures could be captured by descriptors while there were many researchers working on the effects of polymer design parameters on polymers. QSPR models would allow quick and simple predictions about certain mechanical or chemical properties which we are interested in. For our research on investigating effects of polymer design parameters, QSPR might provide unexpected and interesting insight.

5.2 Motivation of investigating QSPR

The motivation for this project starts from the fact that doing infinite number of experiments to find a desired polymer suitable for an application is impossible. New plasticizers are desired while there are too many possible variations of current existing plasticizers. To investigate and explore the rationale of plasticizers, an extremely large number of experiments and simulations would be required. Instead of wasting time on repeating meaningless experiments or simulations, using a data-driven modeling method to predict properties from information those we already know would be more efficient for polymer design. Further studies about this correlation method could be motivated by chemo-informatics. With the assistance of computers, more problems could be solved.

Many researchers worked on the application of QSPR to the prediction of glass transition temperatures of different polymers. Mathias et al.[54] showed their work on calculating glass transition temperatures for acrylate and methacrylate polymers with QSPR model. The QSPR correlation was a limited approach for this prediction due to the high sensitivity of glass transition temperatures to sample history and experimental conditions. After that, Chandola et al.[13] offered an application of their QSPR model. This

model is based on an equation derived from a training dataset encompassing energy-based and formation-based descriptors. Through this model, it is possible to predict the low-temperature flex point of polyvinyl chloride (PVC) plasticizers using parameters such as potential energy of the molecule, effective dipole moment, and molecular volume. Importantly, the predicted values consistently fall within an acceptable range. These studies show an idea of predicting properties with existing data. An expansion of the QSPR method might be helpful in the area of prediction of properties without doing too many experiments or simulations. These applications would serve as a guideline for design of new plasticizers.

5.3 Motivation of investigating descriptors

As a start, the simple counting number of atoms would be used to represent some basic features of molecules. This can distinguish the molecule size directly. In this category, the simple counting could be divided into several different descriptors such as the number of carbons in a specific region of a molecule. In this case, these descriptors would reflect the sub-structures of molecules. However, these simple descriptors cannot distinguish the difference between molecules with similar structures in the leg part if the number of atoms is the same, which means comparing the isomers could not rely on simple count descriptors. Molecular shape is also important to be discussed. Therefore, new descriptors need to be found in this step to reflect the minor differences between isomers. Then there are some characteristics required for these descriptors.

The extraction or calculation of such descriptors do not require extra simulations or experiments. Usually this means that the descriptors are describing the structure or sub-structure properties instead of physicochemical properties. We already use simulation results as the model training data. The extra simulations or experiments to extract the values of descriptors. Thus, the values of such descriptors should be easy to calculated

directly from the structure information of molecules. Another point is that the descriptors should tell the difference between two isomer plasticizers. Simple counts could not achieve this purpose. For example, suberate (hexane), phthalates (benzene) and hexamolls (cyclohexane) all have six carbon atoms in the torso group while the structures are different. Two examples of such kinds of descriptors are Wiener index and Randic Branching index.

Wiener index and Randic Branching index are two topological indices. Both of them belong to Graph Theory[53]. Wiener index was first introduced by Wiener in 1947, which was used for the study in boiling points of paraffin[55]. It was firstly defined as “the sum of the distances (the length of the shortest route between 2 atoms) between any two carbon atoms in the molecule, in terms of carbon-carbon bonds”. Then it was expanded to all non-hydrocarbons. In the chemical graph, it is identical to the half of sum of distance matrix. The equation for Wiener Index is shown below, where i and j represents the i th and j th atoms. One-half is to eliminate the over calculation since the distance between atom i and atom j and is equivalent to the distance between atom i and atom j .

$$Wiener \ index = \frac{1}{2} \sum_{i,j} d_{i,j} \quad (5.2)$$

Randic Branching index, which is a widely used molecular structure descriptor, was firstly presented by Milan Randic in 1975[56]. The definition equation is shown below. The variables m and n represent the number of the adjacent (bonded to each other) points joined by each bond. The pairs of m and n represent the decomposition of bond types. For example, the carbon-carbon bond in ethane has the m and n values as 1-1. For the C-C bonds in propane, the bond type of each bond can be represented as 1-2. Based on the illustration of bond types, the Randic Branching index includes the

hydrogen-suppressed structure information of all bonds.

$$R = \sum_{\substack{\text{all} \\ \text{bonds}}} \sqrt{\frac{1}{mn}} \quad (5.3)$$

One more thing need to be aware is that these topological indices could not contain all structure information of a molecule. For example, the cis/trans isomerism or chirality cannot be reflected through Wiener index and Randic Branching index. However, as we mentioned before, our descriptors do not need to include all structure information of a molecule. We only need the information related to distinguish the plasticizers.

For the purpose of exploring the dependency of plasticizers to substructures, we want to minimize the correlation among the descriptors we used. For example, we do not desire the size of a molecule to be included in every descriptors. Therefore, these topological indices were modified to be used as descriptors. The details of modification of these indices would be introduced in the next chapter.

Finding perfect descriptors for property prediction involves extensive trial and error, while our work provides early exploration towards that goal. Hopefully the exploration of such descriptors would become motivating for the further study of QSPR models. With the ongoing research, the fundamental understanding of effects of chemical structures to the mechanics of plasticization would be explored.

Chapter 6

Methodology: Modeling Preparation

6.1 QSPR Data Selection

For the QSPR, we select the data on the solubility parameters from an identical simulation protocol. Building on the earlier progress, the simulation protocol is identical to section 3.1 for plasticizer molecules. Since most simulation results were driven before the investigation of force fields and protocol comparison, all results in this Part were simulated based on PCFF and the original protocol from Li et al.[8]. Details of the descriptors and solubility parameters of the plasticizers were listed in Table 6.1. Most plasticizers are commercially used, where their chemical structures are listed in Appendix A.

Since the dependency of properties to the chemical structure is our interest, only the descriptors which could reflect the chemical structures were included in the data-set. Some simple count descriptors such as number of carbons in the leg (NCL) or in the torso (NCT) were used. Number of carbons in the aromatic group (NCA) was used as an indicative descriptor which reflects the existence of benzene ring.

TABLE 6.1: Descriptors and solubility parameters of plasticizers

Name	NCT torso	NCL leg	NCA aromatic	Wiener normalized	Randic normalized	δ (J/cc) ^{0.5}
succinate_13	3	13	0	0.9	0.94	17.0
succinate_10	3	10	0	0.9	0.94	17.29
succinate_8	3	8	0	0.9	0.94	17.77
succinate_6	3	6	0	0.9	0.94	18.48
hex_4	6	4	0	0.71	0.97	18.84
hex_7	6	7	0	0.71	0.97	17.88
hex_9	6	9	0	0.71	0.97	17.6
hex_13	6	13	0	0.71	0.97	17.17
DITA	4	13	0	1	1	17.2
DEHS	8	8	0	1	1	17.46
DINA	4	9	0	1	1	17.74
DEHA	4	8	0	1	1	17.88
JJDIUP	6	11	6	0.71	0.97	17.97
JDIUP	6	11	6	0.71	0.97	17.99
DUP	6	11	6	0.71	0.97	18.12
TOTM	6	8	6	0.7	0.95	18.16
911P	6	10	6	0.71	0.97	18.31
DDP	6	10	6	0.71	0.97	18.46
DIHA	4	6	0	1	1	18.7
DOTP	6	8	6	0.74	0.97	19.0
DIHP	6	7	6	0.71	0.97	19.37
DIMP	6	6	6	0.71	0.97	19.55
DEHP	6	8	6	0.71	0.97	19.2

Wiener index and Randic Branching index were also included in the data set. For the purpose of comparing the effects of torso, these topological indices in this report would be defined with some differences. We propose our special and unique definitions of Wiener index and Randic Branching index for substructure of molecules for the purpose of reflect substructures of these molecules. In our definitions, both values only consider the torso part of plasticizers. The indices are calculated based on the separate middle head group which attached to the two ester groups. This tells the shape and size of the torso group. Meanwhile, the carbons in the carbonyl group connecting the oxygen and carbon chain legs are treated as extra atoms to the torso group for non-misleading calculation. This tells the connecting position of torso and ester groups. For example, the Wiener index for DEHP in this project is reported as the Wiener index for an o-Xylene (1,2-Dimethylbenzene) while the the Wiener index for DOTP in this project is reported as the Wiener index for a p-Xylene (1,4-Dimethylbenzene).

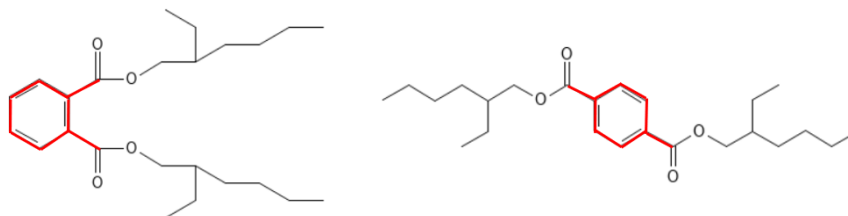


FIGURE 6.1: Chemical structure of DEHP (left) DOTP (right). Red lines represent the carbon atoms used for topological indices

For the purpose of comparing only the attachments of carbon bonds instead of including the whole structure, both Wiener index and Randic Branching index were normalized. The origin values calculated from definition equations were divided by a linear structure value so that the effects of the chain length would be eliminated in this descriptor comparison for our QSPR model. The new normalized values were demonstrated in Table 6.1 and then they were used in the prediction models.

Not only some commonly used plasticizers are estimated in the QSPR project, we added more molecules which were similar to the structure of DEHP for the purpose of extend a larger data-base for comparison. These molecules are called "imaginary plasticizers" in this project. The leg part of these "plasticizers" are the same as DEHP, which is a ethylhexyl group. The torso group is different from DEHP. In this project, the imaginary plasticizers all have alkane carbon chains with or without branches in the torso group. Two examples are shown in Figure 6.2.

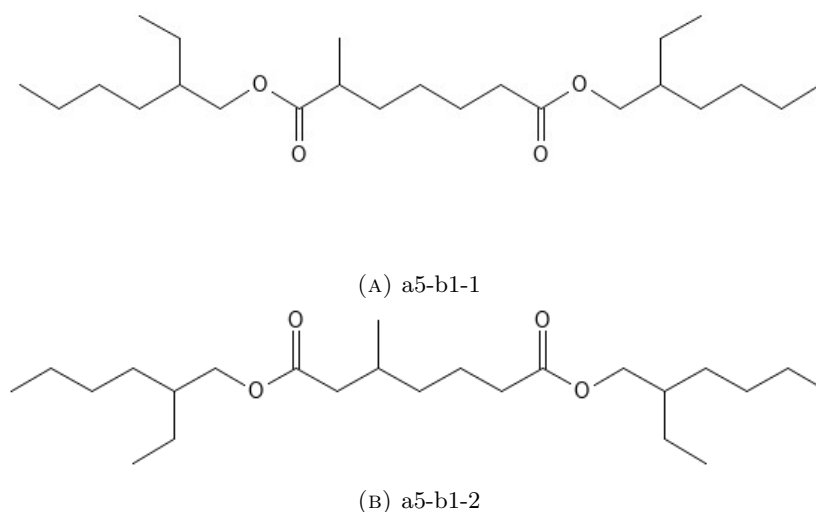


FIGURE 6.2: Chemical structure of examples of "imaginary plasticizers": a5-b1-1 (A) and a5-b1-2 (B)

In the molecule "a5-b1-1", the aromatic group of DEHP is replaced by a carbon chain. "a5" means the distance between the two ester group is 5 carbon atoms. "b1-1" means there is one branch attached to the main body while this branch is at carbon atom number 1 of the main chain. Similarly, for the molecule "a5-b1-2", there is a branch at carbon atom number 2 of the main chain. The detailed structure information of all other imaginary plasticizers could be found in the Appendix A4.1. The details of the descriptors and solubility parameters of the imaginary plasticizers were listed in Table 6.2. The solubility parameters were calculated based on the identical simulation protocol in section 3.1.

TABLE 6.2: Descriptors and solubility parameters of imaginary plasticizers

Name	NCT torso	Wiener	Wiener normalized	Randic	Randic normalized	δ (J/cc) ^{0.5}
a4-b1-1	5	52	0.93	3.27	0.96	16.16
a4-b1-2	5	50	0.89	3.31	0.97	17.78
a4-b2-13	6	71	0.85	3.66	0.94	15.7
a4-b2-22	6	67	0.8	3.62	0.93	17.62
a5	5	56	1	3.41	1	17.75
a5-b1-1	6	79	0.94	3.77	0.96	16.25
a5-b1-2	6	76	0.9	3.81	0.97	17.45
a5-b1-3	6	75	0.89	3.81	0.97	17.51
a5-b2-22	7	98	0.82	4.12	0.93	17.5
a5-b2-24	7	100	0.83	4.2	0.95	17.28
a5-b2-33	7	96	0.8	4.12	0.93	17.39
a6	6	84	1	3.91	1	17.57
a6-b1-1	7	114	0.95	4.27	0.97	17.46
a6-b1-2	7	110	0.92	4.31	0.98	17.76
a7	7	120	1	4.41	1	17.33
a7-b1-1	8	158	0.96	4.77	0.97	16.4
a7-b1-4	8	149	0.9	4.81	0.98	17.47
a7-b2-11	9	198	0.9	5.06	0.93	15.47
a7-b2-44	9	180	0.82	5.12	0.95	17.15

6.2 QSPR Molecular Modeling

The descriptors used for this study represent the conformational information such as the number of carbons in the leg. Details of descriptors are listed in Table 6.1 and 6.2. Multi-variant linear regression model was used to fit the data set. Input values were the descriptors of training and testing molecules. Cross-validation was applied by dividing the data-set into training and testing data. Training data-set was used to fit the model while the testing data-set was used to validate the model. The values of the coefficient of determination, R^2 and the standard error of the estimate (SEE) were used as the criteria to judge the performance of the QSPR model.

Chapter 7

Results and Discussion: Modeling

7.1 QSPR Results

In this study the solubility parameters of pure plasticizers were used as indicators of the plasticization efficiency to correlate the descriptors.

7.1.1 QSPR Results for commercial plasticizers

As a start, we presented the relation between the calculated solubility parameters and the value of one of the descriptor, number of carbon atoms in the leg (NCL), for the purpose of becoming familiar to the correlation of NCL to the solubility parameters. We start with NCL also because the previous study by Li et al.[8] investigated the same quantity, where NCL was defined as "leg size" in that study. Figure 7.1 showed the results and a simple linear regression model was included in this Figure.

As shown from Figure 7.1, there is an obverse trend for the correlation of NCL to the solubility parameters. With increasing number of carbons in the leg in the plasticizers, the solubility parameters decrease. This trend is more obvious for adipates. The number of carbon in the leg represents the "size" of the non-polar part of a plasticizer. Increasing plasticizer's size results in increasing the non-polar parts and decreasing solubility

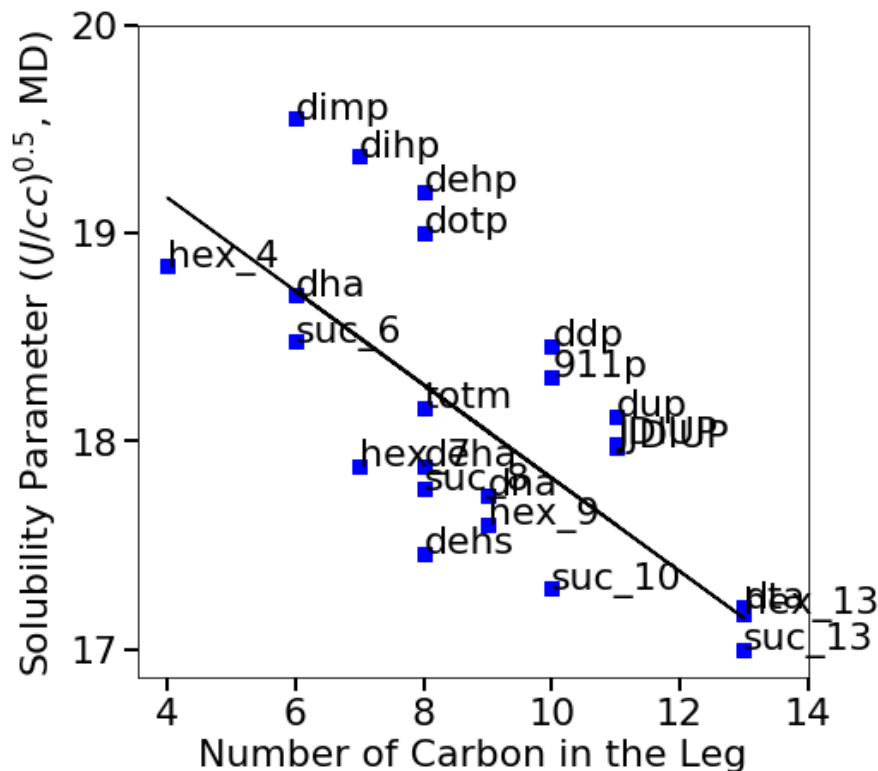


FIGURE 7.1: Solubility parameters of plasticizers vs. NCL. A linear regression line was shown in black.

parameter. Polymers with similar solubility parameters would be more easily to mixed together. A reference solubility parameter value for pure PVC is $19.35 (J/cc)^{0.5}$. This means it is harder to mix PVC with long leg chain plasticizers. It would be easier for DITP to escape from plasticized PVC than DIBP. This observation matched with previous studies by Li et al.[8] on the effects of leg size to the mobility of plasticizers. Now we validated the correlation between NCL to the solubility parameters. This descriptor was confirmed to play an important role in the plasticizer design. Meanwhile, we also want to explore the efficiency of other simple descriptors. The descriptor screenings of NCT and NCA were shown in Figure 7.2.

Clearly, the prediction results based on NCT and NCA were not as good as the linear regression model based on NCL. The trend for NCT was not clear. The data

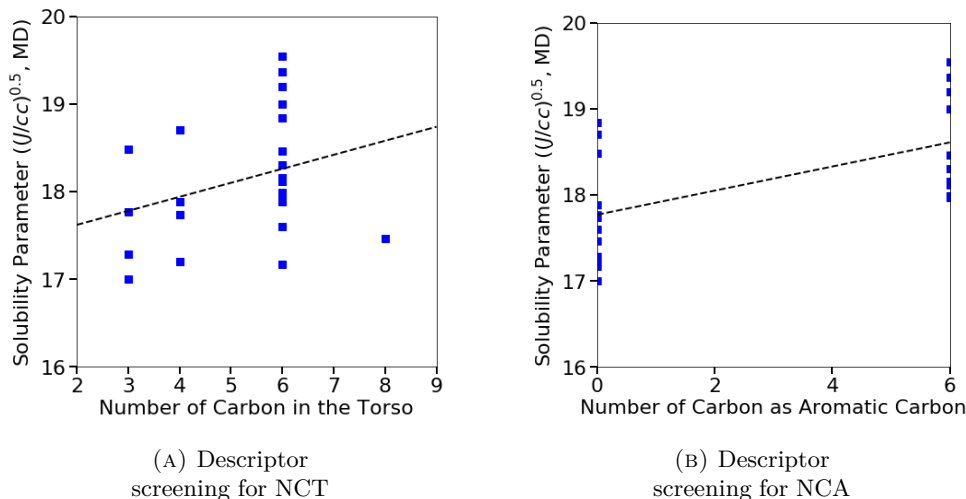


FIGURE 7.2: Solubility parameters of plasticizers vs. NCT (A) and NCA (B). Linear regression lines were shown in black.

points were scattered around the regression line with a slope close to zero. Changes in the independent variable NCT have little impact on the dependent variable, which means it was hard to tell whether NCT would affect solubility parameters. The same scenario also applied to NCA.

Obviously, a simple linear regression could not capture most of the chemical structure information of these plasticizers. For the above three descriptors, even the best linear regression model with NCL would not form a satisfactory prediction model. The standard error of the estimate (SEE) was 0.59 and the coefficient of determination, R^2 , was 0.48 for this linear regression model with NCL. The quality based on these criteria was not sufficient. Thus, we tried to combine all the three descriptors in our QSPR model. Our QSPR model would perform better with more descriptors involved simultaneously.

We use multi-linear regression model, which was our QSPR model, to correlate the first three descriptors listed in section 6.1 which were simply counting descriptors with the commercially used plasticizers. The regression equation obtained was

$$\begin{aligned} \text{Solubility Parameter } \delta &= 19.86(+/- 0.44) \\ &\quad -0.06(+/-0.06) \text{ NCT} \\ &\quad -0.21(+/-0.03) \text{ NCL} \\ &\quad +0.17(+/-0.03) \text{ NCA} \end{aligned} \tag{7.1}$$

where NCT, NCL and NCA represented the number of carbon atoms in the torso part, the number of carbon atoms in the leg part, and the number of aromatic carbon atom respectively. The standard error of the estimate (SEE) was 0.33 and the coefficient of determination, R^2 was 0.86. With these simple parameters involved in the prediction equation, the value of the R^2 implied a good prediction model.

Figure 7.3 presented the relation between the calculated and the observed solubility parameters. The usefulness of the QSPR equation could be illustrated.

The prediction results from QSPR model were on the y-axis and the results from molecular dynamic simulation were on the x-axis. All data points were close to the 45 degree line, which means a great prediction behavior and this was expected from the values of the coefficient of determination. The leg and torso size of the plasticizer molecule seemed to have negative contributions to the solubility parameters. A larger molecular size would decrease the solubility parameters, which means it was difficult for big molecules to mix with other molecules. The existence of benzene ring would also affect the solubility parameters. Since the value of NCA had a positive contribution in the equation, aromatic carbons would increase the solubility parameters compared with a cyclohexane group. Obviously, the contributions from NCL and NCA were pretty strong.

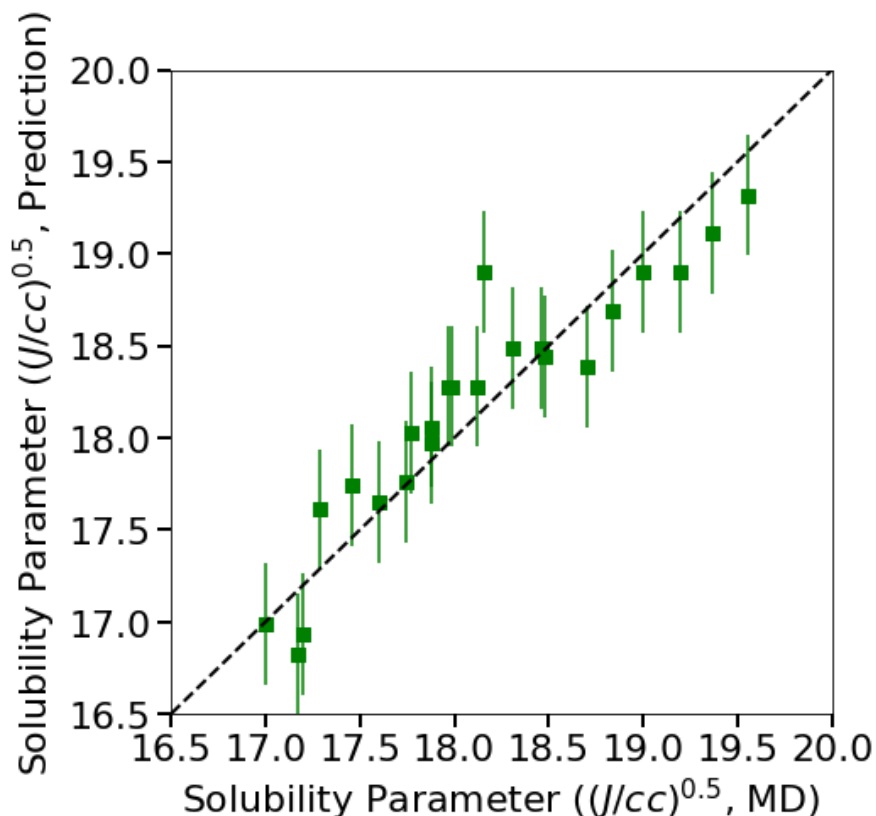


FIGURE 7.3: Plot of δ (predicted from the QSPR model) vs. δ (calculated from the MD simulation). Three simple count descriptors were included in the model.

Thus, NCL and NCA were found to be suitable descriptors when used in the current QSPR model and they should be considered importantly in the design of plasticizers.

According to the equation, one of the descriptors, NCT, had a contribution less than a threshold. The coefficient with 95% confidence interval was too close to zero. Such descriptors should be removed from our QSPR model since their contribution to the model were minimal. However, the effects of the torso group to the solubility parameters should not be ignored.

To improve this QSPR model, we would like to overcome the current naive descriptor, NCT. This descriptor could not distinguish the effects of the torso group. The number of

carbon atoms in the torso could only display the overall size of the torso group. It cannot represent the conformations of the torso. Especially for the branching information in the torso. More specific descriptors with a reflection to the local structure of molecules were needed. Therefore, two more topological indices, Wiener Index and Randic Branching Index, are introduced to our model.

To understand the effects of torso part of plasticizers, two more topological indices were added to enhance the QSPR model. The Wiener Index and Randic Branching Index were introduced in section 6.1. With the two new descriptors, another QSPR model was generated and the new regression equation obtained was

$$\begin{aligned} \text{Solubility Parameter } \delta = & 8.33(+/- 4.16) \\ & -0.21(+/-0.08) \text{ NCT} \\ & -0.22(+/-0.03) \text{ NCL} \\ & +0.17(+/-0.03) \text{ NCA} \\ & -1.79(+/-0.94) \text{ W} \\ & +14.30(+/-5.04) \text{ R} \end{aligned} \tag{7.2}$$

where W was the normalized Wiener index and R was the normalized Randic Branching index. The SEE was 0.27 and the R^2 was 0.92. The R^2 implied a good prediction model while all the descriptors provided good contribution to the prediction. Wiener Index was negatively correlated with solubility parameters while Randic Branching Index was positively correlated with solubility parameters. Figure 7.4 presented the relation between the calculated and the observed solubility parameters.

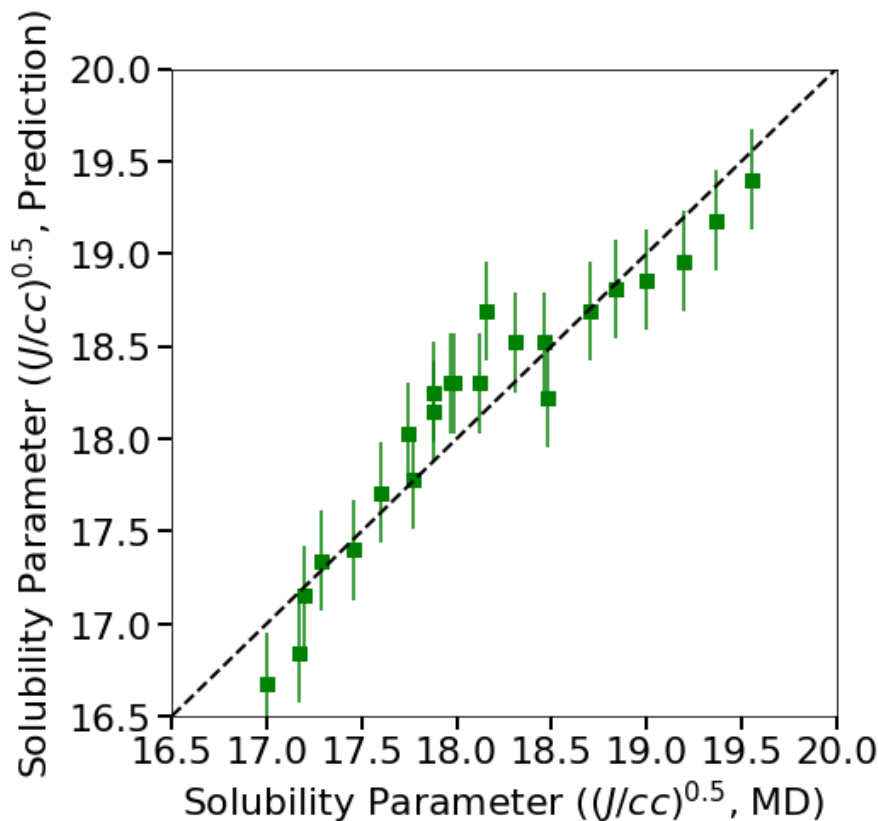


FIGURE 7.4: Plot of δ (predicted from the QSPR model) vs. δ (calculated from the MD simulation). 5 descriptors including W and R were used in the model.

The prediction results from QSPR model were on the y-axis and the results from molecular dynamic simulation were on the x-axis. All data points were also close to the 45 degree line. The higher value of R^2 implied that Wiener Index and Randic Branching Index enhanced our QSPR model a lot. These conformational descriptors contributed a good QSPR equation for plasticizer design directly relating the chemical structures of the plasticizers. The relatively high value of coefficient of W and R implied that they are the most significant descriptors in our QSPR model. Wiener index had negative contributions while Randic Branching index had positive contributions to the solubility parameters.

Another interesting finding was that the coefficient of NCT with 95% confidence interval was no longer close to zero. This finding provided an insight to the selection of descriptors. The explanation would be that the addition of Wiener Index and Randic Branching Index revealed hidden linear relationship between the variables NCT and solubility parameters, which suggested that NCT in QSPR model was still meaningful for making predictions.

The current data base for our QSPR model was relatively small and the plasticizers were split into different categories. A lot plasticizers shared the same values of Wiener Index and Randic Branching Index. For example, all phthalates shared the value of 0.71 for W and 0.97 for R. In the next section, we would extend our data size to explore the effects of Wiener Index and Randic Branching Index.

7.1.2 QSPR Results for imaginary plasticizers with Wiener Index and Randic Branching Index

To overcome the issue where the value of Wiener Index and Randic Branching Index of the plasticizers listed in Table 6.1 occupied few positions on the number axis, some “imaginary plasticizers” were introduced to our QSPR model. As mentioned in Section 6.1, these “imaginary plasticizers” were all conformed by alkane chains in their torso group with different chain length and different branch structure. Therefore, their Wiener Index and Randic Branching Index would vary in a larger number range. We added these “imaginary plasticizers” so that the contribution of Wiener Index and Randic Branching Index would be more clear in our QSPR model.

We combined the data set of the commercially used plasticizers and the “imaginary plasticizers” (in Table 6.2) to make a new QSPR model for prediction. All 5 descriptors were included and the regression equation obtained for this model was

$$\begin{aligned} \text{Solubility Parameter } \delta &= 4.52(+/- 4.80) \\ &-0.24(+/-0.07) \text{ NCT} \\ &-0.21(+/-0.06) \text{ NCL} \\ &+0.14(+/-0.06) \text{ NCA} \\ &-3.56(+/-1.31) \text{ W} \\ &+19.80(+/-5.33) \text{ R} \end{aligned} \tag{7.3}$$

where the SEE was 0.54 and the R^2 was 0.73. All the coefficients remained the same kind of contribution as in previous section. W and R were the most significant descriptors in the model.

Figure 7.5 concludes the simulation results for the commercially used plasticizers (in green) and the “imaginary plasticizers” (in orange). Solubility parameters of pure plasticizers were still chosen as the predicted property.

With the added “imaginary plasticizers”, the behavior of our QSPR model did not improve a lot. Some of the added “imaginary plasticizers were away from the 45 degree line and there were large uncertainties for the constant values in the regression equation 7.3. The large uncertainties implied that there was a possible overfitting issue in our regression model. Compared with the data-set size, our descriptors were too many. The complexity of our regression model needs to be simplified by reducing some less important features and increasing the data-set size. Another possible reason is the low solubility parameters of added “imaginary plasticizers”. The variety of Wiener Index and Randic Branching Index was extended but the prediction values were too concentrated

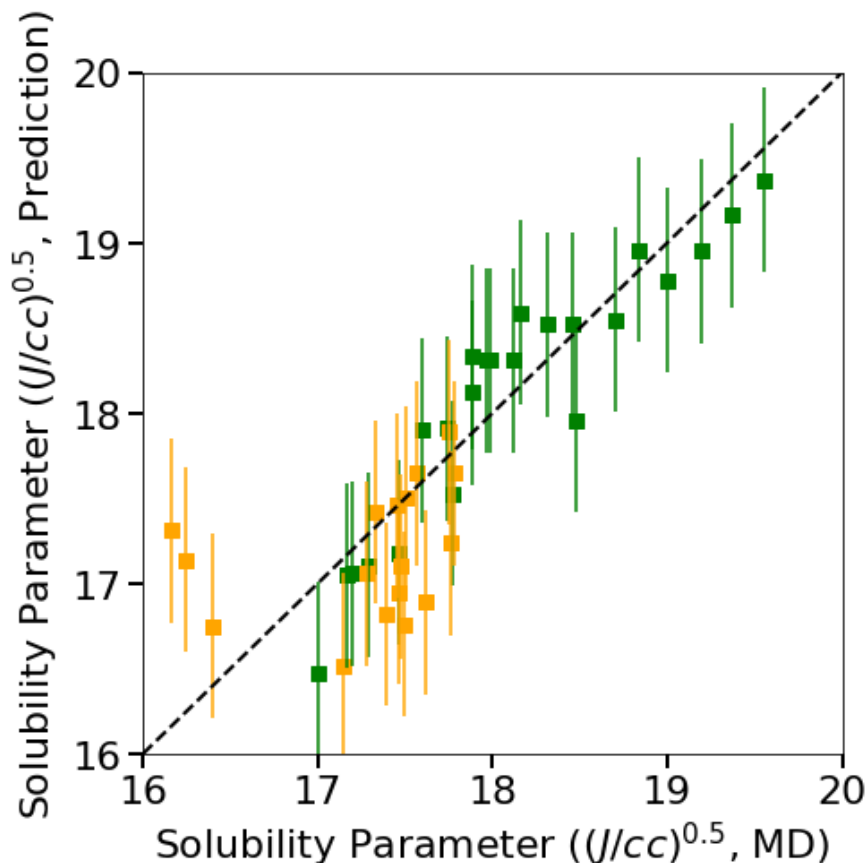


FIGURE 7.5: Plot of δ (predicted from the QSPR model) vs. δ (calculated from the MD simulation). 5 descriptors including W and R were used in the model.

in a relatively small range. The idea of add these plasticizers in the model was correct, but more such kind of plasticizers with higher solubility parameters should be added as well. Since the NCL had been proved to be a significant descriptor in our QSPR equations, the improvements could be adding more “imaginary plasticizers” with longer leg length.

We also would like to take a deeper look at the effects of Wiener Index and Randic Branching Index. Figure 7.6 concludes the simulation results for the solubility parameters of the commercially used plasticizers and “imaginary plasticizers”. The x-axis

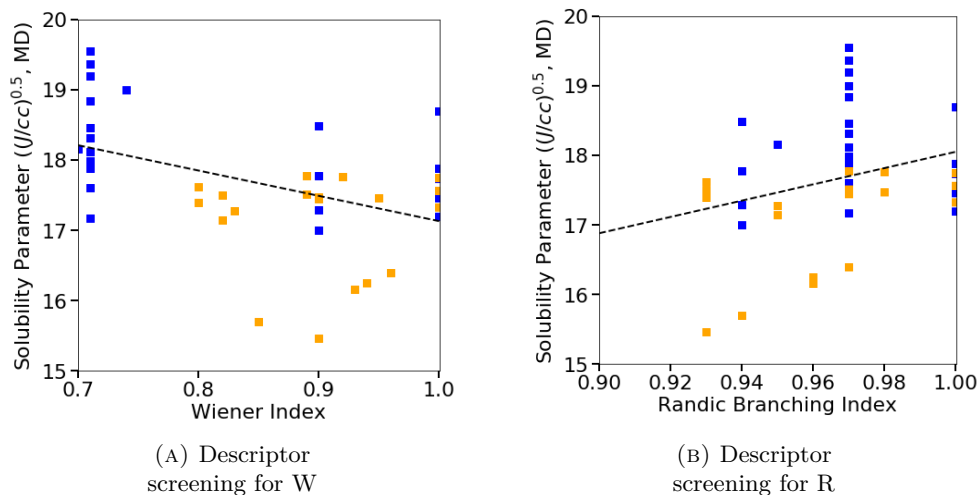


FIGURE 7.6: Solubility parameters of plasticizers vs. Wiener Index (A) and Randic Branching Index (B). Linear regression lines were shown in black. Commercially used plasticizers were in blue and imaginary plasticizers were in orange.

represents the normalized Wiener Indices and Randic Branching Indices, which were designed specifically for the torso part of plasticizers. The trends of negative contribution of Wiener Index and positive contribution of Randic Branching Index were clearer with the added “imaginary plasticizers”. Our QSPR model in Figure 7.5 correlated solubility parameters better compared to the linear regression models using single descriptors.

Meanwhile, one of the disadvantages of using Wiener Index and Randic Branching Index was concluded from these figures. The normalized Wiener Index and Randic Branching Index were designed for telling the differences of bond branching and atom’s position. A large portion of plasticizers were concentrated in one group (a thick blue vertical “line”) which were phthalates and hexamolls. These descriptors could not perfectly distinguish the differences of aromatic carbons and normal carbons. Even the significance of Wiener Index and Randic Branching Index was the highest, the other descriptors like NCA should not be deleted.

7.1.3 QSPR Results conclusion and discussion

This project provides an insight into the QSPR for the plasticizer design. Firstly we compared some basic descriptors like simple count of atoms. The results were acceptable but we were not satisfied with the simple descriptors. To understand the effects of torso part of plasticizers, some modified plasticizers were introduced for comparison. The current descriptors I was using were all formation-based of polymers. One of the advantages of using such descriptors is that there would be unnecessary to do experiments or simulations to extract information of structures. I have tested Wiener Index and Randic Branching Index in the Section 7.1.1. As a initial attempt, these topological indices were applied to the chemical structures of the torso groups in the plasticizers. The results were satisfactory. These topological indices were proved to be helpful for this project.

The QSPR approach was still limited by the size of the data base. Current data base could not cover all kinds of plasticizers. Citrates and epoxides were not included in Table 6.1. For the descriptors only relating to few plasticizers, they could not be included in our current model. Thus, the effects of these descriptors could not be represented through our QSPR model. Meanwhile, for each category of plasticizers, the number of plasticizers were small compared with a usual regression model. The data set in Table 6.1 was too concentrated in phthalates. Adding "imaginary plasticizers" was a good start in this project. Now the added "imaginary plasticizers" were also concentrated in an alkane torso structure. If more such "imaginary plasticizers" with different substructures in their torso or leg parts, the variety of our data set would improve the behavior of our QSPR model.

Another limitation of our QSPR model is the validity of the experimental data used to derive the coefficients of the QSPR correlation. Since literature values reported for solubility parameters of plasticizers were hard to find by experiments, we use the values

from molecular dynamic simulation as our reference results. One improvement that could greatly enhance the QSPR model would be the collection of new types of properties which could be driven by experiments or simulation more easily while they could also represent certain behaviors which we were interested in.

The idea of using computational tools to form QSPR equations was valuable. However, further studies are needed if people want to obtain an improved correlation model for plasticization efficiency. We must be cautious and rational to the progress of machine learning tools.

7.2 Future plans

With the construction of proper simulation protocol for atomistic molecular simulation, we tested our QSPR model for the prediction of plasticizers. There is still a large improving space for us to study. The possible improvements of our prediction model could be discussed from 3 aspects: the preparation of data set, the descriptor selection and treatment, and the mathematical prediction tools.

7.2.1 Extension of data base

As the initial step, a training data set need to be built as the benchmark for this project. Table 6.1 and 6.2 were just a start for the application of our QSPR model.

In the Part I, we already demonstrated the application of using MD simulations to generate solubility parameter results. Table 6.1 and 6.2 were all generated through this procedure, while the force field PCFF was used. The most important problem would be setting up the protocol for the simulation. This protocol should include the choice of force field, the simulation time etc. The comparison and validation of force fields showed more accurate analysis on density and heat of vaporization for COMPASS. In the future, we may change our simulation procedure based on our conclusions in Part I.

As mentioned earlier, right now the training data base was still too small. A larger training data set would always be welcomed to a prediction model. With our current small data base, there was clear advantage in the prediction by QSPR model compared to simple linear regressions. But we were not satisfied with just "better". Adding more plasticizers in our training model is necessary. In section 7.1.2, with more "imaginary plasticizers" added to our QSPR model, the prediction results did not enhance as we expected. This did not prove extending the data set was useless. On the contrary, this proved that the data set was not large enough, which added biases to our model. In the scenario where limited types of training data were added, the variety actually decreased, which could not make the prediction model better. Feeding more training data with higher variety instead of just feeding data with modified torso group of plasticizers could solve the bias issue.

Another advantage of extending the data set is that the output from the training model can deal with more complex demands. The more data we have, the more correlations can be represented in the model. Right now we have limited amount of descriptors which can be included in the QSPR equation. In the polymer design, we can only modify these descriptors. The output from our model will be limited in a small range of plasticizers. By expanding the data base where more plasticizers which have different functional groups are included, the contribution of such new functional groups would be correlated in our prediction model. Thus, more reliable prediction results would be expected.

7.2.2 New descriptors and encoder system

A challenge of this project is to find the proper descriptors for predicting properties of polymers. There are some characteristics required for these descriptors. The extraction or calculation of such descriptors do not require simulations or experiments to avoid the prediction become useless. Usually this means that the descriptors are describing

the structure or sub-structure properties instead of physicochemical properties. The descriptors do not need to be commonly applicable for all chemical molecules since we are interested in the polymer mixtures. One thing to notice is that even if some of the descriptors do not show linear correlation with the target properties, they would still be kept in the dataset in case of being used in the non-linear correlation or being used or predict other properties. Due to the non-linear model which might be used in the future, some of the descriptors found to be useless may be useful in the future.

An idea of extending the list of descriptors is from the study of graph theory and machine learning. In the machine learning process, the encoder and decoder system is commonly used, especially for image processing. One of the applications of neuron network is junction tree autoencoder system. Research on this topic might be helpful. Junction tree variational autoencoder is a method presented by Jin et al.[57]. All kinds of atoms, bonds and such components of molecules are saved in a vocabulary. In this system, the graph of original molecule is represented by two parts, the tree structure with the vocabulary information and the graph to capture connectivity. Then an encoder system will be applied. To extract the original molecule, decoder will be applied to extract the real molecule structure. Both encoder and decoder systems are based on neuron network. An application of this method would be the optimization of finding new molecules with best performance in a large pool of similar polymers.

By applying QSPR with this junction tree autoencoder system, a prediction model would be established for polymer design. The effects of functional chemical structures of plasticizers would be represented by the encoding system in this model. The variation of the code represents different chemical structures. By changing the code with just a few modifications, new polymers can be found with similar chemical structures as the base polymers since similar code means similar structures. The new polymers may have similar properties as the old ones but this new polymer is different from before and

may work better or worse for a certain target purpose. A large polymer pool where numbers of polymers with similar chemical structures and similar characteristics could be built with modifying the code. By applying optimization method, a new polymer with desired properties might be found from this large polymer pool. This approach might be valuable to future plasticizers design.

7.2.3 Numerical analysis method for prediction

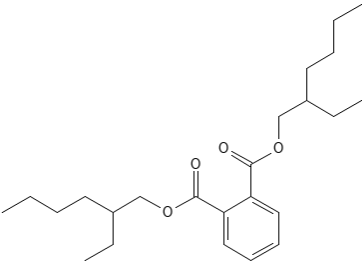
There are a few disadvantages of using QSPR for predicting behaviors of polymers. One of them is the assumption of linear correlations between descriptors and properties. Since QSPR can be considered as an advanced method for linear regression, this limitation can be overcome by introducing more advanced analysis method like neuron network. Additional computational tools might also help the development of prediction methods.

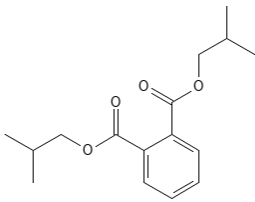
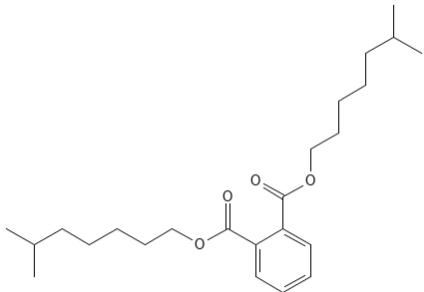
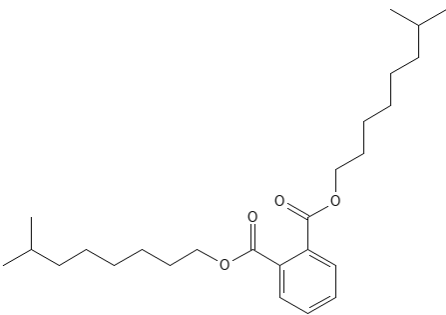
Artificial neuron network (ANN) is a tool for machine learning. The idea is based on artificial neurons[58]. It can be considered as a black box having multivariate input and multi-response output. ANN can only be used if a comparably large set of multivariate data is available which enables one to train an ANN by example. It serves best for non-linear relationship between complex inputs and outputs, which is often the case in real-world prediction problems. This flexibility allows them to capture patterns that linear models may miss. ANN has already become a popular tool for drug design nowadays[59]. Hopefully ANN could be the main improvement of this project. The method to translate from some of properties of pure polymers or polymer mixtures to other desired properties would save time and work from consuming time and resources to extract these properties.

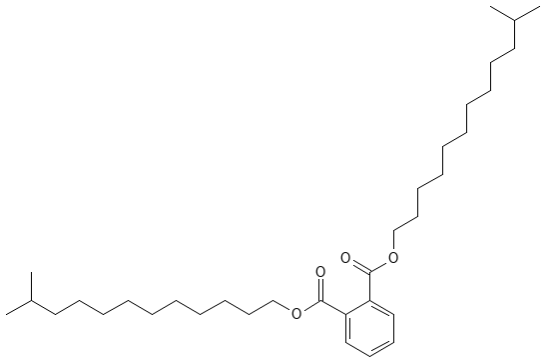
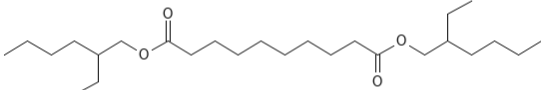
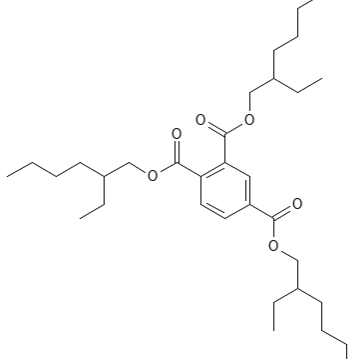
Appendix A

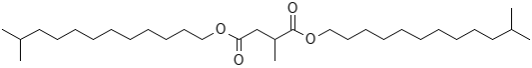
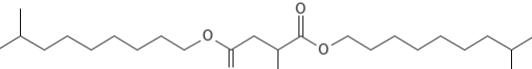
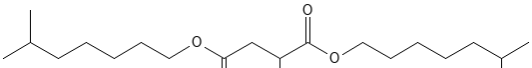
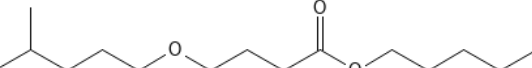
Appendix A: Chemical structures of plasticizers

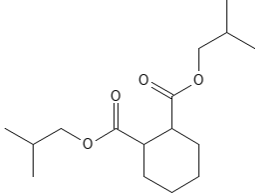
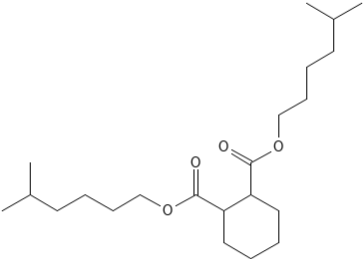
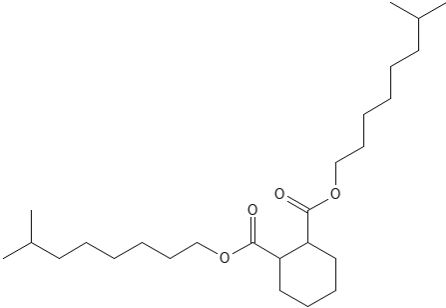
TABLE A1.1: Chemical structures of common plasticizers

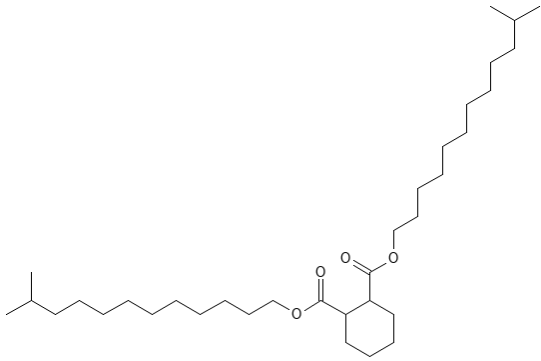
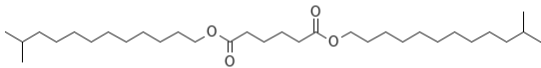
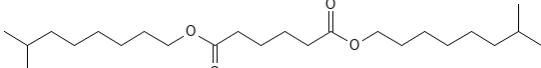
Name	Chemical structure
DEHP	 <chem>CCCCC(CC)COC(=O)c1ccccc1C(=O)OCCCC(CC)C</chem>

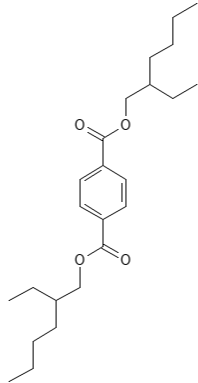
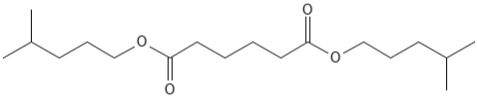
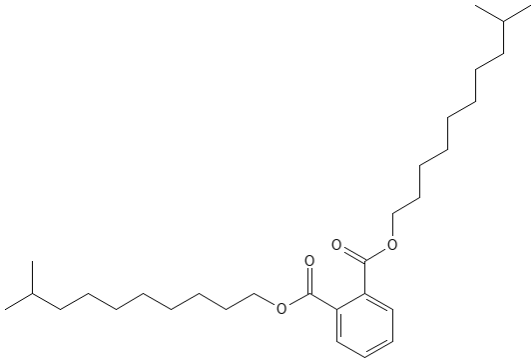
DIBP	 <p>The chemical structure of Diisobutyl phthalate (DIBP) consists of a central benzene ring with two phthalate groups attached at the 1 and 2 positions. Each phthalate group is formed by a carbonyl group (C=O) bonded to an oxygen atom, which is in turn bonded to an isobutyl chain (a three-carbon chain with a methyl branch on the second carbon).</p>
DIOP	 <p>The chemical structure of Diisooctyl phthalate (DIOP) features a central benzene ring with two phthalate groups at the 1 and 2 positions. Each phthalate group consists of a carbonyl group (C=O) bonded to an oxygen atom, which is bonded to an isooctyl chain (an eight-carbon chain with a methyl branch on the second carbon).</p>
DINP	 <p>The chemical structure of Diisododecyl phthalate (DINP) has a central benzene ring with two phthalate groups at the 1 and 2 positions. Each phthalate group is formed by a carbonyl group (C=O) bonded to an oxygen atom, which is bonded to an isododecyl chain (a twelve-carbon chain with a methyl branch on the second carbon).</p>

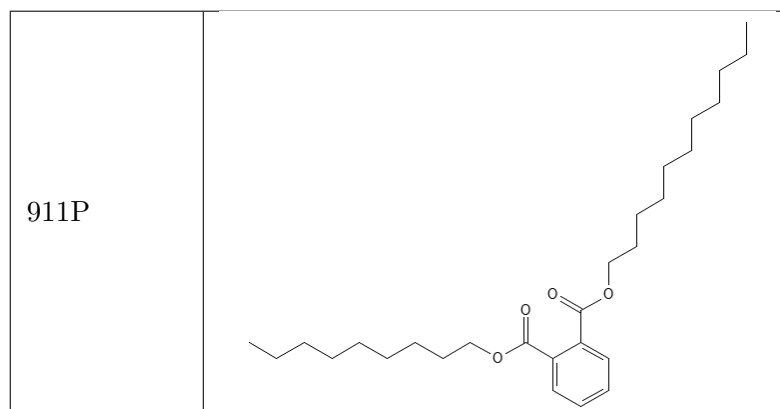
DITP	
DEHS	
TOTM	

Succinate-13	 <chem>CC(C)CCCCCCCCCCCCCOC(=O)CC(C)C(=O)OCCCCCCCCCCCCC(C)C</chem>
Succinate-10	 <chem>CC(C)CCCCCCCCCOC(=O)CC(C)C(=O)OCCCCCCCCC(C)C</chem>
Succinate-8	 <chem>CC(C)CCCCCOC(=O)CC(C)C(=O)OCCCCC(C)C</chem>
Succinate-6	 <chem>CC(C)CCCCOC(=O)CC(C)C(=O)OCCCC(C)C</chem>

Hexamoll-4	 <chem>CC(C)CCOC(=O)C1CCCCC1C(=O)OC(C)C</chem>
Hexamoll-7	 <chem>CC(C)CCCCCOC(=O)C1CCCCC1C(=O)OC(C)CCCCC</chem>
Hexamoll-9	 <chem>CC(C)CCCCCCCOC(=O)C1CCCCC1C(=O)OC(C)CCCCC</chem>

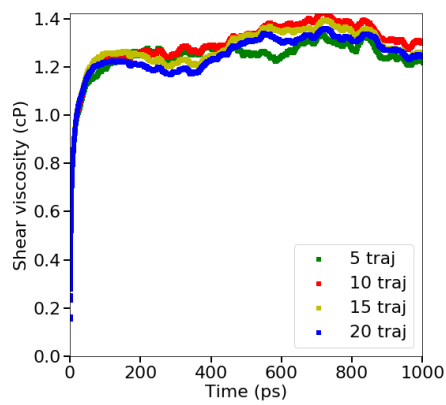
Hexamoll-13	 <p>The chemical structure of Hexamoll-13 is a diester. It features a central cyclohexane ring. One carbon of the ring is bonded to a long, branched alkyl chain (13-hydroxytridecyl) via an ester linkage (-COO-). The adjacent carbon on the ring is bonded to another long, branched alkyl chain (3,7-dimethylundecyl) via an ester linkage (-COO-).</p>
DITA	 <p>The chemical structure of DITA is a diester. It consists of a central chain of four carbon atoms with two carbonyl groups (C=O) at the second and third positions. Each carbonyl group is bonded to an oxygen atom, which is in turn bonded to a long, branched alkyl chain (3,7-dimethylundecyl).</p>
DINA	 <p>The chemical structure of DINA is a diester. It consists of a central chain of four carbon atoms with two carbonyl groups (C=O) at the second and third positions. Each carbonyl group is bonded to an oxygen atom, which is in turn bonded to a long, branched alkyl chain (3,7-dimethylundecyl).</p>

DOTP	
DIHA	
DIUP	

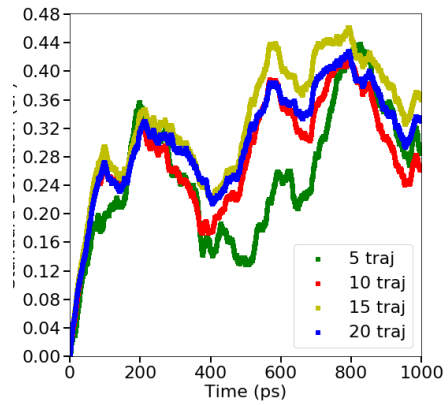


Appendix B

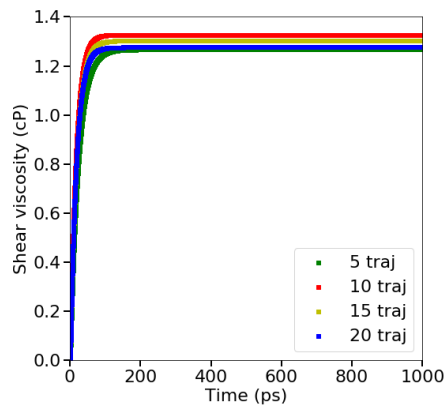
Appendix B: Raw Data and Raw Plots



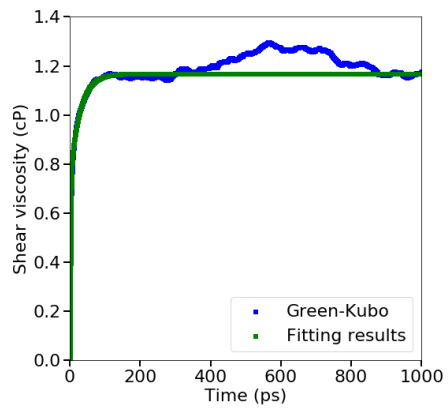
(A) Green-Kubo results from calculation



(B) Standard Deviation of viscosity from simulation



(C) Fitting results



(D) Comparison of simulation results and fitting results

FIGURE A2.1: Detailed calculation process for shear viscosity of ethanol with SciPCFF.

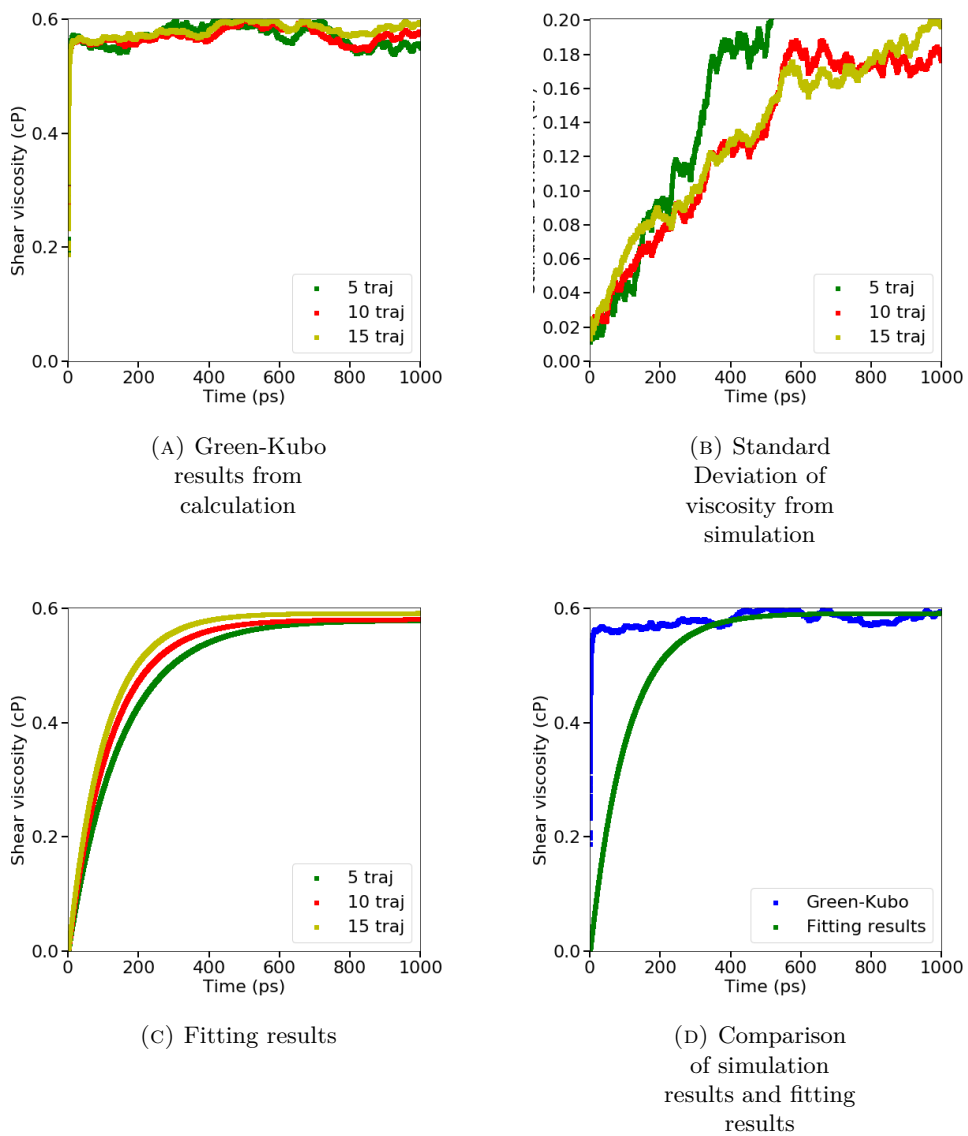
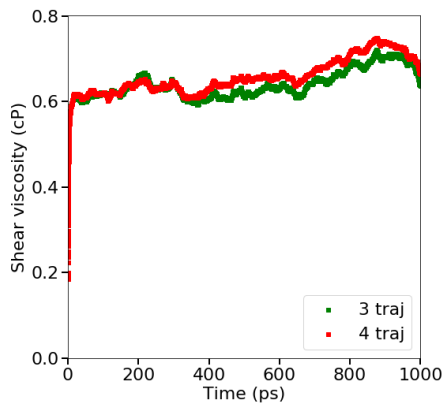
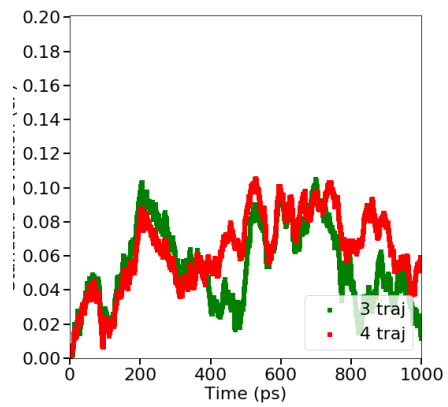


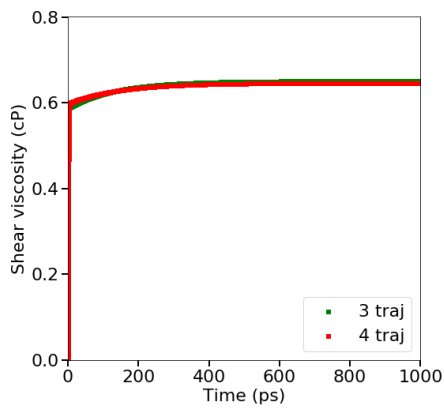
FIGURE A2.2: Detailed calculation process for shear viscosity of methyl acetate with COMPASS.



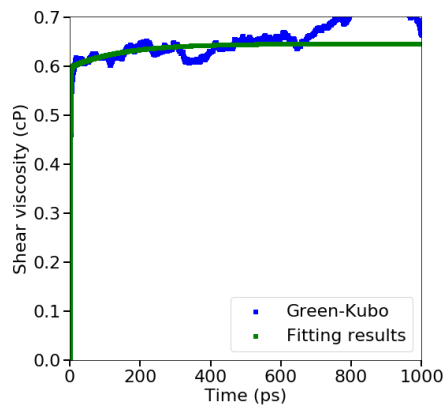
(A) Green-Kubo results from calculation



(B) Standard Deviation of viscosity from simulation

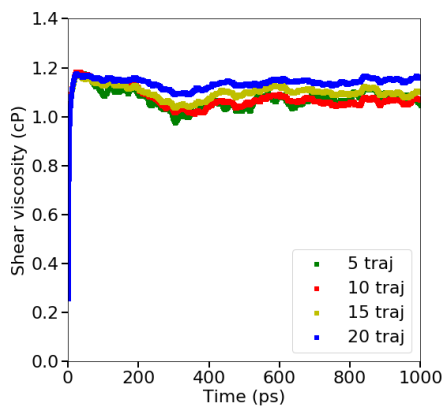


(C) Fitting results

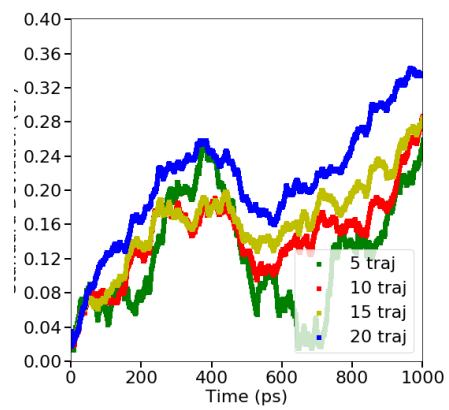


(D) Comparison of simulation results and fitting results

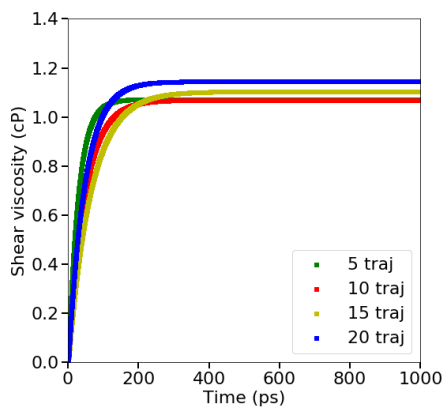
FIGURE A2.3: Detailed calculation process for shear viscosity of methyl acetate with PCFF.



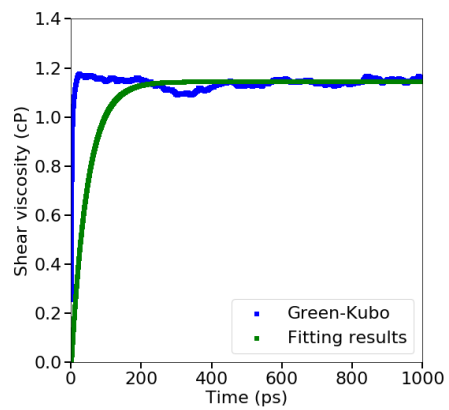
(A) Green-Kubo results from calculation



(B) Standard Deviation of viscosity from simulation

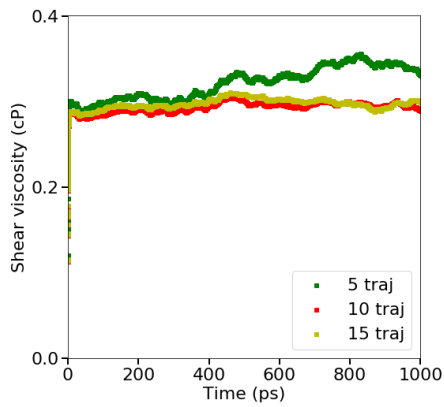


(C) Fitting results

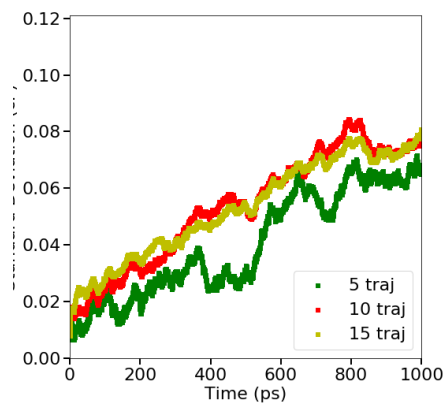


(D) Comparison of simulation results and fitting results

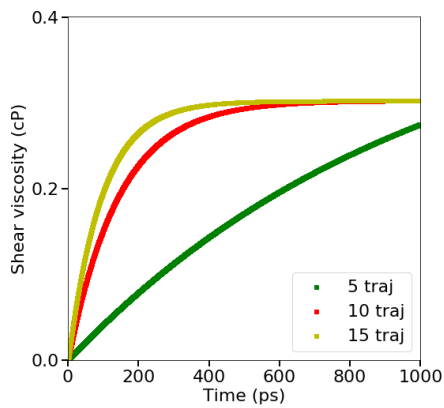
FIGURE A2.4: Detailed calculation process for shear viscosity of methyl acetate with SciPCFF.



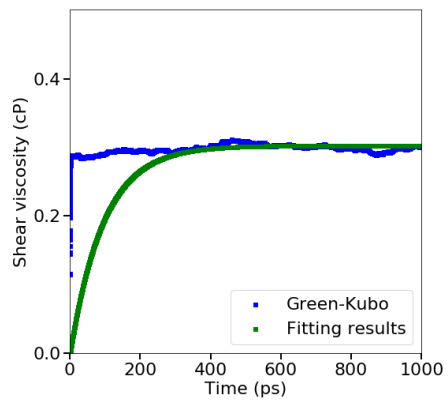
(A) Green-Kubo results from calculation



(B) Standard Deviation of viscosity from simulation

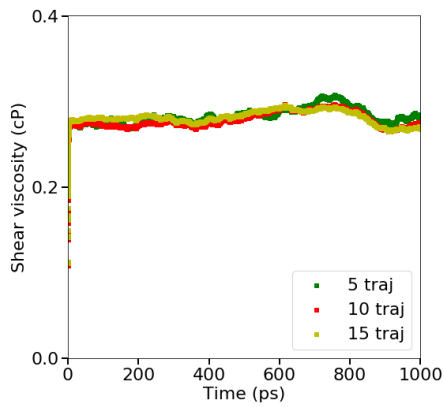


(C) Fitting results

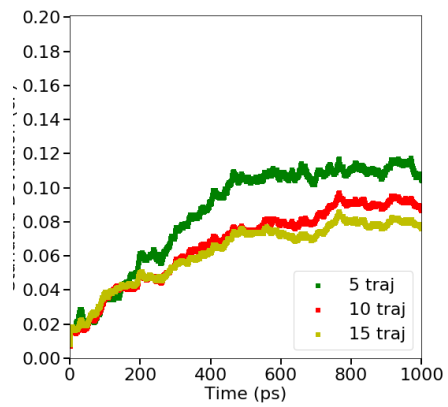


(D) Comparison of simulation results and fitting results

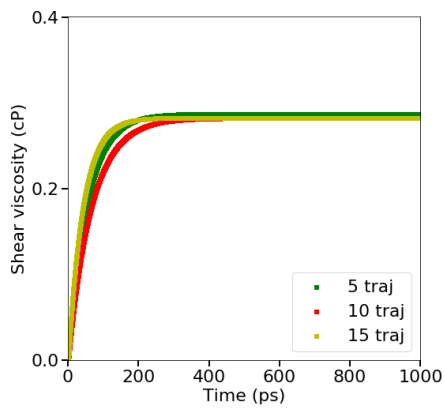
FIGURE A2.5: Detailed calculation process for shear viscosity of acetone with COMPASS.



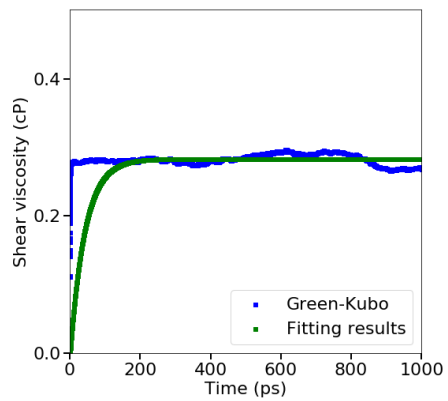
(A) Green-Kubo results from calculation



(B) Standard Deviation of viscosity from simulation

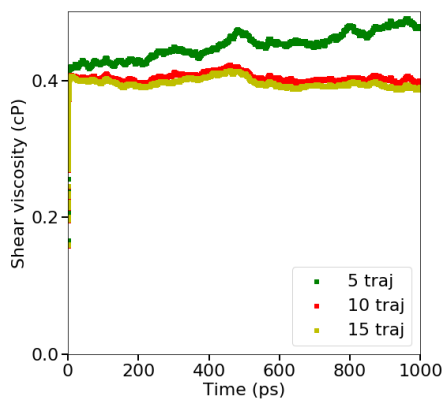


(C) Fitting results

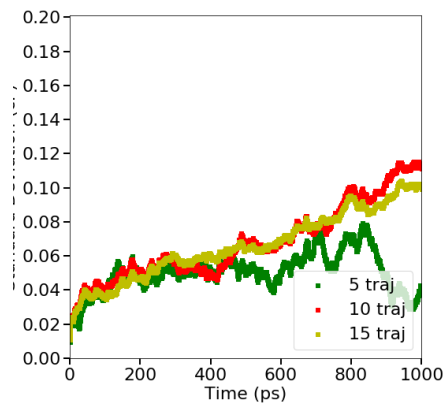


(D) Comparison of simulation results and fitting results

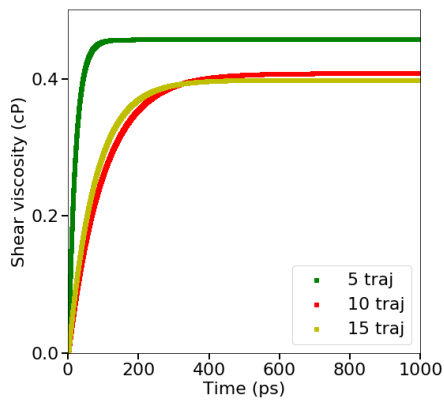
FIGURE A2.6: Detailed calculation process for shear viscosity of acetone with PCFF.



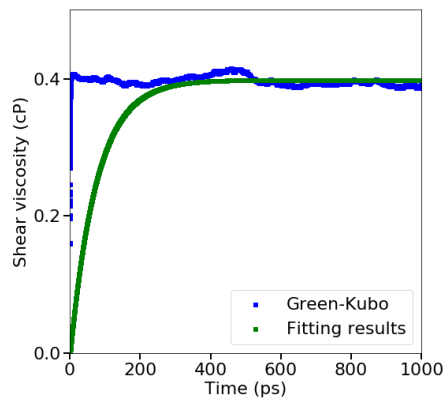
(A) Green-Kubo results from calculation



(B) Standard Deviation of viscosity from simulation

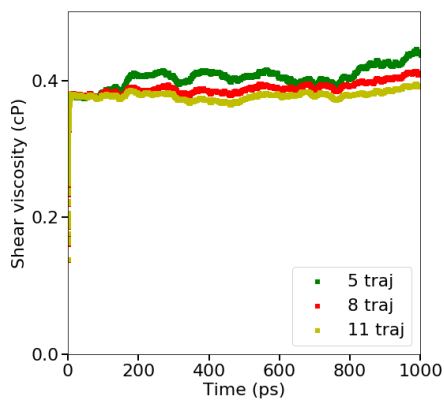


(C) Fitting results

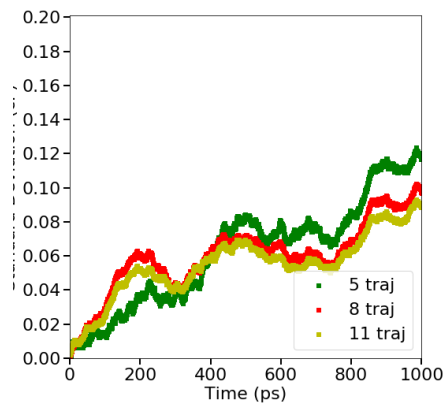


(D) Comparison of simulation results and fitting results

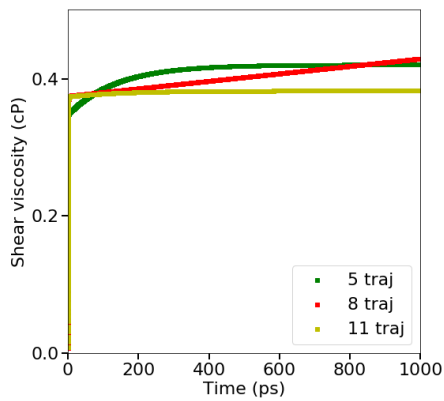
FIGURE A2.7: Detailed calculation process for shear viscosity of acetone with SciPCFF.



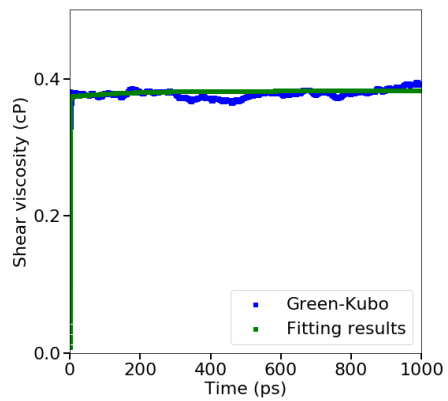
(A) Green-Kubo results from calculation



(B) Standard Deviation of viscosity from simulation

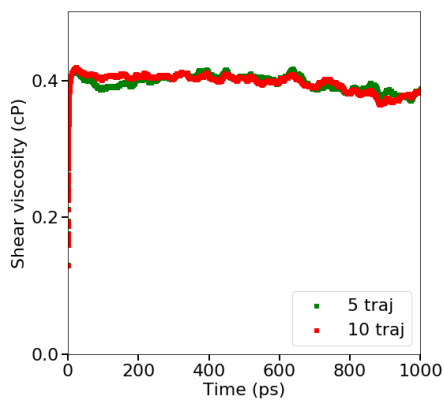


(C) Fitting results

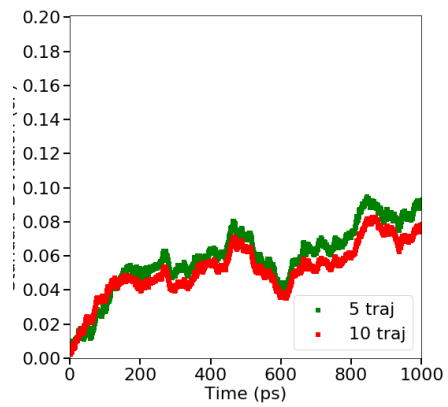


(D) Comparison of simulation results and fitting results

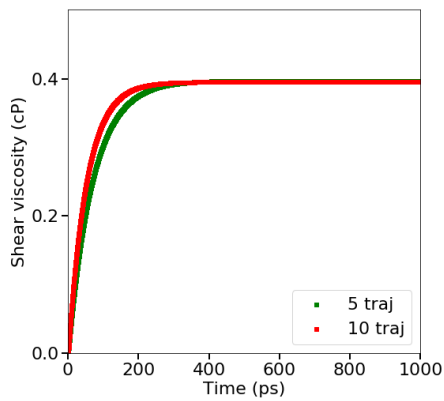
FIGURE A2.8: Detailed calculation process for shear viscosity of diethyl ether with COMPASS.



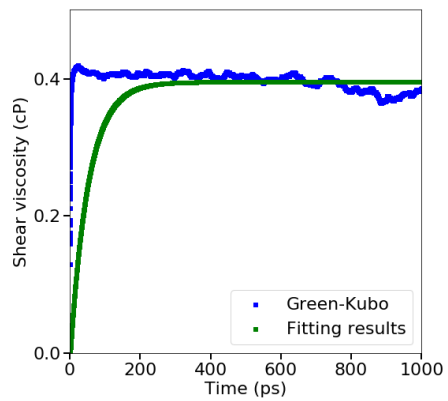
(A) Green-Kubo results from calculation



(B) Standard Deviation of viscosity from simulation

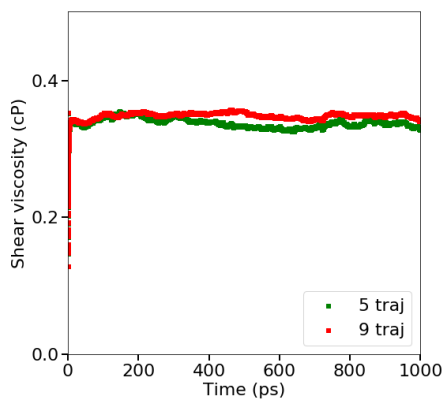


(C) Fitting results

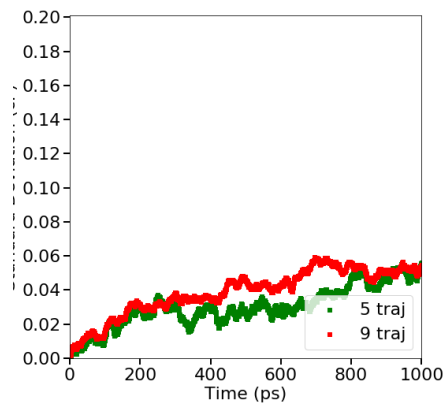


(D) Comparison of simulation results and fitting results

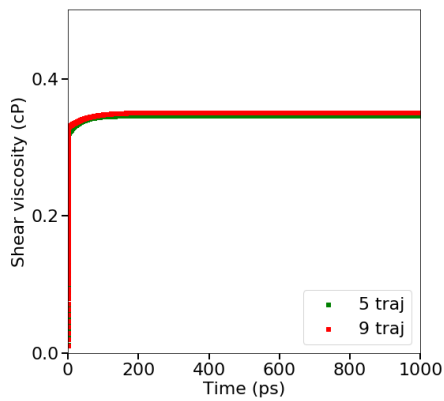
FIGURE A2.9: Detailed calculation process for shear viscosity of diethyl ether with PCFF.



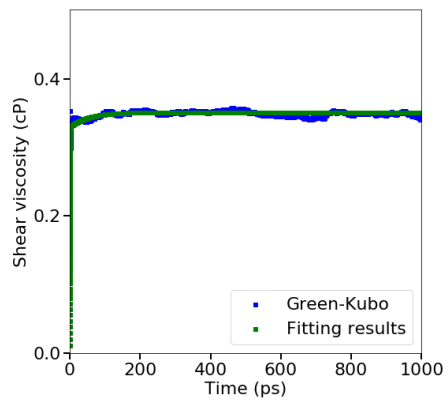
(A) Green-Kubo results from calculation



(B) Standard Deviation of viscosity from simulation



(C) Fitting results



(D) Comparison of simulation results and fitting results

FIGURE A2.10: Detailed calculation process for shear viscosity of diethyl ether with SciPCFF.

Appendix C

Appendix C: LAMMPS (31 Mar 2017) Input Scripts for Molecular Dynamics Simulation

```
# Energy minimization script

# System
units          real
atom_style     full
dimension      3
newton         on
boundary       p p p

# Styles
pair_style     lj/class2/coul/long 15.0
pair_modify    tail yes
kspace_style   ewald 1.0e-4
```



```
bond_style      class2
angle_style     class2
dihedral_style  class2
improper_style  class2
special_bonds   lj 0 0 1 coul 0 0 1

# Data file input # read initial config
variable        data index 70a4-b1-1.lmp
read_data       ${data}

# Settings
velocity         all create 300 58531416
neighbor         2.0 bin
neigh_modify     delay 0 every 1 check yes
timestep         1.0

# Output
thermo_style     custom step vol temp density press enthalpy etotal pe ke evdwl ecoul
thermo           100

# Minimization Step
min_style        sd
minimize         1.0e-3 1.0e-3 1000 100000
min_style        cg
min_modify       line quadratic
minimize         1.0e-4 1.0e-4 1000 100000
```

Data file output

write_data min.lmps

write_restart min.restart.*

Molecular dynamics NPT script init1

System&Styles

dimension 3

newton on

boundary p p p

read_restart min.restart.*

kspace_style ewald 1.0e-4

reset_timestep 0

Styles

Data file input

Settings

neighbor 2.0 bin

neigh_modify delay 0 every 1 check yes

Output

thermo_style custom step vol temp density press enthalpy etotal pe ke evdwl ecoul

thermo 10000

```
# MD step1
fix          7 all nvt temp 600 600 100.0 tchain 3 pchain 3 drag 0.0
fix          r7 all recenter 0.5 0.5 0.5 units fraction
run          100000
unfix       7
unfix       r7

write_restart nvt0.restart.*

# Data file output
write_data   nvt0.lmps

# Molecular dynamics NPT script init2

# System&Styles
dimension    3
newton       on
boundary     p p p
read_restart nvt0.restart.*
kspace_style ewald 1.0e-4
reset_timestep 0

# Styles

# Data file input
```

```
# Settings
neighbor          2.0 bin
neigh_modify      delay 0 every 1 check yes

# Output
thermo_style      custom step vol temp density press enthalpy etotal pe ke evdwl ecoul
thermo            10000

# MD step1
fix               7 all nvt temp 500 500 100.0 tchain 3 pchain 3 drag 0.0
fix               r7 all recenter 0.5 0.5 0.5 units fraction
run               5000000
unfix             7
unfix             r7

write_restart     nvt.restart.*

# Data file output
write_data        nvt.lmps

# Molecular dynamics NPT script

# System&Styles
dimension         3
newton            on
boundary          p p p
```

```
read_restart      nvt.restart.*
kspace_style      ewald 1.0e-4
reset_timestep    0

# Styles

# Data file input

# Settings
neighbor          2.0 bin
neigh_modify      delay 0 every 1 check yes

# Output
thermo_style      custom step vol temp density press enthalpy etotal pe ke evdwl ecoul
thermo            100000

# MD step1

fix               7 all npt temp 600 600 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag 0
fix               r7 all recenter 0.5 0.5 0.5 units fraction
run               5000000
unfix             7
unfix             r7

fix               c0 all npt temp 300 300 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag 0
fix               rc0 all recenter 0.5 0.5 0.5 units fraction
run               2000000
```

```
unfix          c0
unfix          rc0

fix            7 all npt temp 600 600 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag 0
fix            r7 all recenter 0.5 0.5 0.5 units fraction
run            5000000
unfix          7
unfix          r7

fix            c0 all npt temp 300 300 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag 0
fix            rc0 all recenter 0.5 0.5 0.5 units fraction
run            2000000
unfix          c0

fix            7 all npt temp 600 600 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag 0
fix            r7 all recenter 0.5 0.5 0.5 units fraction
run            5000000
unfix          7
unfix          r7

fix            c0 all npt temp 300 300 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag 0
fix            rc0 all recenter 0.5 0.5 0.5 units fraction
run            2000000
unfix          c0
unfix          rc0

fix            7 all npt temp 600 600 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag 0
```

```
fix          r7 all recenter 0.5 0.5 0.5 units fraction
run          5000000
unfix       7
unfix       r7

fix          c0 all npt temp 300 300 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag
fix          rc0 all recenter 0.5 0.5 0.5 units fraction
run          2000000
unfix       c0
unfix       rc0

fix          7 all npt temp 600 600 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag 0
fix          r7 all recenter 0.5 0.5 0.5 units fraction
run          5000000
unfix       7
unfix       r7

fix          c0 all npt temp 300 300 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag
fix          rc0 all recenter 0.5 0.5 0.5 units fraction
run          2000000
unfix       c0
unfix       rc0

fix          7 all npt temp 600 600 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag 0
fix          r7 all recenter 0.5 0.5 0.5 units fraction
run          5000000
unfix       7
```

```
unfix          r7

fix            c0 all npt temp 300 300 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag
fix            rc0 all recenter 0.5 0.5 0.5 units fraction
run            2000000
unfix          c0
unfix          rc0

write_restart annealing1npt.restart.*

# Molecular dynamics NPT script cooling

# System&Styles
dimension      3
newton         on
boundary       p p p
read_restart   annealing1npt.restart.*
kpace_style    ewald 1.0e-4
reset_timestep 0

# Styles

# Data file input

# Settings
neighbor       2.0 bin
```



```
neigh_modify      delay 0 every 1 check yes

# Output
thermo_style      custom step vol temp density press enthalpy etotal pe ke evdwl ecoul
thermo            10000

# MD step1
fix               7 all npt temp 600 600 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag 0
fix               r7 all recenter 0.5 0.5 0.5 units fraction
run               2000000
unfix             7
unfix             r7

fix               c0 all npt temp 570 570 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag 0
fix               rc0 all recenter 0.5 0.5 0.5 units fraction
run               200000
unfix             c0
unfix             rc0

fix               c0 all npt temp 540 540 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag 0
fix               rc0 all recenter 0.5 0.5 0.5 units fraction
run               200000
unfix             c0
unfix             rc0

fix               c0 all npt temp 510 510 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag 0
fix               rc0 all recenter 0.5 0.5 0.5 units fraction
run               200000
unfix             c0
```

```
unfix          rc0
fix            c0 all npt temp 480 480 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag
fix            rc0 all recenter 0.5 0.5 0.5 units fraction
run            200000
unfix          c0
unfix          rc0
fix            c0 all npt temp 450 450 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag
fix            rc0 all recenter 0.5 0.5 0.5 units fraction
run            200000
unfix          c0
unfix          rc0
fix            c0 all npt temp 420 420 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag
fix            rc0 all recenter 0.5 0.5 0.5 units fraction
run            200000
unfix          c0
unfix          rc0
fix            c0 all npt temp 390 390 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag
fix            rc0 all recenter 0.5 0.5 0.5 units fraction
run            200000
unfix          c0
unfix          rc0
fix            c0 all npt temp 360 360 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag
fix            rc0 all recenter 0.5 0.5 0.5 units fraction
run            200000
unfix          c0
unfix          rc0
fix            c0 all npt temp 330 330 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag
```

```
fix          rc0 all recenter 0.5 0.5 0.5 units fraction
run          200000
unfix       c0
unfix       rc0
fix          c0 all npt temp 300 300 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag
fix          rc0 all recenter 0.5 0.5 0.5 units fraction
run          200000
unfix       c0
unfix       rc0

fix          c0 all npt temp 300 300 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag
fix          rc0 all recenter 0.5 0.5 0.5 units fraction
run          3000000
unfix       c0
unfix       rc0

write_restart  cooling1npt.restart.*

# Molecular dynamics NPT script production runs

# System&Styles
dimension     3
newton        on
boundary      p p p
read_restart  cooling1npt.restart.*
kpace_style   ewald 1.0e-4
```

reset_timestep 0

Styles

Data file input

Settings

neighbor 2.0 bin

neigh_modify delay 0 every 1 check yes

Output

thermo_style custom step vol temp density press enthalpy etotal pe ke evdwl ecoul

thermo 10000

MD step1

fix c0 all npt temp 300 300 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag

fix rc0 all recenter 0.5 0.5 0.5 units fraction

run 500000

unfix c0

unfix rc0

write_restart production_c1_1.restart.*

fix c0 all npt temp 300 300 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag

fix rc0 all recenter 0.5 0.5 0.5 units fraction

run 500000

```
unfix          c0
unfix          rc0

write_restart  production_c1_2.restart.*

fix           c0 all npt temp 300 300 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag
fix           rc0 all recenter 0.5 0.5 0.5 units fraction
run           500000
unfix        c0
unfix        rc0

write_restart  production_c1_3.restart.*

fix           c0 all npt temp 300 300 100.0 iso 1 1 1000.0 tchain 3 pchain 3 drag
fix           rc0 all recenter 0.5 0.5 0.5 units fraction
run           500000
unfix        c0
unfix        rc0

write_restart  production_c1_4.restart.*

# Withdraw each molecules from simulation cell

# System&Styles
dimension     3
newton        on
boundary      f f f
```

```
variable max_loop equal "100"

label      loop

variable   a  loop  ${max_loop}

print     "A = $a"

if        "$a > ${max_loop}" then "jump md_c.in break"

read_restart  production_c1_8.restart.*

kspace_style  ewald 1.0e-4

reset_timestep 0

# Styles

# Data file input

variable al equal ${a}-1
variable ar equal ${a}+1

group      mole25 molecule 1:${al} ${ar}:1000

delete_atoms  group mole25 compress no bond yes mol yes

# Settings

neighbor      2.0 bin

neigh_modify  delay 0 every 1 check yes

# Output
```

```
thermo_style custom step vol temp density press enthalpy etotal pe ke evdwl ecoul elo
thermo 10000
dump 1 all custom 1 dump_${a}.nemd id mol type q xu yu zu

# Running
fix          1 all nvt temp 300.0 300.0 100.0
fix          r0 all recenter 0.5 0.5 0.5 units fraction
run          0
unfix       1
unfix       r0

# Data file output
write_data   equili_deletion_${a}.data

clear
next        a
jump        md_c.in loop
label       break
variable    a delete

#####
# SIMULATION DONE
print "All done"

# Read data from each molecule
```

```
variable max_loop equal "100"
label      loop
variable   a loop ${max_loop}
print     "A = $a"
if        "$a > ${max_loop}" then "jump cohesive_all.in break"

# System&Styles
units      real
atom_style full
dimension  3
newton     on
boundary   p p p

# Styles
pair_style lj/class2/coul/long 15.0
pair_modify tail yes
kspace_style ewald 1.0e-4
bond_style class2
angle_style class2
dihedral_style class2
improper_style class2
special_bonds lj 0 0 1 coul 0 0 1

# Data file input
read_data eq_${a}.data

reset_timestep 0
```



```
compute pe1 all pe
variable      p1 equal c_pe1

# Settings
neighbor      2.0 bin
neigh_modify  delay 0 every 1 check yes
timestep      1.0

# Output
fix  fix_print all print 1 "pe ${p1}" file pe_${a}.txt
thermo_style  custom step vol temp density press enthalpy etotal pe ke evdwl ecoul
thermo        1

# Running
run           1

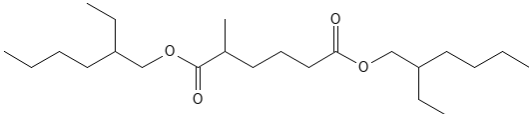
# Data file output
clear
next          a
jump          cohesive_all.in loop
label         break
variable      a delete

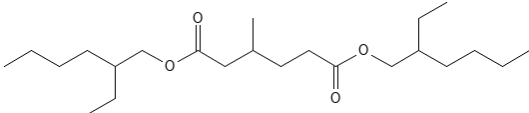
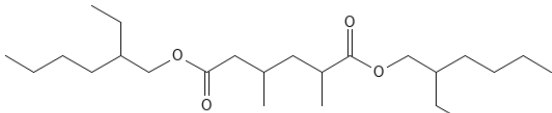
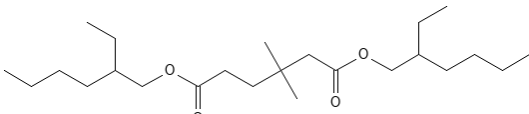
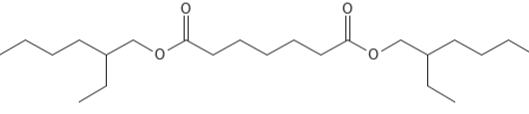
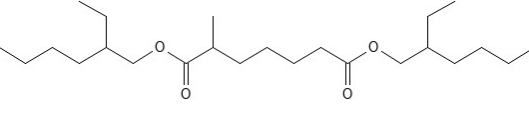
#####
# SIMULATION DONE
print "All done"
```

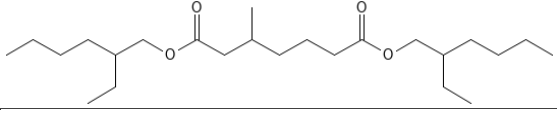
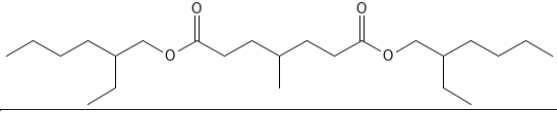
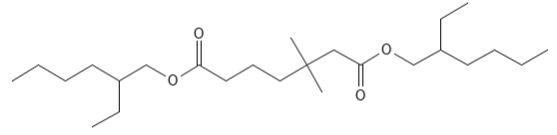
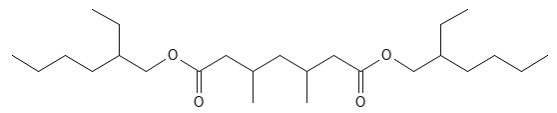
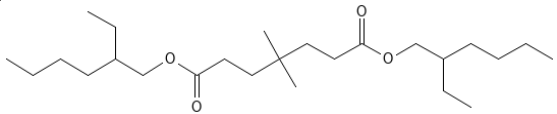
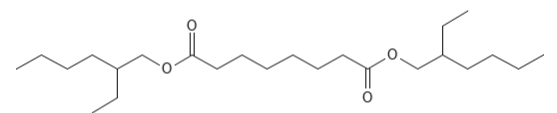
Appendix D

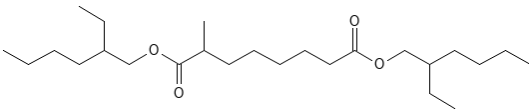
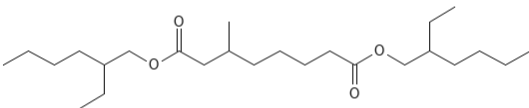
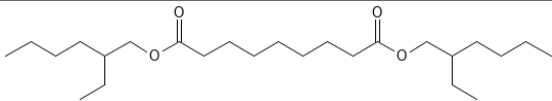
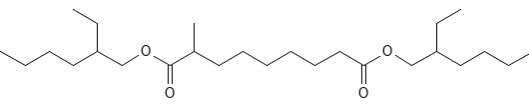
Appendix D: Chemical structures of imaginary plasticizers

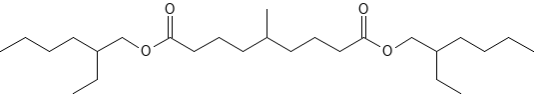
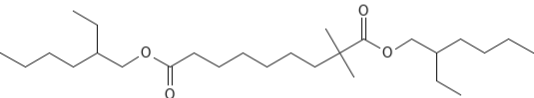
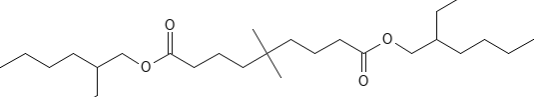
TABLE A4.1: Chemical structures of imaginary plasticizers

Name	Chemical structure
a4-b1-1	

a4-b1-2	
a4-b2-13	
a4-b2-22	
a5	
a5-b1-1	

a5-b1-2	
a5-b1-3	
a5-b2-22	
a5-b2-24	
a5-b2-33	
a6	

a6-b1-1	 <p>The structure shows a central decanedioate chain (10 carbons) with two ester groups. Each ester group is connected to a 2-ethylhexyl group (6 carbons in the main chain, 2 carbons in the ethyl branch).</p>
a6-b1-2	 <p>The structure is identical to a6-b1-1, but the entire molecule is rotated 180 degrees.</p>
a7	 <p>The structure is identical to a6-b1-1 and a6-b1-2, showing a central decanedioate chain with two 2-ethylhexyl groups.</p>
a7-b1-1	 <p>The structure is identical to a6-b1-1, showing a central decanedioate chain with two 2-ethylhexyl groups.</p>

a7-b1-4	
a7-b2-11	
a7-b2-44	

Bibliography

- [1] W. V. Titow. Plasticisers. In: *PVC Plastics*. Springer Netherlands, 1990, 177–257.
- [2] M. Rahman and C. S. Brazel. The plasticizer market: an assessment of traditional plasticizers and research trends to meet new challenges. *Progress in Polymer Science* 29(12) (2004), 1223–1248.
- [3] A. T. Demir and S. Ulutan. Migration of phthalate and non-phthalate plasticizers out of plasticized PVC films into air. *Journal of Applied Polymer Science* 128(3) (2013), 1948–1961.
- [4] L. G. Krauskopf. How about alternatives to phthalate plasticizers? *Journal of Vinyl and Additive Technology* 9(4) (Dec. 2003), 159–171.
- [5] T. Mekonnen, P. Mussone, H. Khalil, and D. Bressler. Progress in bio-based plastics and plasticizing modifications. *Journal of Materials Chemistry A* 1(43) (2013), 13379.
- [6] F. Chiellini, M. Ferri, A. Morelli, L. Dipaola, and G. Latini. Perspectives on alternatives to phthalate plasticized poly (vinyl chloride) in medical devices applications. *Progress in Polymer Science* 38(7) (2013), 1067–1088.
- [7] H. C. Erythropel, S. Shipley, A. Börmann, J. A. Nicell, M. Maric, and R. L. Leask. Designing green plasticizers: Influence of molecule geometry and alkyl chain length on the plasticizing effectiveness of diester plasticizers in PVC blends. *Polymer* 89 (2016), 18–27.

- [8] D. Li, K. Panchal, R. Mafi, and L. Xi. An Atomistic Evaluation of the Compatibility and Plasticization Efficacy of Phthalates in Poly(vinyl chloride). *Macromolecules* 51(18) (Aug. 2018), 6997–7012.
- [9] M. Griebel, S. Knapek, and G. Zumbusch. *Numerical Simulation in Molecular Dynamics*. Heidelberg Germany: Springer Verlag, 2007.
- [10] Z. Luo and J. Jiang. Molecular dynamics and dissipative particle dynamics simulations for the miscibility of poly(ethylene oxide)/poly(vinyl chloride) blends. *Polymer* 51(1) (Jan. 2010), 291–299.
- [11] S. Zhang and L. Xi. Effects of precursor topology on polymer networks simulated with molecular dynamics. *Polymer* 116 (May 2017), 143–152.
- [12] M. Belmares, M. Blanco, W. A. Goddard, R. B. Ross, G. Caldwell, S. Chou, J. Pham, P. M. Olofson, and C. Thomas. Hildebrand and Hansen solubility parameters from Molecular Dynamics with applications to electronic nose polymer sensors. *Journal of Computational Chemistry* 25(15) (2004), 1814–1826.
- [13] M. Chandola and S. Marathe. A QSPR for the plasticization efficiency of polyvinylchloride plasticizers. *Journal of Molecular Graphics and Modelling* 26(5) (Jan. 2008), 824–828.
- [14] M. González. Force fields and molecular dynamics simulations. *École thématique de la Société Française de la Neutronique* 12 (2011), 169–200.
- [15] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* 21 (1953), 1087–1092.
- [16] B. Alder and T. Wainwright. Phase transition for a hard sphere system. *The Journal of Chemical Physics* 27 (1957), 1208–1209.

Bibliography

- [17] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in 't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, and S. J. Plimpton. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications* 271 (Feb. 2022), 108171.
- [18] R. M. Martin. *Electronic Structure: Basic Theory and Practical Methods*. Cambridge University Press, 2004.
- [19] A. Gavezzotti. *Molecular Aggregation: Structure analysis and molecular simulation of crystals and liquids*. Oxford University Press, 2007.
- [20] N. L. Allinger, K. Chen, and J.-H. Lii. An improved force field (MM4) for saturated hydrocarbons. *Journal of Computational Chemistry* 17(5-6) (Apr. 1996), 642–668.
- [21] H. Sun. COMPASS: An ab Initio Force-Field Optimized for Condensed-Phase Applications Overview with Details on Alkane and Benzene Compounds. *The Journal of Physical Chemistry B* 102(38) (Aug. 1998), 7338–7364.
- [22] J. R. Maple, M.-J. Hwang, T. P. Stockfish, U. Dinur, M. Waldman, C. S. Ewig, and A. T. Hagler. Derivation of class II force fields. I. Methodology and quantum force field for the alkyl functional group and alkane molecules. *Journal of Computational Chemistry* 15(2) (Feb. 1994), 162–182.
- [23] H. Sun. Force field for computation of conformational energies, structures, and vibrational frequencies of aromatic polyesters. *Journal of Computational Chemistry* 15(7) (July 1994), 752–768.
- [24] M. J. Hwang, T. P. Stockfish, and A. T. Hagler. Derivation of Class II Force Fields. 2. Derivation and Characterization of a Class II Force Field, CFF93, for the Alkyl Functional Group and Alkane Molecules. *Journal of the American Chemical Society* 116(6) (Mar. 1994), 2515–2525.

Bibliography

- [25] J. Maple, U. Dinur, and A. Hagler. New Approaches to Empirical Force Fields. *Proceedings of the National Academy of Sciences* 85 (1988), 5350.
- [26] D. Rigby, H. Sun, and B. E. Eichinger. Computer simulations of poly(ethylene oxide): force field, pvt diagram and cyclization behaviour. *Polymer International* 44(3) (Nov. 1997), 311–330.
- [27] B. Eichinger, D. Rigby, and J. Stein. Cohesive properties of Ultem and related molecules from simulations. *Polymer* 43(2) (Jan. 2002), 599–607.
- [28] T. A. Halgren. The representation of van der Waals (vdW) interactions in molecular mechanics force fields: potential form, combination rules, and vdW parameters. *Journal of the American Chemical Society* 114(20) (Sept. 1992), 7827–7843.
- [29] M. Waldman and A. Hagler. New combining rules for rare gas van der waals parameters. *Journal of Computational Chemistry* 14(9) (Sept. 1993), 1077–1084.
- [30] X. P. Chen, C. A. Yuan, C. K. Wong, S. W. Koh, and G. Q. Zhang. Validation of forcefields in predicting the physical and thermophysical properties of emeraldine base polyaniline. *Molecular Simulation* 37(12) (Oct. 2011), 990–996.
- [31] P. Prathumrat, I. Sbarski, E. Hajizadeh, and M. Nikzad. A comparative study of force fields for predicting shape memory properties of liquid crystalline elastomers using molecular dynamic simulations. *Journal of Applied Physics* 129(15) (Apr. 2021), 155101.
- [32] D. Hofmann, M. Entrialgo-Castano, A. Lerbret, M. Heuchel, and Y. Yampolskii. Molecular Modeling Investigation of Free Volume Distributions in Stiff Chain Polymers with Conventional and Ultrahigh Free Volume: Comparison between Molecular Modeling and Positron Lifetime Studies. *Macromolecules* 36(22) (Sept. 2003), 8528–8538.
- [33] N. C. Karayiannis, V. G. Mavrantzas, and D. N. Theodorou. Detailed Atomistic Simulation of the Segmental Dynamics and Barrier Properties of Amorphous

- Poly(ethylene terephthalate) and Poly(ethylene isophthalate). *Macromolecules* 37(8) (Mar. 2004), 2978–2995.
- [34] L. Martinez, R. Andrade, E. G. Birgin, and J. M. Martinez. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *Journal of Computational Chemistry* 30(13) (Oct. 2009), 2157–2164.
- [35] P. Ewald. The calculation of optical and electrostatic grid potential. *Annalen der Physik* 64 (1921), 253–287.
- [36] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. 2nd. Oxford: Oxford University Press, 1989.
- [37] W. Shinoda, M. Shiga, and M. Mikami. Rapid estimation of elastic constants by molecular dynamics simulation under constant stress. *Physical Review B* 69(13) (Apr. 2004).
- [38] B. Hess. Determining the shear viscosity of model liquids from molecular dynamics simulations. *The Journal of Chemical Physics* 116(1) (2002), 209.
- [39] Y. Zhang, A. Otani, and E. J. Maginn. Reliable Viscosity Calculation from Equilibrium Molecular Dynamics Simulations: A Time Decomposition Method. *Journal of Chemical Theory and Computation* 11(8) (July 2015), 3537–3546.
- [40] C. Rey-Castro and L. F. Vega. Transport Properties of the Ionic Liquid 1-Ethyl-3-Methylimidazolium Chloride from Equilibrium Molecular Dynamics Simulation. The Effect of Temperature. *The Journal of Physical Chemistry B* 110(29) (July 2006), 14426–14435.
- [41] C. Wohlfarth. *Viscosity of methyl acetate: Datasheet from Landolt-Börnstein - Group IV Physical Chemistry · Volume 25: “Supplement to IV/18” in SpringerMaterials* (https://doi.org/10.1007/978-3-540-75486-2_69). Ed. by M. Lechner. Part of SpringerMaterials.

- [42] J. Gmehling, ed. *Acetone Dynamic Viscosity: Datasheet from "Dortmund Data Bank (DDB) – Thermophysical Properties Edition 2014" in SpringerMaterials* (https://materials.springer.com/thermophysical/docs/vis_c4). Part of SpringerMaterials.
- [43] C. Wohlfarth. *Viscosity of diethyl ether: Datasheet from Physical Chemistry · Volume 29: "Viscosity of Pure Organic Liquids and Binary Liquid Mixtures" in SpringerMaterials* (https://doi.org/10.1007/978-3-662-49218-5_114). Ed. by M. D. Lechner. Part of SpringerMaterials.
- [44] V. Kiryakov, I. Usyukin, and V. Shleinikov. Die Dichte des Dampfes von Aceton bei niedrigen Temperaturen. *Zhurnal Vsesoyuznogo Khimicheskogo Obshchestva imeni D.I. Mendeleeva* 11 (1966), 710–711.
- [45] R. Malhotra and L. Woolf. Thermodynamic properties of propanone (acetone) at temperatures from 278 K to 323 K and pressures up to 400 MPa. *The Journal of Chemical Thermodynamics* 23(9) (Sept. 1991), 867–876.
- [46] A. Kumagai and H. Iwasaki. Pressure volume temperature relationships of several polar liquids. *Journal of Chemical Engineering Data* 23(3) (July 1978), 193–195.
- [47] S. Oswal, P. Oswal, R. Gardas, S. Patel, and R. Shinde. Acoustic, volumetric, compressibility and refractivity properties and reduction parameters for the ERAS and Flory models of some homologous series of amines from 298.15 to 328.15 K. *Fluid Phase Equilibria* 216(1) (Feb. 2004), 33–45.
- [48] R. E. Kirk, D. F. Othmer, and C. A. Mann. Encyclopedia of Chemical Technology. Vol. II. *The Journal of Physical and Colloid Chemistry* 53(4) (Apr. 1949), 591–591.
- [49] A. MARCILLA and J. GARCÍA. RHEOLOGICAL STUDY OF PVC PLASTISOLS DURING GELATION AND FUSION. *European Polymer Journal* 33(3) (Mar. 1997), 349–355.

Bibliography

- [50] P. A. Small. Some factors affecting the solubility of polymers. *Journal of Applied Chemistry* 3(2) (May 2007), 71–80.
- [51] K. Roy, S. Kar, and R. Das. *A Primer on QSAR/QSPR Modeling*. Berlin/Heidelberg, Germany: Springer International Publishing, 2015.
- [52] J. Gasteiger. *Handbook of Chemoinformatics*. Weinheim, Germany: John Wiley & Sons, Inc., 2008.
- [53] P. J. Hansen and P. C. Jurs. Chemical applications of graph theory. Part I. Fundamentals and topological indices. *Journal of Chemical Education* 65(7) (July 1988), 574.
- [54] C. C. Cypcar, P. Camelio, V. Lazzeri, L. J. Mathias, and B. Waegell. Prediction of the Glass Transition Temperature of Multicyclic and Bulky Substituted Acrylate and Methacrylate Polymers Using the Energy, Volume, Mass (EVM) QSPR Model. *Macromolecules* 29(27) (Jan. 1996), 8954–8959.
- [55] H. Wiener. Structural Determination of Paraffin Boiling Points. *Journal of the American Chemical Society* 69 (1947), 17–20.
- [56] M. Randic. Characterization of molecular branching. *Journal of the American Chemical Society* 97(23) (Nov. 1975), 6609–6615.
- [57] W. Jin, R. Barzilay, and T. Jaakkola. Junction Tree Variational Autoencoder for Molecular Graph Generation. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, 2323–2332.
- [58] A. R. Leach and V. J. Gillet. *An Introduction to Chemoinformatics*. Netherlands: Springer, 2007, 1–255.
- [59] B. Begam and J. S. Kumar. A Study on Cheminformatics and its Applications on Modern Drug Discovery. *Procedia Engineering* 38 (2012), 1264–1275.