Development of Transferable Coarse-Grained Models of Amino Acids

Olivia Kristine Conway

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Science
In
Chemical Engineering

Sanket A. Deshmukh
Luke E. K. Achenie
John B. Matson

August 27, 2019
Blacksburg, VA

Development of Transferable Coarse-Grained Models of Amino Acids

Olivia Kristine Conway

ABSTRACT

There are twenty standard amino acids that are the structural units of biomolecules and biomaterials such as proteins and peptide amphiphiles (PAs). The focus of this study was to develop accurate transferable coarse-grained (CG) models of those amino acids. In CG models, several atoms are represented together as a single pseudo-atom or "bead," which can allow the modeling of processes like self-assembly of biomolecules and biomaterials through reduction of degrees of freedom and corresponding increased computational speed. A 2:1 to 4:1 mapping scheme, in which a CG bead is comprised of two to four heavy atoms, respectively, and associated hydrogens, has been employed to represent functional groups in the amino acids. The amino acid backbone atoms are modeled as two beads while the side chains are modeled with one to three beads, and each terminus is modeled as one bead. The bonded parameters for the CG models were obtained from bond, angle, and dihedral distributions from all-atom molecular dynamics (MD) simulations of dipeptides. Non-bonded parameters were optimized using the particle swarm optimization (PSO) method to reproduce experimental properties (heat of vaporization, surface tension, and density) of analogues of the side chains, termini, and backbone groups of the amino acids. These CG models were used to study the self-assembly pathways and mechanisms of the PA c16-AHL$_3$K$_3$-CO$_2$H in the presence of explicit CG water.

Development of Transferable Coarse-Grained Models of Amino Acids

Olivia Kristine Conway

GENERAL AUDIENCE ABSTRACT

In this study, models of the amino acids were developed for computer simulations. In these models, the amino acids are represented as a collection of two or more "beads" bonded together rather than as a collection of atoms bonded together. The beads were created in such a way that their characteristics reflect those of the molecules and atom groups that they represent. This was accomplished in part by selecting parameters for each bead that approximately reproduce experimental properties (density, heat of vaporization, and surface tension) and structures (bonds and angles) of the molecules and atom groups of which they are representative. Amino acids can link together to form short segments, known as peptides, or longer chains that form proteins. The bead models that were developed in this study can be linked together in the same way. They can also be linked with other beads that represent other atom groups—carbon groups of a carbon chain, for example. Certain types of molecules known as peptide amphiphiles (PAs) are often composed of amino acids and a carbon chain. The amino acid bead models were created especially to study these molecules, so once the models were developed, they were used in computer simulations to represent PAs. Many types of PAs can automatically assemble into structures that resemble fibers, and it is this behavior in particular that was studied. By using these models in computer simulations, we are able to see things that cannot be seen in a lab with a microscope or other lab tools. This may help with future efforts to study and design molecules such as PAs which show promise for medical applications like drug delivery.

Acknowledgements

I would like to thank the members of the Department of Chemical Engineering at Virginia Tech who specifically contributed to making my time as a graduate student in the department go more smoothly, especially Diane Cannaday (Graduate Coordinator) for all her help and support, Tina Russell (Communications/Program Support), and Dr. David Cox (Professor and Department Head). Their friendliness and willingness to offer assistance and advice were continuous all throughout my time at VT, and for that I am grateful.

I would also like to thank the members of my group for their help and contributions: Dr. Sanket Deshmukh, my advisor, for remaining positive even when the research wasn't developing as desired; Dr. Karteek Bejagam, who provided so much useful help and insight regarding the research and research tools during his time as a post-doctoral researcher in the lab; and Yaxin An, fellow graduate student, who also provided help especially in formulating the research strategy. Thanks also to committee members Dr. John Matson and Dr. Luke Achenie for the parts they played in bringing the research to completion.

This work would never have been possible without the support of my family throughout my entire educational experience. I am forever grateful for them. There were also many individuals in the Blacksburg area who surrounded me with friendship and support while I lived there and made my Virginia experience so much better for it. My thanks to them all.

Table of Contents

1. Introduction

There are twenty common amino acids which are the fundamental structural units of proteins and peptides—molecules that can have important biological functions such as catalyzing biochemical reactions (enzymes), or serving as signaling molecules, transport molecules, or structural molecules, for example. The chemical diversity of the amino acids as "building blocks" in combination with an ability to form secondary structures like pleats (β-pleats) and helical twists (α-helices) enables the multiplicity of structures and functions that is necessary for the wider spectrum of forms and operations characteristic of biological systems. Amino acids can also be chemically altered, through synthetic additions for example, for an even wider array of features.

A class of molecules based on amino acids that in recent years has been the subject of much investigation is peptide amphiphiles (PAs). Combining the chemical diversity of the amino acids with a hydrophobic group, such as an alkyl chain, PAs possess the capacity to self-assemble into structures like micelles, sheets, and fibers.[1] The peptide portion provides potential for biocompatibility and biodegradation, and also bioactivity if created to possess sequence similarity to proteins with biological significance. The amphiphilicity also enables the possibility of encapsulating small hydrophobic drugs.[2] For these reasons, studies involving PAs have predominantly focused on biomedical applications (e.g. PAs as drug delivery agents and biomaterials for regenerative medicine and tissue engineering, among others).[2,3] Research such as this is particularly relevant as the world seeks advanced medical therapies that are personalized and specific, with the capability to be engineered for unique situations.

A variety of methods are used to study PAs and their self-assembly. There are experimental, *in vitro* techniques such as microscopy (e.g. scanning electron microscopy (SEM), transmission electron microscopy (TEM), and atomic force microscopy (AFM)) and spectroscopy (e.g. nuclear magnetic resonance (NMR), transmission infrared spectroscopy (IR), circular

dichroism (CD), dynamic light scattering, and UV-vis absorption).[2,4] While these methodologies are indispensable to the investigation and elucidation of PA self-assembly structures and characteristics, they do not provide us with the ability to directly monitor the self-assembly process on the nanosecond and microsecond timescales. They also do not fully show individual PAs in this process and their interactions with other molecules in their environment like solvent water molecules. Such molecular-level information would be, however, useful for the design of PAs. The computational methodology of molecular dynamics (MD) may help in this respect through modeling, simulation, and visualization of nanometer-level phenomena. In order to conduct MD simulations, appropriate models must first be developed or selected for the system that is to be studied. While models for amino acids and their polymer structures, proteins, exist—as well as models for various other molecules such as hydrocarbons and water—each model has advantages and disadvantages. This thesis provides a discussion of some of the models already in existence that may be used and have been used to study PAs, as well as their limitations, and it also represents work done to design models that may be more appropriate to the study of PA self-assembly.

Background information on MD, MD modeling, and amino acids and peptides/proteins is given in the next few sections, following which an overview of the literature, methodologies, study results, and study conclusions are provided. It is important to note that the development of the amino acid/peptide model that is described in the following pages of this manuscript represents the foundations of such a model, and that frequently, model development is a continuous, ongoing process. Suggestions for future work are also thus provided in the Section 6.

2. Background

2.1 Molecular Dynamics

MD is a methodology for modeling molecular systems based on numerically solving Newton's equations of motion. In MD, each atom or particle is represented with mass and radius. Each particle is both acted upon by a potential energy function, also known as a *force field*, and acts on other particles through that force field. The force fields used for biomolecule and polymer simulations, as shown in equation 1, often include terms describing the bonded interactions—the bonds, angles, dihedrals, and improper dihedrals (terms 1 through 4 in equation 1)—and the non-bonded interactions as modeled with a Lennard Jones (LJ) potential function (term 5, equation 1) and a term describing electrostatics (term 6, equation 1). Atoms (or particles) in MD are assigned initial positions—often based on molecular geometry as elucidated by experimental analysis of the molecular structure—and initial velocities. In accordance with equations 2-5, the new forces acting upon the particles are calculated and the positions and velocities are updated.

$$U = \sum k_b(b_0 - b)^2 + \sum k_\theta(\theta_0 - \theta)^2 + \sum k_\varphi(\varphi_0 - \varphi)^2 + \sum k_\phi(1 + cos(n\phi - \delta))$$
$$+ \sum 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum \frac{q_i q_j}{4\pi D r_{ij}}$$

(1)

As has been mentioned, MD offers the advantage of enabling the study of systems on much smaller time and length scales than can feasibly be done using experimental, *in vitro* techniques. This makes it an especially useful technique for studying mechanisms of processes occurring on nano scales such as peptide amphiphile self-assembly.

$$F = ma$$

(2)

$$F(x) = -\nabla U(x)$$

(3)

$$\frac{dx}{dt} = v \tag{4}$$

$$\frac{dv}{dt} = \frac{F(x)}{m} \tag{5}$$

2.2 Coarse-Grained Modeling

There are two general ways to model molecules for MD simulations: 1) all-atom (AA) modeling, which represents molecules in atom-level detail; and 2) coarse-grained (CG) modeling, which models molecules through the representation of groups of atoms in a molecule as "pseudo-atoms" or "beads."[5] In this work, the term "bead" will be used. There are usually multiple choices about which atoms to represent as a single bead. To coarse-grain a benzene molecule, for instance, one could choose to represent the entire molecule as one bead; or it could be modeled with two beads, each containing three of the six carbons in benzene, with these two beads connected by a single bond; or a model of three beads could be created, each containing two carbons, such that the benzene molecule would take the form of a triangle with each bead as a point and the sides the bonds between the beads. These are all possible "mapping schemes" for a CG benzene model, as illustrated in Figure 1. Most commonly in coarse-graining, each bead is represented as a sphere with its center at one of the atoms or at the center of mass of the atoms it represents. The number of heavy atoms (e.g. carbon, nitrogen, oxygen) per bead is represented by the terminology "N:1" (read: N to 1), in which N is the number of heavy atoms per bead. The hydrogen atoms associated with the heavy atoms are included in the bead. Typically, a bead only represents molecules of a similar type (e.g. four water molecules in one CG water bead) or atoms connected to each other through bonds,[5,6] as in the benzene example. Modeling molecules this way—by grouping atoms into beads—results in a decrease in the number of degrees of freedom involved in simulating a system containing those molecules. The advantage of that is a corresponding increase in the computational speed for the system which enables one to run a simulation at greater time and length scales than are possible with AA simulations with each atom modeled individually. This

opens up the possibility of gaining insights into systems and processes that are too large (or long) to be simulated with an AA model or that are too small (or short) to be observed experimentally. CG simulations can span volumes as large as 100 x 100 x 100 nm[3] that in atomistic detail contain millions of particles, but contain many fewer with coarse-graining. Coarse-graining also makes it possible to study processes that span microseconds to milliseconds in range. Additionally, they can enable high-throughput studies of thousands of parallel simulations, and they may provide insights into the physical nature of a process by revealing which details matter and which do not in comparison with AA simulations.[7]
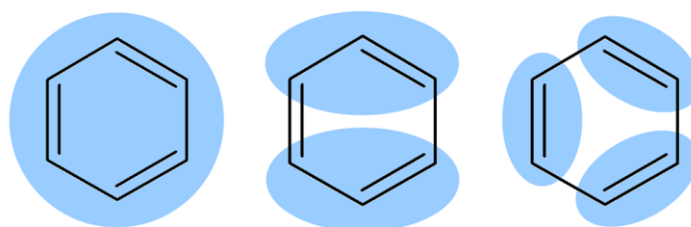

Figure 1: Examples of CG mapping schemes for benzene

When creating a CG model, care must be exercised to capture the essential properties of the molecule(s) insofar as is possible, realistic, and relevant to what is being studied. Coarse-graining smooths out the energy landscape in comparison with AA models,[6] changes the geometry of the molecule(s), and may result in a variety of other differences depending on the methodology used.

2.3 Amino Acid, Peptide, and Protein Characteristics

Amino acids are comprised of three notable chemical groups, two of which are common to all amino acids; these are an amine group and a carboxyl group connected through a carbon atom called the alpha-carbon (C_α) (shown in Figure 2 in the amino acid Serine as an example). The particular structure of the chemical groups bonded to the alpha-carbon (if any) gives each amino acid its unique identity; this group is known as the side-chain. Only one amino acid, glycine, has no side chain. The rest all exhibit unique groups bonded to the alpha-carbon such that the alpha-carbon is chiral, and the standard form found in biological systems is the L-conformation[8]

Because of the chemical nature of the amine and carboxyl groups, amino acids can exist in charged and neutral states, and they exist as zwitterions at neutral pH. Amino acids can be classified based on the properties of their side chains: hydrophobicity/hydrophilicity, charge, and polarity. The amino acids lysine, histidine, arginine, aspartic acid, and glutamic acid possess a charge at neutral pH.[8] A table of all twenty standard amino acids in their neutral forms is given in the appendix (Table A1.1).
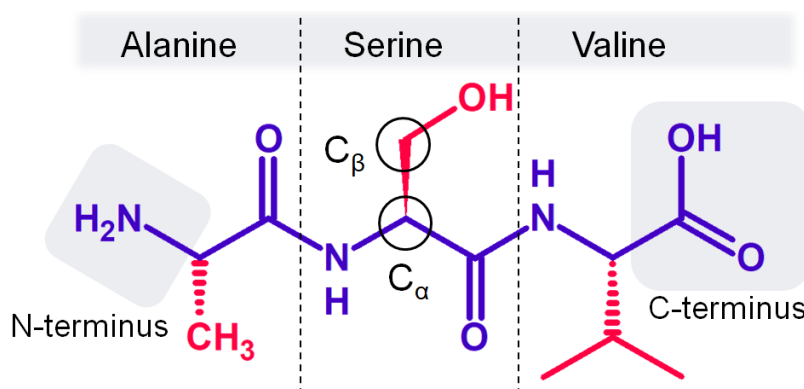

Figure 2: Peptide characteristics shown on an example tripeptide. Blue – peptide backbone; Red – amino acid side chains. Alpha-carbon and Beta-carbon shown on serine as an example.

Amino acids become linked together by the formation of the peptide bond—a condensation reaction in which the nitrogen of the amine group of one amino acid bonds with the carbon of the carboxyl group of another amino acid.[8] Larger structures that result from this (e.g. proteins and peptide amphiphiles) can be influenced by steric effects of side chains, secondary structures arising from hydrogen bond formation between peptide backbone oxygen and nitrogen atoms, and side chain interactions (e.g. pi-pi stacking). Secondary structures include α-helices and β-sheets which Figure 3 shows.

When developing a CG MD amino acid/protein model, as this work represents, these are some of the characteristics that must be taken into consideration. Section 4 following the literature review will describe the efforts made to capture the desired characteristics.
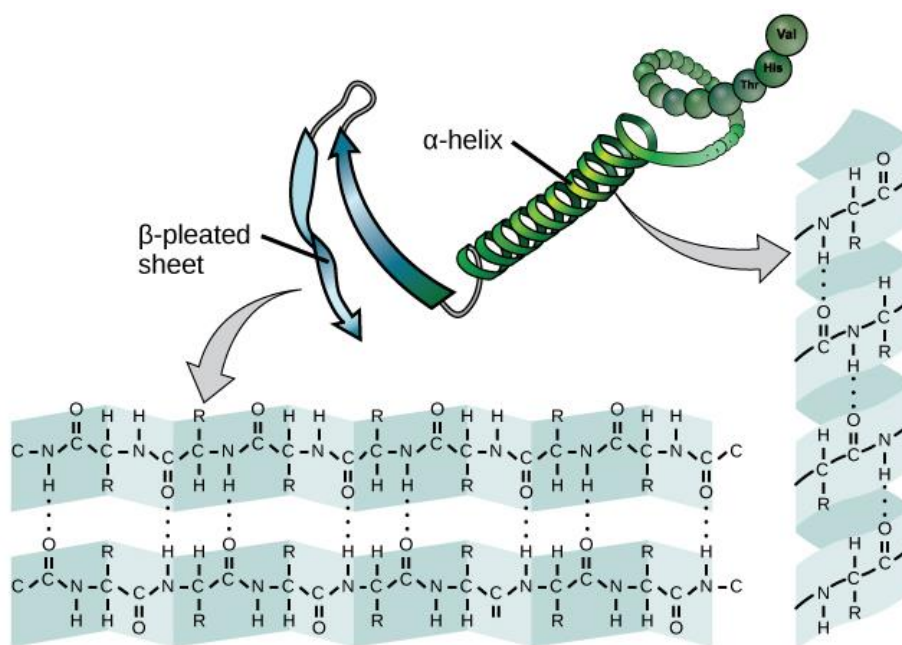
Figure 3: β-pleated sheet and α-helix secondary structures. Hydrogen bonds between peptide backbones are shown with dotted lines. Image from ref 9, licensed under creative commons.

3. Literature Review

3.1 Experimental PA Studies

There are many examples in the literature of experimental PA design with self-assembly characteristics and useful applications. A common self-assembled structure is a fiber as shown in Figure 4. In one study, microparticles were synthesized that contained the drug doxorubicin inside a shell of PAs that were functionalized to target the folate receptor. A 60-fold higher cytotoxicity was reported of the microparticles against a particular type of breast cancer cells, MDA-MB-231, compared with non-targeting particles and commercial doxorubicin.[10] In another study, self-assembly was demonstrated of a PA nanofibrous system with an antisense oligonucleotide (a potential therapeutic for several disorders) that exhibited enhanced cellular uptake and activity as compared with the oligonucleotide delivered without the peptide amphiphile.[11] The Stupp lab has investigated PAs for neural regeneration. They incorporated the laminin-derived IKVAV amino acid sequence into PAs and showed rapid, selective differentiation of neural progenitor cells into

neurons when cultured within the IKVAV PA.[12] They then used the material to treat spinal cord injury and demonstrated a reduction in cell death at the site of the injury as well as an increase in oligodendroglia (cells that form the myelin sheath) and regeneration of neurons.[13] More recently, the Matson lab designed hydrogels of functionalized self-assembling aromatic PAs for controlled release of H_2S with potential biological applications.[14] This is representative of the leading edge of research in drug delivery: to create biocompatible delivery systems that enable localized drug delivery and controlled release rates, and which undergo biodegradation in the process of—or shortly following—drug release. The PA-based hydrogels created by the Matson group have potential in all these areas.

PAs may also be used for photochemical/photophysical and electronic applications such as light-harvesting complexes. For example, Matsui and MacCuspie were able to fabricate metalloporphyrin nanotubes by coating PA nanotubes with the metalloporphyrin protoporphyrin IX Zn(II).[15] Fry, Stupp, and coworkers also experimented with peptide amphiphiles and protoporphyrin IX Zn(II). They synthesized self-assembled PA metalloporphyrin arrays that showed strong exciton interactions influenced by the peptide.[16] Garifullin, et al., reported the encapsulation of a derivative of zinc phthalocyanine within self-assembled peptide amphiphile nanofibers. The resulting complexes showed very fast intermolecular energy transfer, demonstrating the potential of PAs for the fabrication of supramolecular organic electronic devices.[17] Thus it can be seen that a substantial amount of research has shown PAs to be promising candidates in the quest for improved and enhanced therapies and technologies.

3.2 MD PA Studies

Considering the diversity of characteristics and form as well as the potential applications that PAs have, research efforts have been made in the area of mechanisms of PA self-assembly into supramolecular structures such as fibers. In MD studies, both AA models (e.g. CHARMM,[18] Amber,[19] OPLS[20]) and CG models (e.g. Martini,[21] PRME,[22] and SDK[23]) have been used. Experimental and MD

studies have both shown that for most PA fibers, their assembly occurs first through the formation of micelles followed by the merging of micelles to form short cylindrical structures followed by elongation into fibers through further merges.[24–26] Early experimental and MD studies indicated that hydrophobic interactions of the alkyl tails of PAs and the formation of β-sheets between the peptide portions of PAs are integral to the formation of fibers.[27] A CG MD study performed by Velichko, Stupp, and Olvera de la Cruz showed that under conditions of pure hydrophobic interactions without hydrogen bonding, micelles will form.[28] When the hydrophobic interactions were turned off and hydrogen bonding interactions were turned on, the PAs organized into one-dimensional β-sheets. Fibers were only observed when both significant hydrophobic attraction and hydrogen bonding resulting in β-sheet formation were at play. Later studies, however, cast doubt on the importance or even necessity of β-sheet formation for fiber formation, showing instead that β-sheet content can be negligible in PA fibers and that van der Waals forces—and in some cases electrostatic interactions—may have a much more significant role in fiber formation, along with hydrophobic interactions.[24,29,30] Research has also suggested that water ordering near the comparatively hydrophilic peptide portion of a PA provides structure to the PAs and thereby facilitates fiber formation—insights gleaned from combined AA and CG MD studies supported by experimental validation.[4] Altogether, the particular stable form (e.g. micelle, fiber, tube) of different PA self-assembled structures results from the particular strengths of and interplay between van der Waals forces, hydrogen bonding, and electrostatic interactions among the chemical constituents of the PAs and their surroundings.[31,32] While insights thus have been gained through the use of MD, the particular models implemented in the study of amino acid based molecules do have various limitations. In order to understand some of the strengths and weaknesses of these models, however, there must first be an understanding of MD and MD model development. Thus the next few sections provide an overview of these topics, as well as some details on amino acids and their

characteristics which must be considered when modeling them, and then will be described some of the models that may be used for PAs.

3.3 Methods for Development of CG Amino Acid & Protein Models

Several methods have been used to develop amino acid/protein models. They generally fall into two categories based on the mathematical formulas used to define the potentials: 1) physics-based – in which the CG potential is described by bonded terms relating bond deformation and angles, as well as non-bonded terms describing Lennard-Jones interactions (dispersion and repulsion effects) and electrostatic interactions; and 2) knowledge-based – an approach in which the CG potentials are determined by statistical analysis of known protein structures and defined by statistical models. Physics-based methods can be further subdivided into "bottom-up" and "top-down" approaches. The former is one in which AA models are used as the basis by which CG potentials are derived; while the latter is thermodynamics-based, in which the CG potentials are derived by reproducing (as closely as possible) key experimental data, especially thermodynamic properties such as surface tension, free energy of solvation, and density.[6,7]

With respect to AA-based CG force field development, there are three primary techniques that have been employed: the inverse Monte Carlo (IMC) method, iterative Boltzmann inversion (IBI), and force matching (FM), with the latter two being the most popular.[7] With the IBI method, the goal is to match the CG bead radial distribution function (RDF) with that of its AA counterpart using a CG pair potential. As its name suggests, it is an iterative process that proceeds until there is a convergence of the potential yielding the correct RDF. However, a particular RDF may be reproduced by many different pair potentials such that the solution is not unique. Yet the IBI method has been very popular because of its simplicity and quickness of convergence.[6] The IMC method also attempts to match the CG RDF with the AA RDF, and it too is an iterative process. It differs from IBI in that it handles correlations between CG force field parameters in an explicit way—such that their values are updated in an interdependent way, whereas in the IBI method the

updates are independent—and with each iteration a set of linear equations is solved to find a better approximation for the force field parameters.[6,33] Force matching, on the other hand, attempts to match CG pairwise forces to AA forces calculated based on a set of reference conformations. A later development used AA molecular dynamics (MD) trajectories as the reference. This process also may be iterative.[6]

The AA-based, physics-based approaches are more likely to capture some of the fine details of the system while the top-down approach of using experimental thermodynamic data is more likely to provide CG potentials that are transferable to other systems or states.[7] A protein-structure statistically-derived method can work well for computational biology projects because CG force-fields derived in this manner are simple and efficient. Yet their weakness is also in a lack of transferability; as Ingólfsson et al. describe it, "While the force field derived for single domain globular proteins will work well for a vast majority of single protein and peptide structures, the interaction between independent domains, interactions between proteins and nucleic acids, etc. require derivation of a new component of knowledge-based force fields."[7]

Another method of CG model development, as mentioned, is to optimize model parameters to reproduce experimental, thermodynamic properties. Examples of models that utilize this method include the Martini model, in which partitioning free energies of amino acid side chains between water and oil (cyclohexane and butane) were used to develop non-bonded parameters;[21] and the Shinoda-Devane-Klein (SDK) model which used experimental values of density and surface tension for non-bonded amino acid/protein parameters of side chain analogues.[23] Both of these models, along with some others, will be further described in the following section.

3.4 Notable Protein Models

Amino acid CG models have predominantly been developed for examining the folding, characteristics, and behavior of proteins in large systems.[6] Because such systems consist of a large number of particles, it is quite common to incorporate solvent effects into the model so that solvent

molecules are not simulated—thereby decreasing the total number of degrees of freedom and significantly speeding up the simulation. In those cases, the model is an implicit solvent model, whereas a model that requires explicit representation of solvent molecules is called an explicit solvent model. As can be seen in Table 1, many CG protein models are implicit solvent models: AWSEM (associative memory, water mediated, structure and energy model)[34]; Bereau & Deserno[35]; CABS (c-alpha, c-beta, side-chain)[36]; Hills, Lu, & Voth[37]; OPEP (optimized potential for efficient protein structure prediction)[38]; PaLaCe[39]; PRIME (protein intermediate resolution model)[22]; PRIMO (protein intermediate model)[40]; Rosetta[41]; and UNRES (united residue)[42] models. Many of these models place a greater emphasis on the protein backbone than on the side chains due to the influence of the backbone on secondary structure characteristics and protein folding. Exceptions are the Martini[21]; PRIMO; and Hills, Lu, and Voth models, which allow a higher resolution for the side chains than many of the other models. Of the explicit solvent models, the Martini model is certainly more widely used than the SCORPION model[43], especially because the model has already been well developed for other types of molecules—such as lipids[44] and DNA[45]—which is of use when modeling proteins in biological systems.

The Shinoda-Devane-Klein (SDK) model[23] is another CG force field that can be used for protein and peptide simulations, yet only non-bonded parameters were developed for the model and not bonded parameters (only generic bonded parameters were used in the parameterization of the model). Those who wish to use the SDK model thus may need to develop their own bonded parameters as in reference [46]. Regarding the non-bonded interactions, the softer LJ (9-6) potential form was used rather than the LJ (12-6) form shown in equation 1. Additionally, the backbone was modeled in a somewhat crude way by simply using the Asparagine and Glutamine side chain beads for the standard and Alanine backbone beads.[23] However, there is also an SDK CG water model that can be used in conjunction with the SDK protein model, in which parameters were developed for a CG water bead comprised of three water molecules.[46]

12

Table 1: Select CG Protein Models

Model	Solvent	Backbone Beads	Side Chain Beads	Development
AWSEM	implicit	2 (C_α, O)	1 (C_β)	mixed: knowledge-based and physics-based
Bereau & Deserno	implicit	3	1 (C_β)	knowledge-based
CABS	implicit	3	1	knowledge-based
Hills, Lu, & Voth	implicit	1 (C_α)	up to 4	physics-based, force-matching
Martini	explicit	1	up to 4	physics-based, thermodynamics-based
OPEP	implicit	5	1 (proline: 3)	mixed: physics-based and knowledge-based
PaLaCe	implicit	3	1-2	physics-based, iterative Boltzmann inversion
PRIME	implicit	3	1	knowledge-based
PRIMO	implicit	3	up to 5	physics-based
Rosetta	implicit	all backbone atoms	1-2	mixed: knowledge-based and physics-based
SCORPION	explicit	1	1-2	physics-based, force-matching
UNRES	implicit	2	1	physics-based

It has already been shown that water plays a critical role in the self-assembly of PAs,[4] and

for that reason it seems best to employ explicit solvent in PA self-assembly simulations. As noted

already, the majority of protein models developed have been developed with implicit solvent

parameters. SCORPION, an explicit solvent CG protein model, only models amino acid structures

with one backbone bead per amino acid residue (the monomeric unit), and side chains are modeled

with at most two beads for a given amino acid. This is one of the lower resolution options, whereas

a more intermediate resolution model of both the side chains and backbone may be more effective

at closely capturing the behavior and interactions of both during PA self-assembly. Additionally,

the SCORPION model utilizes an elastic network rather than explicit bonded terms for the bonded

interactions.[43] The Martini model has a higher resolution for the side chains with up to four beads

for a given amino acid. However, the backbone is modeled with just one bead per residue, and

when using the Martini model, it is required that one specify the type of backbone bead—whether it

is in an alpha helix or beta sheet or free in solution, for example. Furthermore, the Martini CG water

model shows freezing behavior between 280-300 K which the developers corrected by introducing

antifreeze particles along with the water to reproduce liquid behavior in that temperature range.[44]

A set of CG water models that do not have this limitation has been recently developed in our group

by Bejagam et al.[33]

While any of the protein models discussed in the preceding paragraphs could be chosen for

studying PAs, it may be that none of them captures PA-water interactions accurately, considering

some of the limitations addressed. If that is the case, then it is possible that the self-assembly

pathways demonstrated from the use of those models may be wrong. Yet a thorough knowledge of

the actual steps involved in the process of self-assembly would be advantageous for designing PAs

and controlling the self-assembly process. Therefore we considered the development of new CG

amino acid models a worthy endeavor—models whose interaction parameters could be tuned with

respect to the aforementioned water model developed in our group.[33]

4. Model Development

4.1 Overview

In addition to utilizing the water models developed by Bejagam, et al.[33] in order to create

amino acid models for use with explicit solvent, we were interested in using other CG models that

members of our lab had recently developed: CG models for hydrocarbons[47,48] and benzene.[33] These

could be used for modeling hydrocarbon tails of PAs and parts of amino acids as well. We

determined to create an intermediate-resolution model of both the amino acid side chains and the

peptide backbone so as to more closely capture the behavior of both during PA self-assembly. The

peptide backbone would be parameterized such that it would not be necessary to set a

predetermined secondary structure, unlike in the Martini model.[21] Additionally, a 12-6 LJ

functional form was chosen so as to harmonize with the water, hydrocarbon, and benzene models

since those were developed with the same LJ functional form.

Our mapping schemes would be 2:1 where possible, and only larger (e.g. 3:1 or 4:1) when

constrained by the atomic bonding pattern within the molecule or in the event of other constraints

necessary for creating a stable CG model. A 2:1 mapping scheme would allow for greater accuracy

as compared with beads containing larger numbers of heavy atoms per bead, while simultaneously

offering an advantage in a decrease in the number of degrees of freedom which increases

computational speed as compared with AA models as previously described. The maximum number

of heavy atoms represented in a given bead in the models developed is four (4:1 beads), as can be

seen in the mapping schemes given in Table A1.1 in the appendix. The overall ring structures of

Phenylalanine, Tyrosine, Tryptophan, and Histidine were preserved. The centers of the beads are

located at the center of mass of the atoms they represent. The mapping schemes that were

developed also provide for transferability across amino acids where possible, such that in some

cases the same bead type could be used in more than one amino acid (e.g. C2M in methionine,

glutamine, and glutamic acid).

We determined to develop a physics-based model as opposed to a knowledge-based model

since the latter is derived from statistical analysis of large amino acid assemblies in the form of

proteins and protein complexes, neither of which represents the target application for our model.

We also determined that the base, initial model would only possess neutral beads corresponding to

neutral (uncharged) amino acid side chains and termini. The terms that our model would need to

account for are thus all of those given in equation 1 with the exception of the fifth term representing electrostatic interactions which is important for force fields (e.g. AA force fields) that possess particles with full or partial charges. We also chose a hybrid method for parameterizing our model that utilizes AA distributions for developing the CG bonded parameters and experimental thermodynamic data for developing the non-bonded parameters of the CG force field. This methodology was successfully used in our group to develop the hydrocarbon models.[47] The bonded parameters we could obtain by mapping distributions from AA molecular dynamics simulation trajectories to CG distributions. The non-bonded parameters were determined in a manner similar to that used in the SDK model development, in which the non-bonded parameters or side chain beads were optimized to reproduce thermodynamic properties of amino acid side chain analogues.[23] Side chain analogues are molecules that closely approximate the structure of the amino acid side chains. For example, the side chain analogue of neutral Lysine is N-butylamine as is shown in Figure 16 on page 26. While the SDK model was reported to have performed well in predicting the native structures of a set of proteins, bonded parameters were not specifically developed for the model, and beads that represented the backbone were patterned off of the Asparagine and Glutamine side chains[23] as previously mentioned. Additionally, while the SDK model was parameterized using only experimental values for density and surface tension, we included a third optimization experimental value, heat of vaporization, when such information was available for an analogue. Including heat of vaporization in the optimization process would help ensure that the non-bonded parameters were really ideally optimized to reproduce the appropriate thermodynamics of the chemical constituents of the amino acids.

Following development of the bonded and non-bonded parameters, we would be able to run CG PA simulations to test for self-assembly and the appropriateness of the parameters, and to determine what subsequent steps should be taken in the model parameterization. Tuning the

interaction parameters with the water model also could be evaluated. Obtaining hydration free

energies for amino acid analogues was also used for that purpose.

4.2 Mapping Schemes and Analogues

Determining the mapping scheme is one of the first steps in CG model development. In this

section, the mapping schemes for the amino acids will be shown. The choice of analogues used to

represent the amino acid side chains, backbone beads, and termini beads will be discussed. Any

particularities for a given amino acid or analogue will also be conveyed. More details of the

mapping schemes are given in Table A1.1 (for the amino acids) and Table A1.2 (for the analogues)

in the appendix.

4.2.1 Alanine
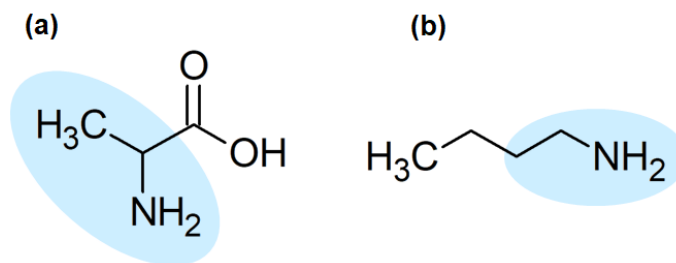
Figure 5: (a) Alanine with mapping scheme and (b) N-butylamine

Alanine is the amino acid with the smallest side chain—just a methyl group. This made it

necessary to model the alanine backbone differently than the standard backbone associated with

most of the other amino acids, with the side chain grouped in with the alpha carbon and nitrogen,

as shown in Figure 5. To have modeled the side chain separately would have resulted in a bead so

lightweight that it would have been unstable at larger time steps than are possible when

implementing a 2:1 or greater ratio mapping scheme. A lower mass leads to a greater vibrational

frequency in connection with any beads it is bonded to, and this correlates with a smaller time step

required for stability.[49] N-butylamine was chosen for parameterizing the alanine beads because of

its structural similarity, with both the backbone Alanine bead (AB) and the N-terminal alanine bead (ANT) modeled with the same non-bonded parameters.

4.2.2 Arginine
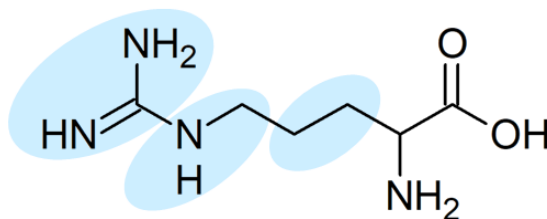
Figure 6: Arginine with side chain mapping scheme

The structure of the arginine side chain is the most unique as compared with the other amino acid side chains. The closest analogue to the nitrogenous end group is guanidine, for which not a large quantity of experimental thermodynamic data are available. The pKa of this end group is 13.8,[50] meaning it is almost never in neutral form, and almost always it is protonated. DeVane, et al., during the creation of the SDK model, used experimental data from guanidinium salt solutions to parameterize the charged end group of arginine.[23] Yet since the base model development described in this thesis is only of neutral species, we modeled a neutral form of arginine as well. We used three beads to describe the side chain, as shown in Figure 6. For the outermost bead in the side chain—which represents two nitrogen atoms and one carbon with associated hydrogens—we used an AA model of the arginine side chain and ran simulations to calculate values for the density, heat of vaporization, and surface tension at 298 K for optimizing the non-bonded parameters. Because the CHARMM AA force field does not include a model of a neutral arginine, we obtained partial charges for the atomic model from Gaussian[51], an *ab initio* quantum chemistry software package that can be used for predicting many properties of molecules, including atomic charges. Some of these charges were modified so as to harmonize better with the CHARMM values for the side-chain atoms and assign no difference between hydrogens bonded to the same carbon atom, as shown in Table A4.1 in the appendix which provides the atom names (as given in the CHARMM files), the

18

Gaussian charges, the CHARMM charges, and the charges that were assigned. Regarding the other beads in the CG model, the middle side chain bead was modeled with the same non-bonded parameters as the NC (nitrogen plus C_α) bead of the backbone, and the inner side chain bead was modeled with the same non-bonded parameters as those of the two-carbon C2M bead previously developed by a member of the group from modeling hydrocarbons.[47]

4.2.3 Asparagine
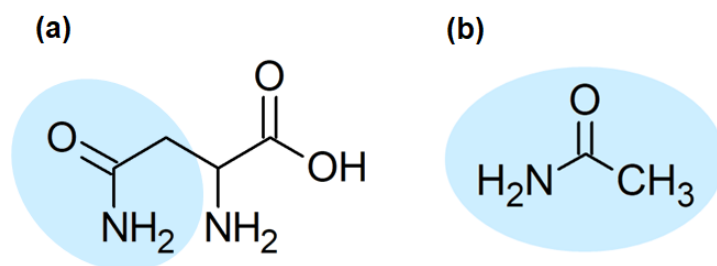
Figure 7: (a) Asparagine with side chain mapping scheme, and (b) acetamide

Because of the structure of the asparagine side chain, it was necessary to model the entire side chain as one bead as shown in Figure 7(a) above. Experimental values of acetamide were used for optimizing the non-bonded parameters. However, because the melting point of acetamide occurs at 81 °C (354 K),[52] all simulations were run at 358 K at which temperature the molecule is in liquid form. This was for the calculations performed in the optimization of the non-bonded parameters, as is described in Section 4.4.

4.2.4 Aspartic Acid
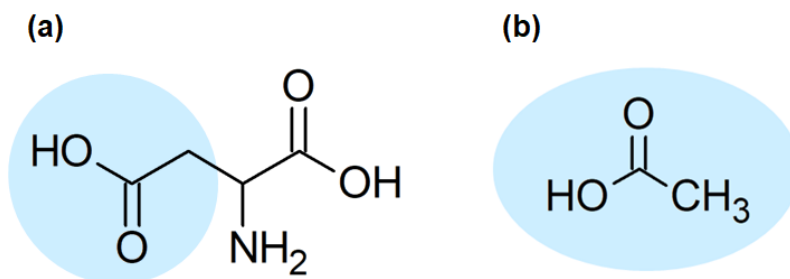
Figure 8: (a) Aspartic acid with side chain mapping scheme and (b) acetic acid

Like asparagine, the aspartic acid side chain was modeled as one bead (Figure 8(a)) because of the bonding pattern between the side chain atoms. Acetic acid experimental values were used for optimizing the non-bonded parameters.

4.2.5 Cysteine

(a)   **(b)**
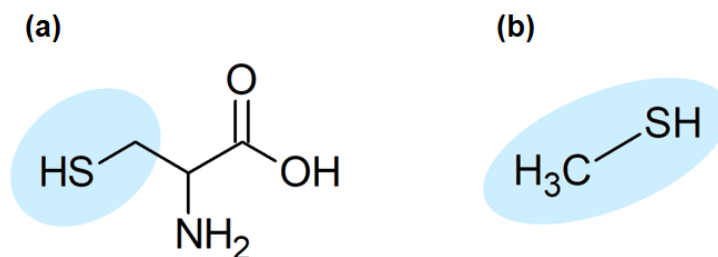
Figure 9: (a) Cysteine with side chain mapping scheme and (b) methanethiol

The analogue for the cysteine side chain is methanethiol (methyl mercaptan); its similarity to the cysteine side chain is shown in Figure 9 above. Its boiling point is 6 °C (279 K)[52] so all simulations were run at 278 K. The value for the heat of vaporization used was that at the boiling point, which is given in Section 4.5, as a close approximation to the value at 278 K since a value was not found for the slightly lower temperature.

4.2.6 Glutamic Acid

(a)   **(b)**
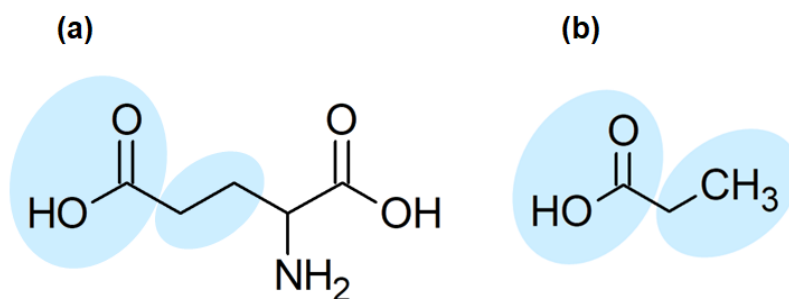
Figure 10: (a) Glutamic acid with side chain mapping scheme and (b) propionic acid

Two beads were used to model the glutamic acid side chain (Figure 10(a)). Non-bonded parameters for the outer bead—representing the carboxyl group of the side chain—had already been obtained by previous work done in the group (unpublished results) from optimization with

respect to propionic acid experimental values. The inner bead was modeled with the same parameters as the C2M bead obtained in previous work.[47]

4.2.7 Glutamine

(a)   **(b)**
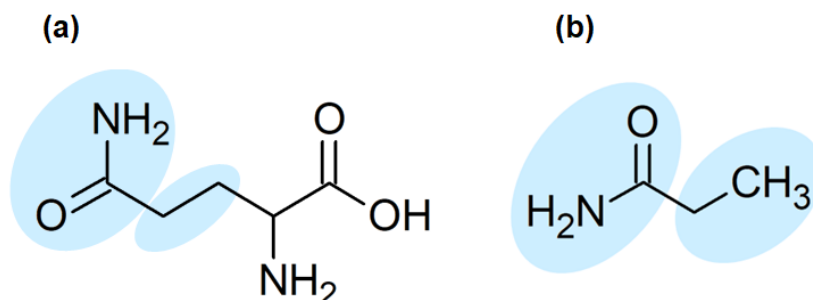
Figure 11: (a) Glutamine with side chain mapping scheme and (b) propionamide

Propionamide was used as the analogue for glutamine as it represents the side chain very accurately as can be seen in Figure 11. Its melting point is 81 °C (354 K),[52] so all simulations were run at 356 K to target experimental thermodynamic properties of the liquid phase for optimizing the CG force field parameters. The inner side chain bead was assigned the C2M type as in the case of the inner side chain bead of glutamic acid.

4.2.8 Glycine

(a)   **(b)**
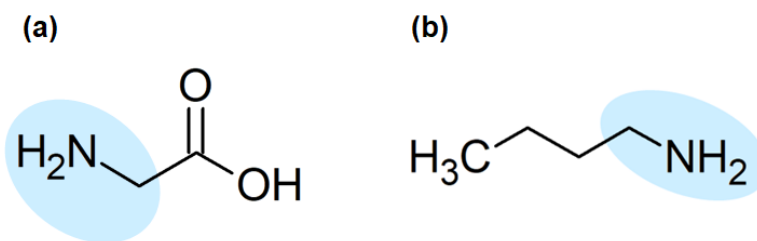
Figure 12: (a) Glycine with mapping scheme and (b) N-butylamine.

Glycine is the only amino acid that does not have a side chain. Its N-terminal C_α-amine group was modeled the same as any other N-terminal group, and its backbone C_α-amine group was modeled the same as the standard backbone C_α-amine group, the only difference in both cases being an extra hydrogen as compared with the other N-terminal and backbone groups. The standard N-terminal analogue is N-butylamine which is shown in Figure 12 above to be a good analogue for the glycine N-terminal bead. However, analysis of the AA CG-mapped distributions would indicate

21

whether the bonded parameters should be modeled the same as for the other standard amino acids or differently.

4.2.9 Histidine
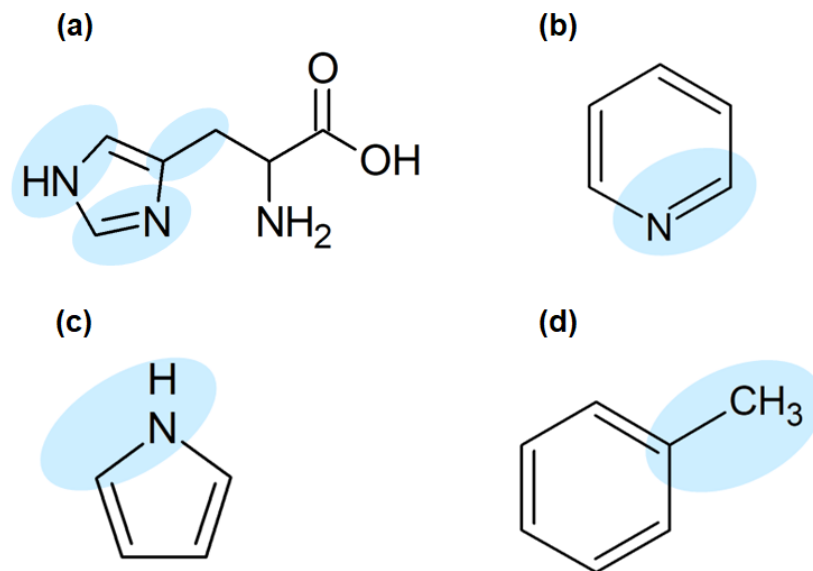
Figure 13: (a) Histidine with side chain mapping scheme, and side chain analogues (b) pyridine, (c) pyrrole, and (d) toluene.

The analogue of the histidine side chain is 4-methylimidazole. Because experimental thermodynamic properties for the analogue are lacking in its liquid phase, we used pyridine, pyrrole, and toluene for optimizing the parameters of the three histidine side-chain beads. The similarities of some of the constituents of these analogues to the histidine side chain mapped constituents can be seen in Figure 13. The r_{min} values for the analogue and amino acid beads were compared to check for similarity as a validation (or contradiction) of the appropriateness of those analogues for parameterizing the histidine beads.

4.2.10 Isoleucine

(a)   (b)
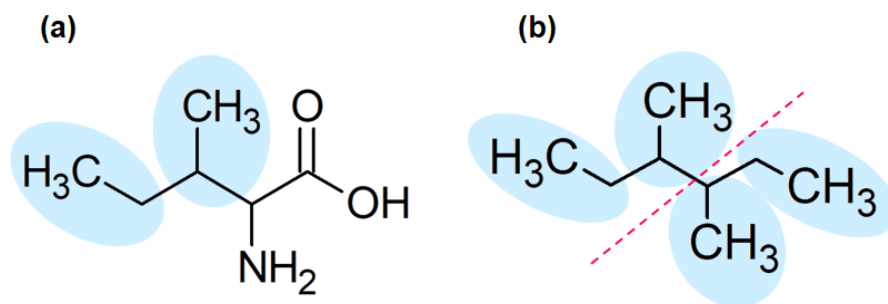
Figure 14: (a) Isoleucine with side chain mapping scheme and (b) 3,4-dimethylhexane. Dashed line shows that 2,5-dimethylhexane is a symmetric analogue of the isoleucine side chain.

Butane is a close analogue of the isoleucine side chain, but it exists as a gas at room

temperature, and its condensation point is -0.5 °C.[52] Therefore, we chose 3,4-dimethylhexane—

which is a liquid at room temperature[52]—for optimizing the non-bonded parameters of the

isoleucine side chain. This is the same approach as that taken in the development of the SDK

model[23] We gave the outermost side chain bead the same parameters as the C2E bead in the

hydrocarbon optimized previously by An, et. al.[47] Yet because of the slight difference between the

inner side chain bead and the outer side chain bead, as shown in Figure 14, we decided to model the

inner side chain bead as different from the outer side chain bead.

4.2.11 Leucine

(a)   (b)
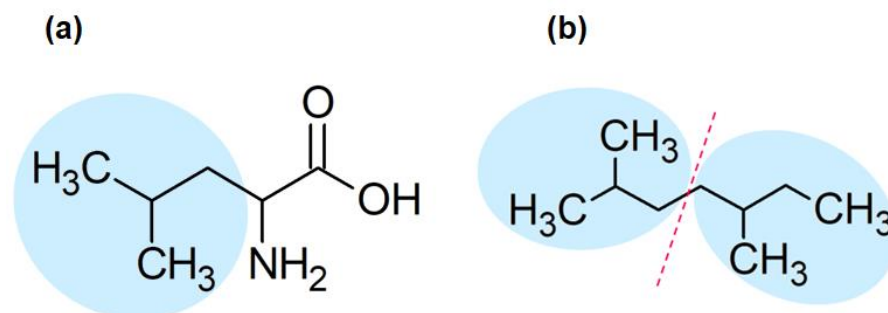
Figure 15: (a) Leucine with side chain mapping scheme and (b) 2,5-dimethylhexane. Dashed line shows that 2,5-dimethylhexane is a symmetric analogue of the leucine side chain.

The side chain of leucine had to be modeled as one bead because of the bond pattern. As

with the isoleucine side chain, the most obvious analogue for leucine—isobutane—is a gas at room

temperature and has the low condensation point of -11.7 °C.[52] So using the same approach as

DeVane, et al.,[23] we used 2,5-dimethylhexane--which is liquid at room temperature[52]--for

optimizing the non-bonded parameters of the leucine side chain. Figure 15 shows that the

symmetry of 2,5-dimethylhexane provides two identical beads that are exact analogues of the

leucine side chain.

4.2.12 Lysine
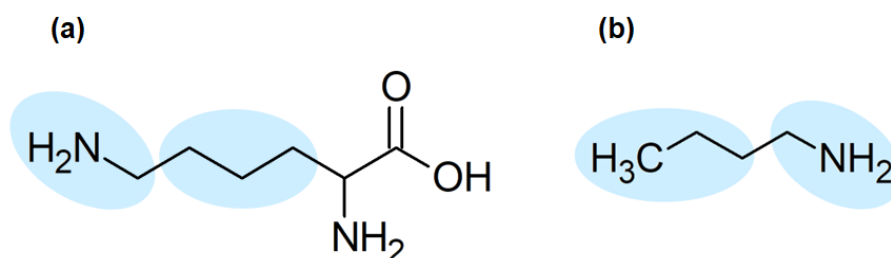
Figure 16: (a) Lysine with side chain mapping scheme and (b) N-butylamine

To model the neutral lysine side chain, we used N-butylamine to obtain non-bonded

parameters for the outer side chain bead along with the C3M non-bonded parameters for the inner

side chain bead. As shown in Figure 16, N-butylamine is a near-exact match to the lysine side chain.

4.2.13 Methionine
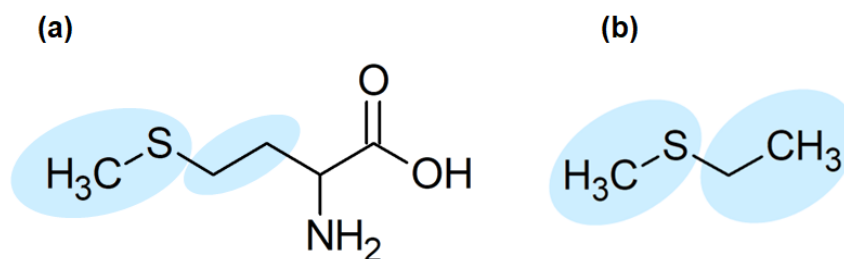
Figure 17: (a) Methionine with side chain mapping scheme and (b) ethyl methyl sulfide

Ethyl methyl sulfide is the analogue of the methionine side chain, as can be seen in Figure

17. The C2M parameters were used for the inner side chain bead, leaving only the outer side chain

bead—representing the sulfur atom and outermost carbon with associated hydrogens—to be

optimized with respect to the experimental values of ethyl methyl sulfide.

4.2.14 Phenylalanine

(a)   **(b)**
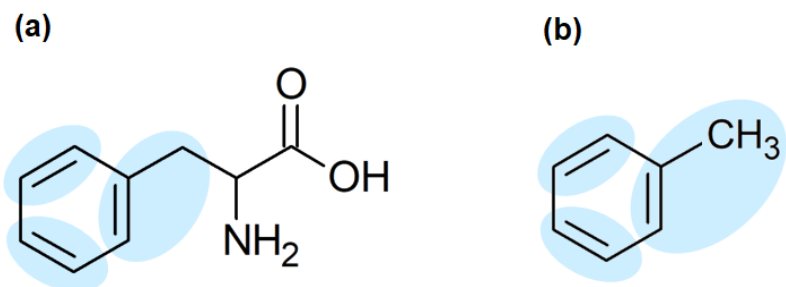
Figure 18: (a) Phenylalanine with side chain mapping scheme and (b) toluene

Toluene is the obvious analogue for phenylalanine, as shown in Figure 18. As our lab had already obtained non-bonded parameters for a two-carbon bead from modeling benzene (type BZ),[33] those non-bonded parameters were used for the two-carbon side chain beads. That left only the three-carbon bead non-bonded parameters to be optimized. As in such cases with the other amino acid side chains, we verified this decision by comparing the RDFs of the toluene beads from using the BZ parameters with those of the AA mapped CG toluene RDFs and the phenylalanine side chain RDFs.

4.2.15 Proline
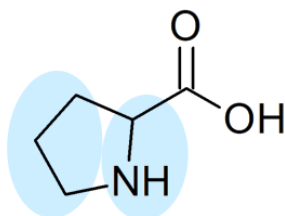
Figure 19: Proline with mapping scheme

Proline has the peculiar feature of having a side chain that is connected back to the amino acid nitrogen, as shown in Figure 19 above. It was decided that the three-carbon side chain bead could be modeled as C3M for its non-bonded parameters and that the nitrogen-C_α bead could be modeled using the same parameters as those developed for the lysine outer side chain bead. Comparing RDFs of the CG proline beads and the RDFs of the AA mapped proline beads shows that this is not an unreasonable approximation. Pyrrolidine could be used to check or tune these

parameters to potentially better fit the structure, but that did not seem likely to be necessary to

capture the essential attributes of the proline beads. The tendency of proline to cause turns or

kinks in a peptide or protein sequence is a more defining attribute of the amino acid, and that is

more likely captured through the bond and angle distributions. The Val-Pro-Leu tripeptide was

used to obtain those distributions.

4.2.16 Serine

(a)   (b)
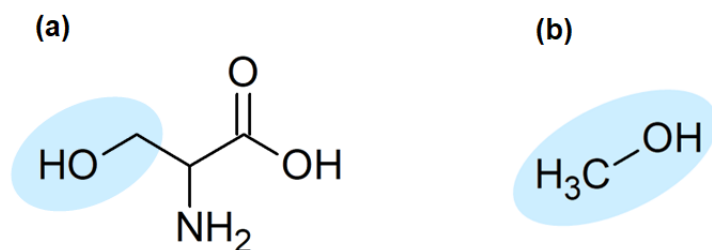
Figure 20: (a) Serine with side chain mapping scheme and (b) methanol

As can be seen in Figure 20 above, the obvious analogue for serine is methanol. Since

methanol is a liquid at room temperature and is well-characterized with respect to its

thermodynamic properties, there was no need to consider any other analogue option for obtaining

the non-bonded parameters for the serine side chain bead.

4.2.17 Threonine

(a)   (b)
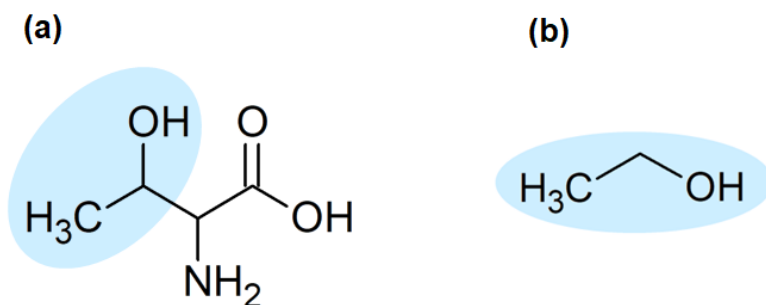
Figure 21: (a) Threonine with side chain mapping scheme and (b) ethanol

Ethanol is the analogue we used for the threonine side chain because of its similarity to the

threonine side chain as shown in Figure 21. Like methanol for the serine side chain, ethanol is

liquid at room temperature and well-characterized with respect to its thermodynamic properties such that no other analogue option was necessary.

4.2.18 Tryptophan
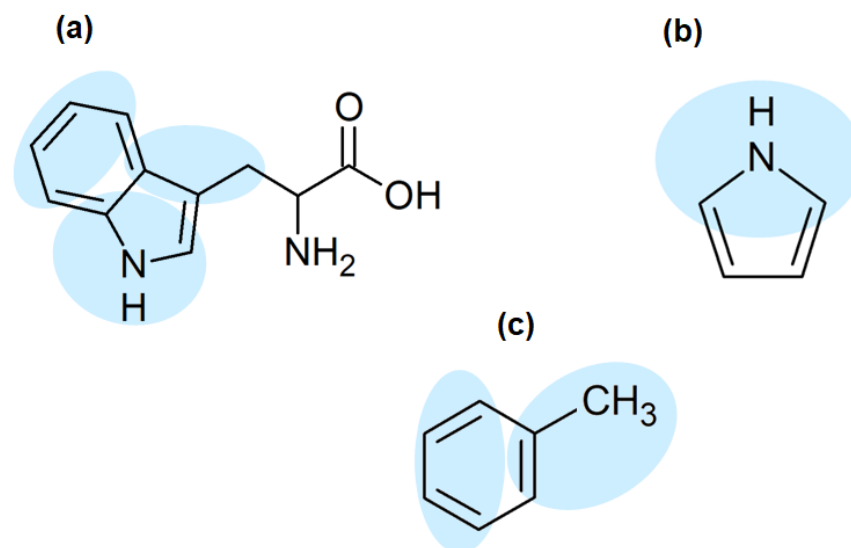
Figure 22: (a) Tryptophan with side chain mapping scheme, and analogues (a) pyrrole and (b) toluene

Tryptophan has the largest, bulkiest side chain. Initially we tried to use five beads and two rings for mapping the side chain, but we were not able to make that mapping scheme functional. When we attempted to run simulations with the five bead side chain, they were so unstable that they crashed immediately. This instability appeared to result from some beads in the side chain being counted in intramolecular interactions multiple times, which happened in part because of the two ring structures (resulting in multiple paths along which a bead could interact with others) and in part because only the interactions between beads directly bonded to each other were excluded. A four-bead ring was also ruled out as problematic, so the mapping scheme was instead finalized as one ring made up of three beads with the masses as similar as possible. As with histidine, however, the non-bonded parameters for these beads were optimized using more than one analogue— toluene and pyrrole in this case. The similarities of some of the constituents of these analogues to constituents in the tryptophan side chain are shown in Figure 22.

4.2.19 Tyrosine
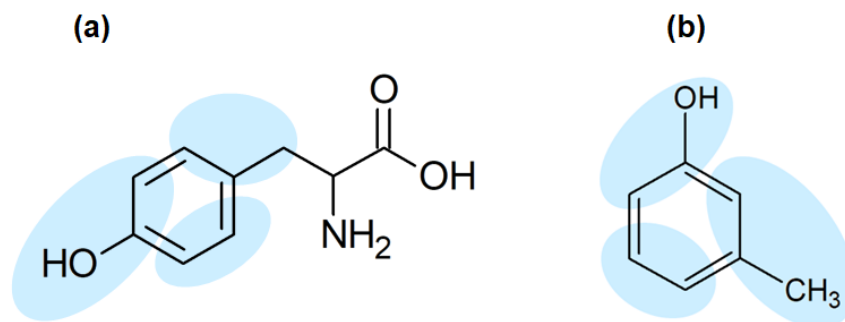
Figure 23: (a) Tyrosine with side chain mapping scheme and (b) m-cresol

The analogue chosen for tyrosine was m-cresol because of its similarities to the tyrosine side chain and because of the availability of experimental thermodynamic data at an appropriate temperature for parameterization. Because of the number of atoms in the tyrosine side chain and because of the way they are bonded to each other, a number of different options exist for the mapping scheme. The particular mapping scheme we settled on, as shown in Figure 23(a), preserves a ring structure for the CG tyrosine. Initially we attempted to model the tyrosine side chain with four beads, adhering to a preference for a 2:1 mapping scheme (two heavy atoms per bead). This mapping scheme proved unstable, however, so the final mapping scheme consists of only three beads. Toluene was used as the analogue for the three carbon bead, the two carbon bead was modeled with the same non-bonded parameters as those of the benzene bead, BZ,[33] and the parameters for the bead with the hydroxyl group were optimized using m-cresol in combination with the already-developed non-bonded parameters for the other side chain beads. Because the structure of m-cresol differs slightly from that of the tyrosine side chain, however, the bonded parameters for tyrosine were set based on the AA CG mapped distribution of the tyrosine dipeptide rather than from the m-cresol analogue.

4.2.20 Valine

(a)   (b)
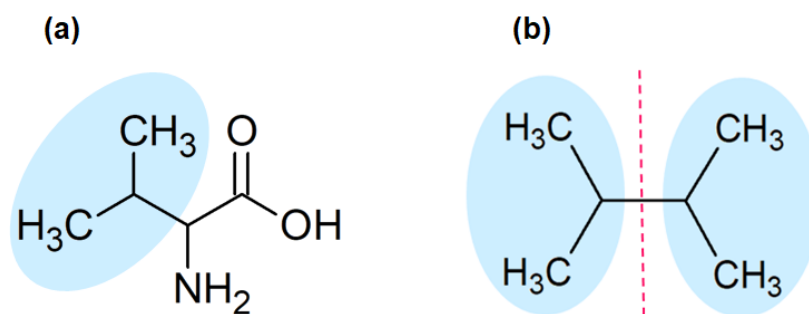
Figure 24: (a) Valine with side chain mapping scheme and (b) 2,3-dimethylbutane. Dashed line shows that 2,3-dimethylbutane is a symmetric analogue of the valine side chain.

The analogue used for valine was 2,3-dimethylbutane since it exists as a liquid at room temperature and since the more obvious analogue, propane, exists in the gas phase at that temperature.[52] This is also the same procedure that DeVane, et al. used in the development of the SDK model.[23] Figure 24 shows that the symmetry of 2,3-dimethylbutane provides two identical beads that are exact analogues of the valine side chain.

4.2.21 Backbone and Termini

(a)   (b)

(c)
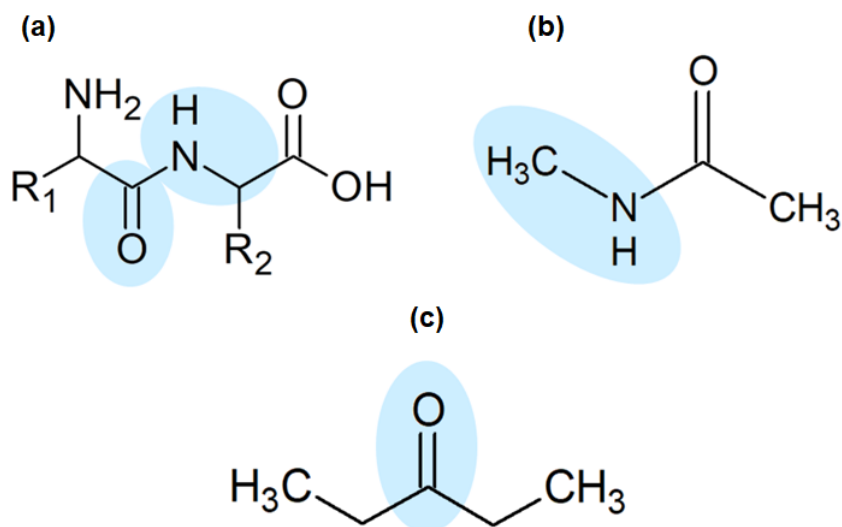
Figure 25: (a) A generic dipeptide showing the backbone mapping scheme, and analogues (b) N-methylacetamide and (c) diethylketone.
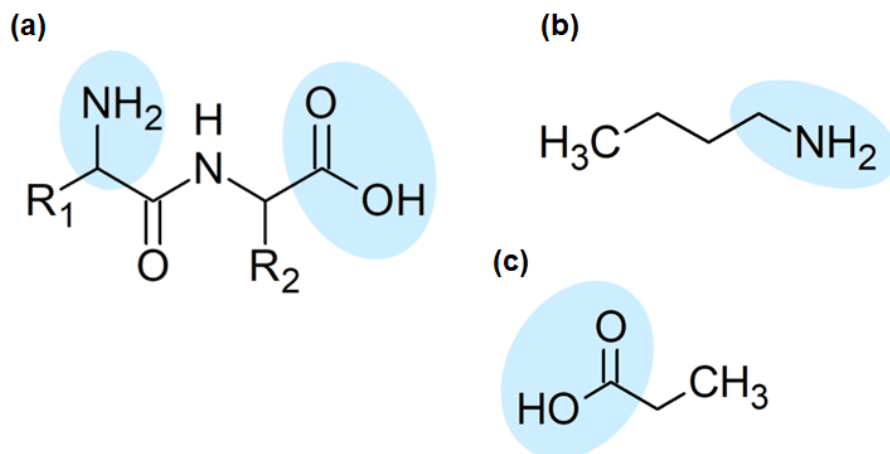
Figure 26: (a) A generic dipeptide showing the mapping scheme for the N-terminus and the C-terminus, and analogues (a) N-butylamine and (c) propionic acid.

In this model, there are two beads for modeling the standard peptide backbone and two beads for modeling the termini, as shown in Figures 25 and 26. This is different from most protein models which tend to use either one bead for the entire backbone repeat unit (Nitrogen, alpha-Carbon, and carbonyl together) or which describe the backbone in AA detail, as shown in Table 1. Non-bonded parameters had already been determined by another group member for a bead with the same structure as the C-terminal bead when modeling propionic acid (unpublished results). The N-terminal bead was given the same parameters as the outer lysine side chain bead. N-methylacetamide was chosen as the analogue for the N-C$_\alpha$ bead (BNC), and diethylketone was chosen as the analogue for the backbone carbonyl (BCO). Simulations for N-methylacetamide were run at 313 K since its melting point is 28 °C (301 K).[52]

4.3 Bonded Parameters

After the mapping schemes were determined, the next step that was undertaken was the running of AA simulations to obtain bond, angle, and distributions for the CG models by mapping the AA molecule to its CG version using the AA simulation trajectories. Considering that we needed to obtain parameters for bonds, angles, and dihedral angles related to the side chain beads, terminal beads, and peptide backbone beads, we decided to run AA simulations of predominantly

homogeneous dipeptides—dipeptides that contain two of the same amino acid, e.g. alanyl-alanine and leucyl-leucine. This would result not only in bonds and angles unique to a given amino acid in a peptide, but it would also produce backbone bonds and angles and their force constants for each dipeptide simulation. Considering the uniqueness of the side chains of the different amino acids, the backbone bonds, angles, and force constants resulting from the AA simulations were likely to differ somewhat across dipeptide simulations. Thus we determined to use all of the dipeptide simulations in addition to a valine tripeptide simulation to define what would seem to be optimal bonded parameters for the backbone beads. Valine was chosen as the homogeneous tripeptide amino acid due to the moderate bulkiness of the side chain and single bead modeling of the side chain for ease of use. SDF (structure) files of the dipeptides were obtained from PubChem[53] and converted to PDB files using the molecular editing software Avogadro.[54] Because of the uniqueness of proline and arginine, non-homogeneous dipeptides were used to model those amino acids. Additionally, because we chose to bundle the side chain of alanine in with its terminal and backbone beads, we used a homogeneous tripeptide, alanyl-alanyl-alanine, in order to capture those angles and dihedrals. We also used structures that contained the L forms of the amino acids.

For each dipeptide or tripeptide, Packmol[55] was used to generate a PDB file containing five hundred dipeptide/tripeptide molecules. The box sizes varied because the sizes of the different dipeptides/tripeptides varied. Molecular dynamics simulations were run with the CHARMM[18] force field and the molecular dynamics software NAMD[56] (version 2.12), with the molecular visualization program VMD.[57] The simulations were conducted at 298 K with a 1 fs time step for 5 ns, in the NPT ensemble (constant moles, pressure, and temperature) using Langevin dynamics for temperature control and the Nosé-Hoover Langevin piston barostat for pressure control. A configuration file is given in the appendix, A26, as an example. The Langevin thermostat is a popular choice for MD simulations. It produces the correct canonical distribution and controls temperature by removing energy from the system with friction forces and by adding energy to the system with random

forces.[58] The frictional and random forces are incorporated by adding them to the Newtonian

equations of motion[56] as shown in equation 6 in which γ is the friction coefficient, v is the velocity,

k_B is the Boltzmann constant, T is temperature, m is mass, and $R(t)$ is the univariate Gaussian

random process. The fluctuating (random) force is described by the last term in equation 6 while

the dissipative (frictional) force is described by the $-\gamma v$ term. Pressure control with the Nosé-

Hoover Langevin piston barostat used in NAMD generates the correct NPT ensemble distribution

when used in conjunction with the Langevin thermostat.[56]

$$ma \; = \; F(r) \; - \; \gamma v \; + \; \sqrt{\frac{2\gamma k_B T}{m}} \; R(t) \tag{6}$$

NAMD implements the velocity-Verlet integration method for advancing the positions and

velocities of each atom/bead in time, as shown in equations 7-10. The velocity-Verlet algorithm is

simple, it is symplectic (demonstrating long-time stability) and time-reversible, and it requires only

one force evaluation for each time step, making it suitable for biomolecular simulations which are

typically large and complex.[56]

$$r(t + \Delta t) \; = \; r(t) \; + \; v(t)\Delta t \; + \frac{1}{2} a(t)\Delta t^2 \tag{7}$$

$$v(t + \frac{\Delta t}{2}) \; = \; v(t) \; + \; \frac{1}{2}(t)\Delta t \tag{8}$$

$$a(t + \Delta t) \; = \; -\left(\frac{1}{m}\right)\nabla V[r(t + \Delta t)] \tag{9}$$

$$v(t + \Delta t) \; = \; v\left(t + \frac{\Delta t}{2}\right) + \frac{1}{2}a(t + \Delta t)\Delta t \tag{10}$$

Particle Mesh Ewald (PME) was used for electrostatic interactions. PME is a method that

divides the electrostatic interactions into a long-range contribution, which is calculated using a

Fourier transform; and a short-range contribution, which is calculated in real space. PME is both

highly accurate and reasonably fast, thus making it a standard choice for MD simulations.[59] For the

non-bonded interactions, a cutoff of 12 Å was used with 1-4 scaling. With respect to the cutoff, this

means that in the Lennard-Jones potential, calculations were performed only on atoms within 12 Å from each other rather than allowing the potential to very gradually drift towards zero at long ranges. The 1-4 scaling means that within a molecule, non-bonded interactions between atoms located within two bonds from each other were ignored while the interactions between an atom and another atom three bonds away from it were calculated with reduced effect. Interactions between atoms farther apart from each other than three bonds in the same molecule would have full effect.

Following these AA simulations, the CG trajectory was created based on the AA simulation trajectory. The way this works is by identifying which atoms in a molecule make up which bead and then by obtaining the center of mass of those atoms which becomes the center of mass of the bead. Using these centers of mass, one can then calculate the length of the vector between two beads bonded to each other (i.e., the length of the bond between two beads bonded to each other), the angle created by three beads bonded to each other, and the dihedral angle created by four beads bonded to one another in series. All the lengths of a bond of a given type, all the angles of a given type, and all the dihedrals of a given type are collected and binned into histogram distributions. The resulting curves are used to obtain the equilibrium bond lengths and angles and approximations of the dihedral angles. These files also create a CG trajectory from the AA trajectory and from the information given about the mapping of the AA molecule to a CG molecule. This CG trajectory for each dipeptide/tripeptide simulation was loaded into VMD to obtain the radial distribution function (RDF) for each bead type with itself. The same procedure was carried out for the side chain and backbone analogues to check for consistency or discrepancy between the amino acid beads and the analogues used for obtaining optimal non-bonded parameters.

4.4 Non-bonded Parameters

The positions of the RDF peaks were used as center values for a narrow range given for optimization of the r_{min} values which are related to the σ values of the LJ potential for the beads

according to equation 11. The RDF is a measure of the probability of finding another particle (e.g.

atom, bead) within any given distance from the particle itself. For any substance, the likelihood of

finding a particle of its own kind within a distance less than that of the diameter of the particle is

essentially zero because of the strong repulsive force that prohibits the overlapping of particles.

But for a closely packed structure like a liquid, there will be a high probability of finding another

particle at a distance approximately equal to the diameter of the particle (Chandler, 1987).[60] This

high probability corresponds to the first peak of a graph of the RDF. In this way, the RDF can be

used to obtain the σ values of the LJ potential since σ represents the distance at which the potential

between two particles is zero. Ranges for the ε values of the LJ potential were given a little wider

and were based on the polarities of the atoms represented by the beads; for example, a maximum

lower bound was typically -1.0 while a maximum upper bound was usually around -0.3.

Hydrocarbon beads tend to have less negative epsilon values while beads containing one or more

nitrogen or oxygen atoms tends to have more negative epsilon values. The optimization algorithm

can then select values from those ranges for testing in simulations, and the parameters were

optimized by comparing the heat of vaporization, surface tension, and density values of the CG

analogue simulations with the experimental values for those analogues. Particle Swarm

Optimization (PSO) was used for this purpose.

$$r_{min} = 2^{1/6}\sigma \qquad (11)$$

PSO is an algorithm that takes inspiration from the movements of a flock of birds or swarm

of insects. To begin with, a set of values (a "particle") for the force-field parameters to be optimized

is randomly selected based on the supplied ranges for use in a test simulation. Many such particles

are generated for simultaneous test simulations. Each particle is analogous to a bird in a flock (or

insect in a swarm). In our case, we chose to generate forty such particles for each iteration of the

PSO for broad coverage of the parameter space to facilitate quicker identification of optimal force

field parameters. Once the heat of vaporization, surface tension, and density were calculated for

each simulation, they were compared with the experimental values and ranked based on the errors. The best set is that which corresponds to the lowest error. For the subsequent iteration, the best particle remains the same, and all other particles have their values adjusted slightly in the direction of the values of the best particle. A record is also kept of an individual particle's best set—and updated when it changes—so that the particle may be moved in that direction as well. The maximum amount by which each value can change is given by the investigator. PSO is implemented using equations (12) and (13) in which V represents how quickly a given value changes (its velocity), x represents the value itself, the subscript n stands for the current step, and $n+1$ stands for the next step. The global best (the best set of values across all iterations) is represented by g_{best}, and p_{best} is for the personal best set of a particle. The value of $rand()$ is a random number between 0 and 1, w is an inertia factor, and $c1$ and $c2$ are swarm and personal constants, respectively. The first term in equation (12) is an inertial term that stabilizes the particle motion, and the constants $c1$ and $c2$ give the relative magnitude or pull in the direction of the global best and personal best, respectively. The position of a particle is updated based on the distance traveled over a discrete time interval ($\Delta t = 1$).[33] The investigator can set the number of iterations to be completed. In the case of this study, PSO was able to minimize the overall error within the first 40 or so iterations when using 40 particles.

$$V_{n+1} = w * V_n + c1 * rand() * (g_{best,n} - x_n) + c2 * rand() * (p_{best,n} - x_n) \tag{12}$$

$$x_{n+1} = x_n + V_{n+1} * \Delta t \tag{13}$$

In addition to the PSO, an artificial neural network (ANN) was incorporated after the first four iterations. The data from those iterations—and all subsequent iterations—were given to the ANN which used the data to produce a parameter set for each following iteration that seemed likely—based on the ANN computations—to result in a reduced overall error. The ANN works in the following manner. A set of input values comprise the first "layer." The final layer is comprised of the output values—target values or predictions. Any layers between the input and output layers

are called hidden layers, and they are comprised of "nodes" which are essentially processing

centers, playing a role like that of biological neurons. Layers are connected to each other, receiving

inputs from the nodes of the previous layer, and sending outputs to the next layer, except in the

cases of the input layer—which has no inputs to itself—and the output layer, which does not send

output on to any further layers. The inputs to a node are individually weighted (with weights that

can be tuned), and they are summed together and passed through an activation function to produce

the output of the node. The activation function acts like an on/off switch such that the output is

only passed on to the next layer if a certain threshold value is reached.[33] For this research study,

TensorFlow[61] (a Python library created by Google developers) was utilized for implementation of

the ANN with the exponential linear unit (ELU) activation function; an input layer, five hidden

layers, and an output layer; and dropout to help prevent overfitting. The number of input nodes

corresponded to the number of target properties for each system (e.g. targeting density, heat of

vaporization, and surface tension corresponds to three input nodes), and the number of nodes in

the output layer corresponded to the number of force field parameters to be fitted. Each hidden

layer contained fifty nodes. Incorporating the ANN into the PSO is not absolutely necessary for

achieving good results, but it can lead to a satisfactorily minimal error sooner than PSO alone and is

not in any way detrimental.[33] In some cases, it was even able to lead to more accurate predictions if

the given ranges for the PSO did not encompass the best values, thus guiding the model

development. Figure 27 illustrates how the ANN is incorporated into the optimization algorithm.
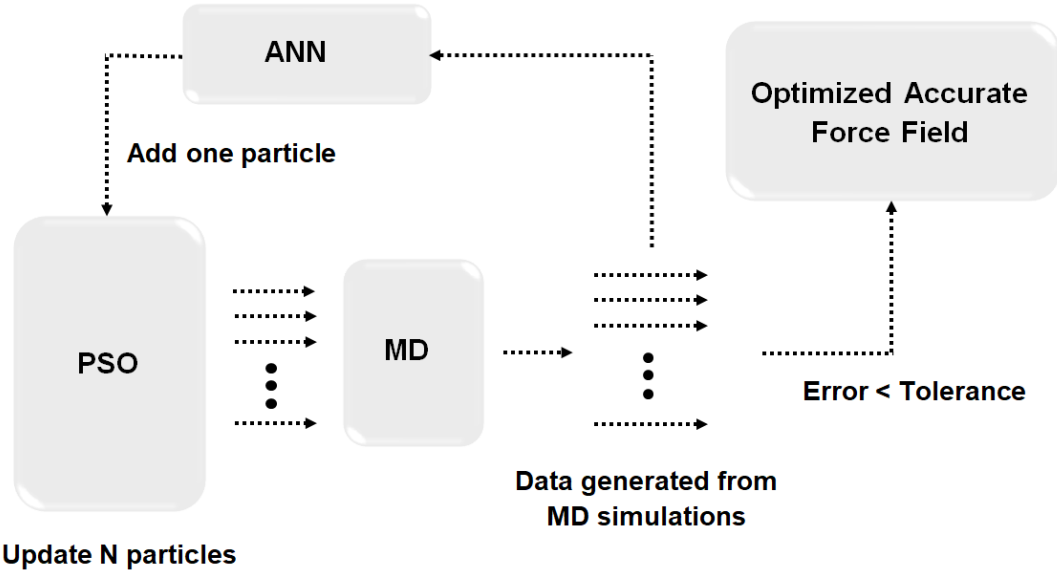
Figure 27: Schematic showing PSO and ANN incorporation with MD simulations to develop optimized force field parameters.

4.5 Calculation of Thermodynamic Properties

To obtain the density of a simulation, the number of molecules, the molecular weight of each molecule, and the simulation box size must be known. The density was calculated according to equation (14) below in which ρ, m, and V have their usual meanings of density, mass, and volume, respectively. N stands for the number of molecules in the system, MW is the molecular weight, N_A is Avogadro's number, and L_x, L_y, and L_z are the box lengths in the x, y, and z dimensions, respectively.

$$\rho = \frac{m}{V} = \frac{\left(\frac{N \times MW}{N_A}\right)}{L_x L_y L_z} \tag{14}$$

The surface tension was calculated using equation 15 in which L_z is the box length in the z dimension, and P_{zz}, P_{xx}, and P_{yy} are the pressure tensors in the z, x, and y dimensions, respectively. The simulation is run at constant volume with the box length extended in the z dimension to the extent that a region of vacuum is created above the liquid surface.

$$\gamma = \frac{L_z}{2}\left[P_{zz} - \frac{\left(P_{xx} + P_{yy}\right)}{2}\right] \tag{15}$$

The heat of vaporization (ΔH_{vap}) was calculated using equation (19). The volume of the liquid phase, V_{liq}, is negligible as compared with the volume of the gas phase, V_{gas}; and with the assumption that the gas is an ideal gas such that the kinetic energies of the molecule in the gas and liquid phases are the same, equation (17) becomes equation (18). C is a correction term that is usually small and neglected, giving equation (19) which is used for calculating the heat of vaporization from MD simulations in which T is the temperature of the simulation.[62] R is the ideal gas constant. The potential energy of the gas phase, $E_{gas}^{potential}$, is obtained by simulating one molecule in a box at constant volume to imitate the lack of intermolecular interactions experienced by a molecule in the gas phase as per the ideal gas law. The potential energy of the liquid phase, $E_{liq}^{potential}$, is obtained from the same simulation conducted to determine the density. Since the potential energy is a combination of the intermolecular and intramolecular energies, the potential energy of the gas phase can be set as zero for CG simulations in which a molecule is represented as a single bead.

$$\Delta H_{vap} = H_{vapor} - H_{liquid} \tag{16}$$

$$\Delta H_{vap} = E_{gas}(T) - E_{liq}(T) + p(V_{gas} - V_{liq}) \tag{17}$$

$$\Delta H_{vap} = E_{gas}^{potential}(T) - E_{liq}^{potential}(T) + RT + C \tag{18}$$

$$\Delta H_{vap} = E_{gas}^{potential}(T) - E_{liq}^{potential}(T) + RT \tag{19}$$

Five hundred molecules were used for the density and surface tension simulations. Each simulation—for density, surface tension, and heat of vaporization—was run for a total of 2 ns. Non-bonded interactions were calculated by excluding 1-2 interactions and with a cutoff of 12 Å. Constant temperature was simulated using the Langevin thermostat, and constant pressure simulations were conducted with the Nosé-Hoover Langevin piston barostat[56] as described in the previous section.

4.5 Results and Discussion

Prior to moving forward with setting the non-bonded parameters for the backbone and termini beads as the same across all dipeptides, a quick comparison of the AA CG mapped RDFs of the backbone and termini beads showed similarities as can be seen in Figure 28. This provided verification that it would be suitable to treat the beads as the same regardless of which amino acid they are associated with. The RDFs also indicate that the environment of a bead—which other beads are attached to the molecule and are in the vicinity, as well as their sizes and crowding effects—can influence the RDF.
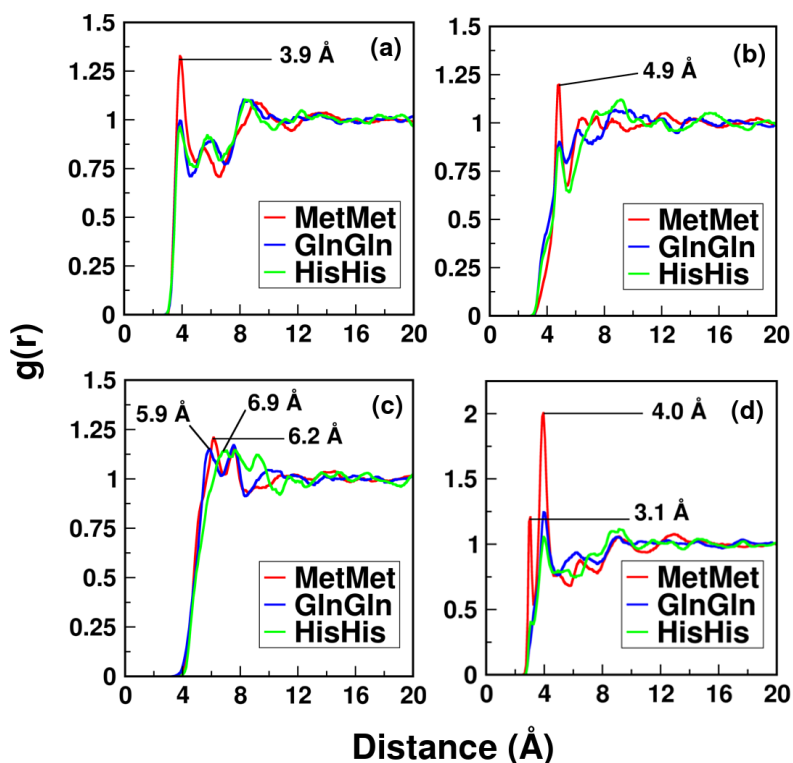
Figure 28: AA RDFs of (a) NCT, (b) CO, (c) NC, and (d) COO1 beads from the homogeneous methionine (MetMet), glutamine (GlnGln), and histidine (HisHis) dipeptides.

After the desired number of optimization cycles (100 at least with a maximum of 200, depending on the dipeptide) the top typically 1-3 sets with the lowest errors were selected for testing over ten nanoseconds. This allowed the selection of optimal non-bonded parameters for each bead type. It was also an important verification step because the top PSO result did not necessarily correlate with the top result after 10 ns simulations. While the density error remained

essentially the same for the 2 ns optimization cycle and the 10ns simulation, the heat of vaporization and surface tension errors in some cases were different such that the error showed an increase up to about 6% (e.g. from 2% error to 8% error). Additionally, we targeted a very low density error—within 2% of the experimental value; a surface tension error within 5% of the experimental value; and a heat of vaporization error as low as possible while optimizing the density and surface tension errors. The heat of vaporization error tends to be much higher, perhaps partly because the energy landscape for a CG model is much smoother than that of an AA model and the energy landscape in reality.[6]

Interestingly, the analogues methanol (for serine), ethanol (for threonine), and m-cresol (for tyrosine) all showed the worst matches to their experimental heat of vaporization values by far, with the errors for the CG models being 52% for both methanol and ethanol and 29% for m-cresol. Since these three analogues all contain a hydroxyl group--and since they are the only analogues that contain a hydroxyl group--it seems likely that the peculiar hydrogen bonding capacity of the hydroxyl group may be the cause of the extremely high heat of vaporization errors. The analogues containing an amine group do not exhibit such poor matches to the experimental heat of vaporization values despite their hydrogen bonding capacity, so in the cases of methanol, ethanol, and m-cresol, it appears to be the oxygen atom that makes the difference. However, the CG water models previously developed by Bejagam, et al.[33] do not have such high errors with respect to the heat of vaporization values. So perhaps the results shown here have not only to do with the character of the hydroxyl group, but also the character of the molecules in that the carbon-containing segments are quite different in their properties from the hydroxyl group itself. Further investigation would be useful in helping to elucidate and overcome this issue.

With respect to the RDFs—which are impacted by the non-bonded parameters, not the bonded parameters—the CG bead peaks are usually much sharper and higher than those of the AA CG-mapped peaks. Frequently, if there is a shoulder or a more gradual slope in the AA RDF initial

peak, the CG initial peak exists where the shoulder is, or in the middle of the slope, or between the shoulder and the initial peak when comparing with the AA CG-mapped distribution. The RDFs for the methionine terminal and backbone beads are shown in Figure 29, comparing the AA CG mapped distributions with the CG distributions. Methionine was chosen as the example for the backbone and terminal beads because its AA RDFs are the most pronounced, as shown in Figure 28. The results for the NCT (neutral N-terminal), CO (backbone carbonyl bead), and NC (backbone N-C_α) beads seem to indicate that the analogues chosen for the backbone and terminal beads were reasonable. The CG RDF for the COO1 (neutral C-terminal) bead (Figure 29(d)) is a rather poor match to the AA RDF. However, because the COO1 bead represents a carboxyl functional group, and the carboxyl functional group of propionic acid was used as the analogue, there is no reason to suspect that propionic acid was not a sufficiently good analogue to represent the C-terminal bead, COO1. Furthermore, the result shown is essentially the same when comparing the CG RDF of the propionic acid bead with the AA CG-mapped RDF of the propionic acid bead. This indicates that a simple coarse-graining of the carboxyl group fails to capture the AA details of the carboxyl group. Comparisons of the CG RDFs for each dipeptide side chain bead type with their AA mapped CG counterparts show that the analogue choices for those bead types were suitable as well; all such figures are given in the appendix for the reader's examination. Considering the unusual nature of the proline side chain, as discussed previously, it should be noted that the CG RDFs of the proline backbone (PNC) and side chain (PS) beads correspond well to their AA CG-mapped RDFs, as shown in Figure A18.6 in the appendix.
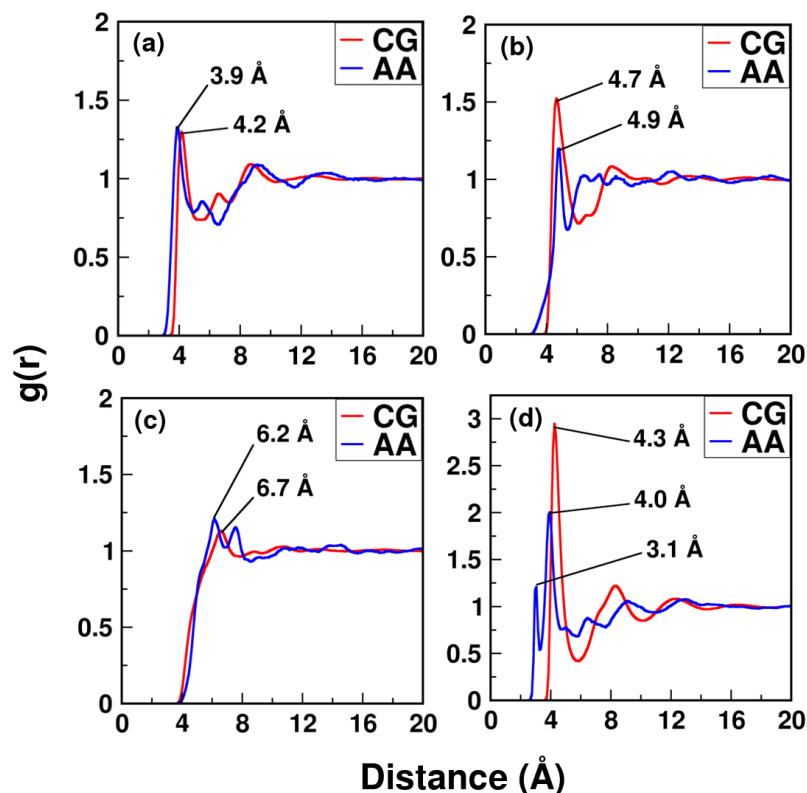
Figure 29: Comparisons of AA coarse-grain mapped (AA) RDFs vs coarse-grained (CG) RDFs of (a) NCT, (b) CO, (c) NC, and (d) COO1 from the methionine dipeptide.

Following the 10 ns simulations to finalize the non-bonded parameters, 100 ns simulations were conducted to further validate these parameters. The trajectory from each simulation was divided into segments from 50-60 ns, 60-70 ns, 70-80 ns, 80-90 ns, and 90-100 ns to obtain block average values for the density, surface tension, and heat of vaporization; these values are reported in Tables A2.1-A2.3 in the appendix and shown in Figures 30-32 below. Percent errors were calculated in accordance with equation 20.

$$\% \ error = \frac{|experimental - CG \ model|}{experimental} \tag{20}$$
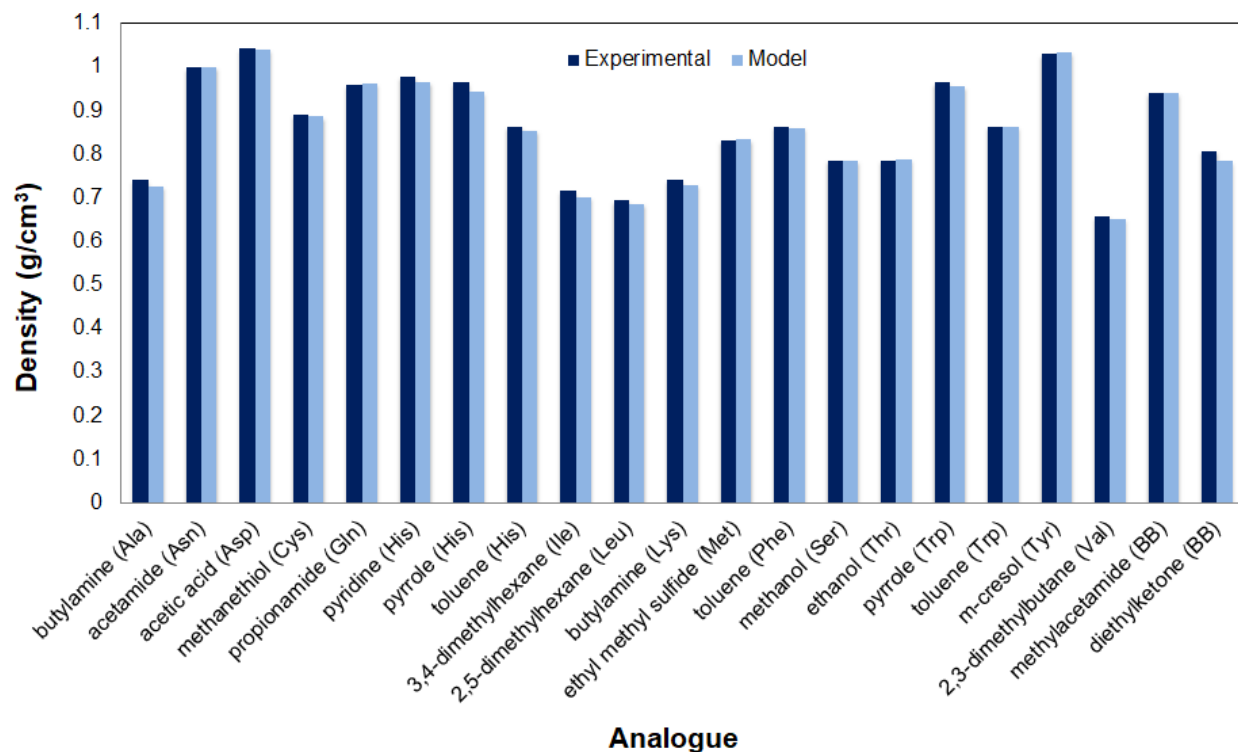
Figure 30: Comparisons of experimental and CG model densities. BB stands for backbone. Experimental data from refs [63–67].
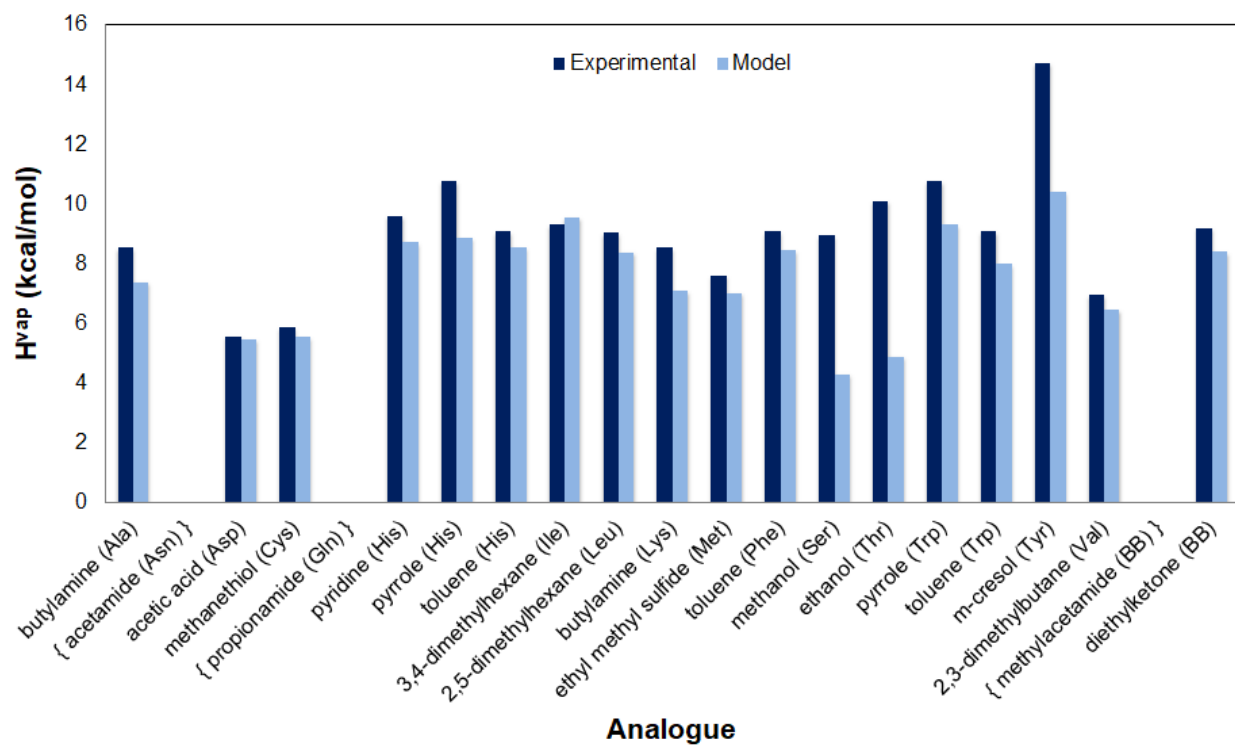
Figure 31: Comparisons of experimental and CG model heats of vaporization. BB stands for backbone. Experimental data from refs [63,65,68]. Experimental values for heat of vaporization for molecules in braces { example } were not found.

43
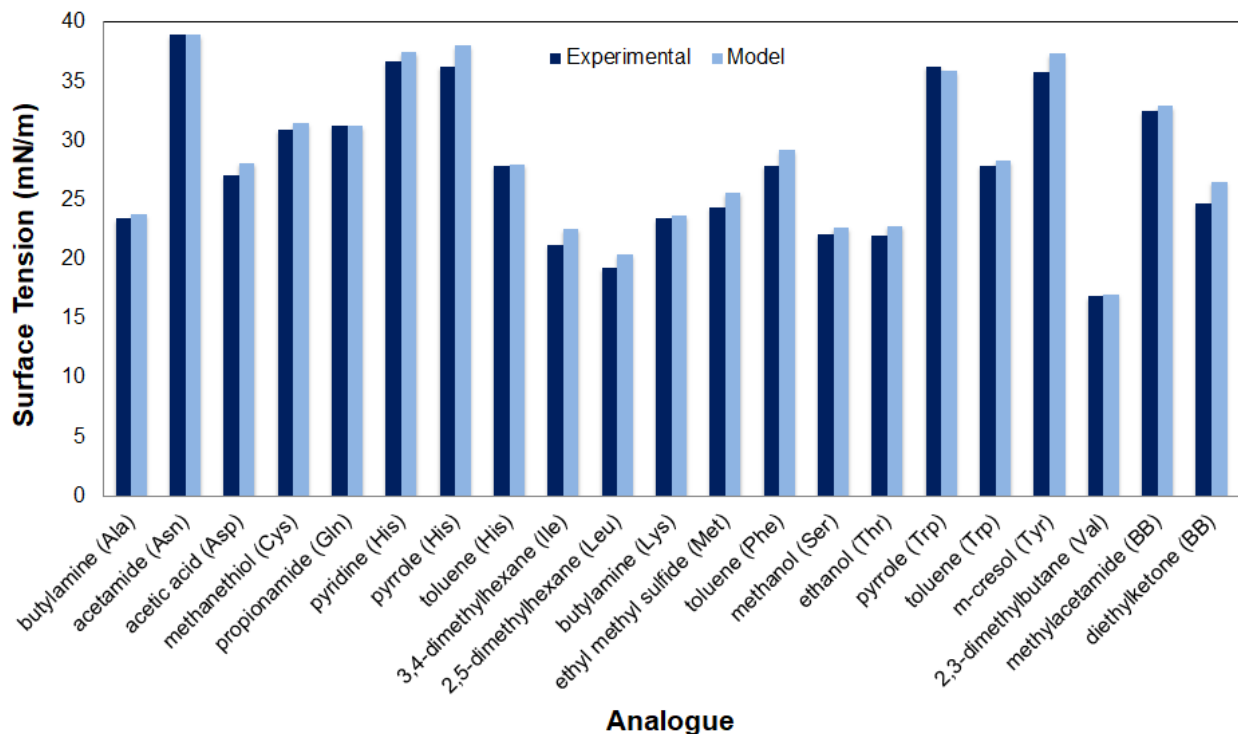
Figure 32: Bar graph comparisons of experimental and CG model surface tensions. BB stands for backbone. Experimental data from refs [63–66,68,69].

With the non-bonded parameters having been determined for the amino acid side chain and backbone beads, those values were plugged back into the dipeptide parameter files to finalize the bonded parameters and test the CG dipeptide simulation densities against the AA dipeptide simulation densities. The density results are given in Table 3 with percent errors calculated in accordance with equation 20 above with the substitution of the AA value in place of experimental. For the most part the density errors were not large, but in the case of the glycine dipeptide the error was 11.73%. As will be shown in some paragraphs that follow, this error—and that for any of the dipeptides, by extension—can vary dramatically based on the non-bonded parameters used and the exclusion criteria provided in the configuration file for a simulation. The bond, angle, and dihedral distributions, along with mapping schemes and RDFs, are included in the appendix, for both analogues and peptides; the graphs represent the best outcomes from optimization and

44

selection of the non-bonded and bonded parameters given some of the constraints described in the

paragraphs below.

Table 3: Density comparisons, all atom (AA) vs coarse-grained (CG) models

Dipeptide/Tripeptide	AA Density (g/cm³)	CG Density (g/cm³)	Percent Error
Ala-Ala-Ala	1.170	1.169	0.16
His-His	1.331	1.250	6.11
Leu-Leu	1.011	0.991	2.00
Lys-Lys	1.120	1.067	4.73
Asn-Asn	1.358	1.290	4.99
Asp-Asp	1.371	1.335	2.66
Ala-Arg	1.207	1.194	1.05
Cys-Cys	1.359	1.259	7.39
Gln-Gln	1.277	1.283	0.47
Glu-Glu	1.289	1.233	4.32
Gly-Gly	1.321	1.166	11.73
Ile-Ile	1.014	0.987	2.71
Met-Met	1.188	1.161	2.25
Phe-Phe	1.134	1.123	1.03
Val-Pro-Leu	1.062	1.021	3.91
Ser-Ser	1.337	1.245	6.87
Thr-Thr	1.179	1.193	1.24
Trp-Trp	1.218	1.244	2.11
Tyr-Tyr	1.204	1.238	2.88
Val-Val	1.039	0.977	5.98
average % error	. . . .	. . . .	*3.73*

In this final parameterization for setting the bonded parameters, there were four primary

matters of importance that had to be taken into consideration and addressed. The first was that CG

distributions are much more uniform than AA distributions often are. While an AA distribution

may have, for example, two peaks in a bond distribution, the CG distribution will have only one.

This is also true of the angle distributions. The dihedral distributions may have more than one peak in the CG distribution, but in some cases many different attempts to find a matching dihedral for the CG model failed, and the only outcome that could be obtained was a dihedral that was shifted (as in figure A10.5(e)) or a non-existent (flat) dihedral (as in Figure A9.5(e)). Good matches could be obtained for many of the dihedrals, however, as shown in the dihedral figures in the appendix.

The second matter of note is that higher force constants require a smaller time step because a stiffer force constant engenders a higher vibrational frequency which correlates with smaller time steps.[49] This meant that in cases such as the bonds and angles in ring structures (e.g. the side chain of Histidine), while the all-atom distribution might have required a very high force constant to match it, the final force constant for each bond and angle given as the CG parameter was lower to help ensure stability at a time step of 5 femtoseconds or greater. The lower the time step, the longer it takes for the simulation to reach microseconds in length; therefore it is preferable to have as high a time step as seems to be reasonable considering the parameterization. Figure 33 shows, as an example, the AA and CG distributions for a bond and an angle in the Histidine side chain.
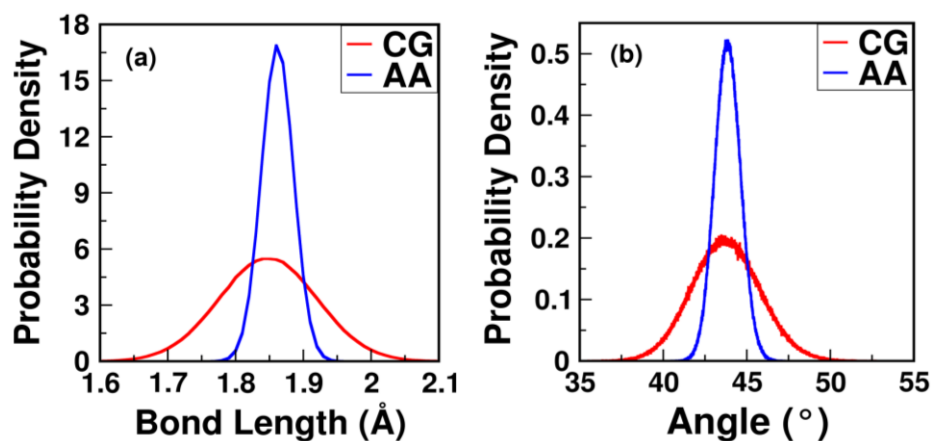
Figure 33: Comparisons of all-atom (AA) mapped and coarse-grained (CG) distributions in (a) bond NCR2-NCR1 and (b) angle NCR2-CCR1-NCR1 in the histidine side chain.

A third matter of importance was to parameterize the bonded interactions in such a way as to harmonize them across the backbones of the different dipeptides. As was mentioned previously, there are certain backbone bonds and angles such as CO-NC and CO-NC-COO1 that were in most of

the dipeptide simulations, and we decided to consider all the dipeptide distributions and attempt to harmonize them to obtain the most reasonable parameters. As it turned out, this was not challenging to do since the common bonds and angles were largely consistent across dipeptides as can be seen in Figures 34, 35, and 36 below.
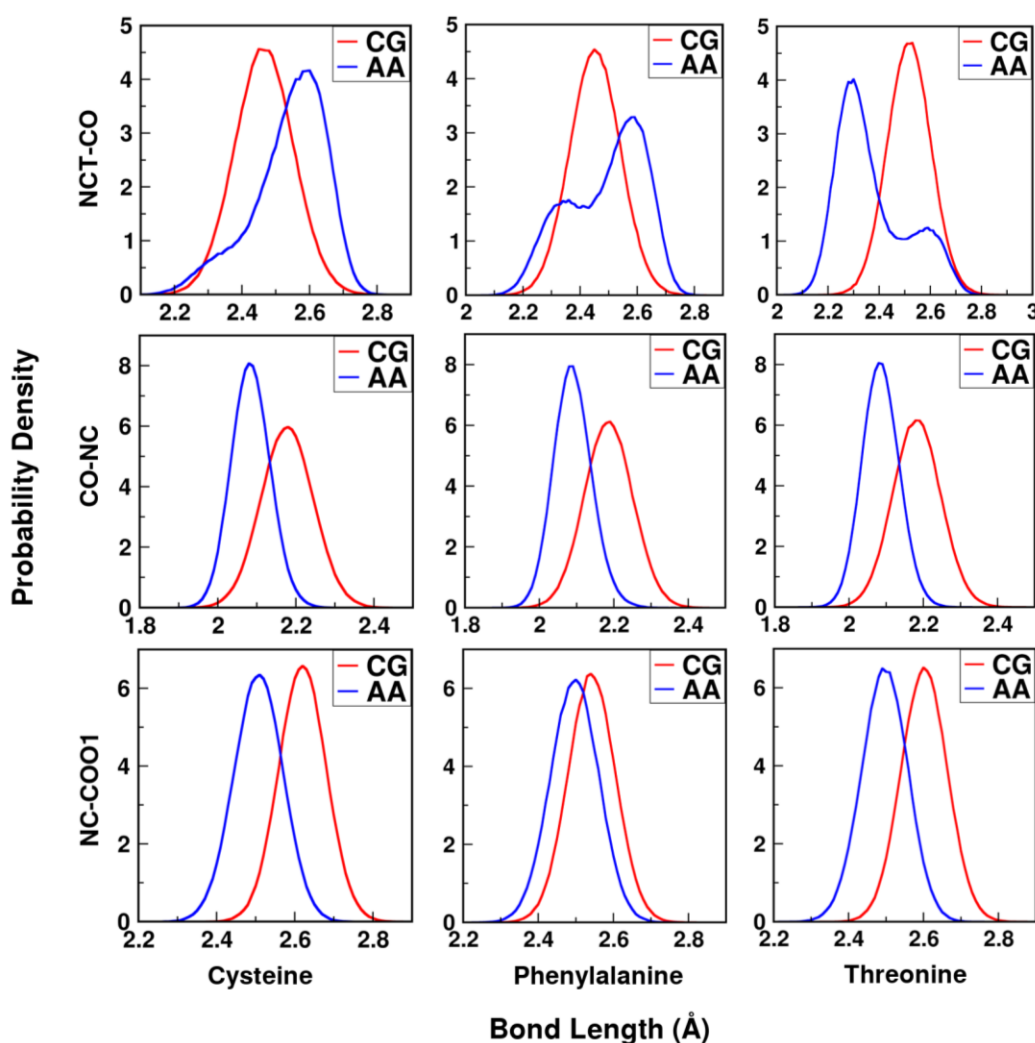
Figure 34: Comparisons of backbone bond distributions for the bonds NCT-CO (top row), CO-NC (middle row), and NC-COO1 (bottom row) from the cysteine (first column), phenylalanine (second column), and threonine (third column) dipeptides.
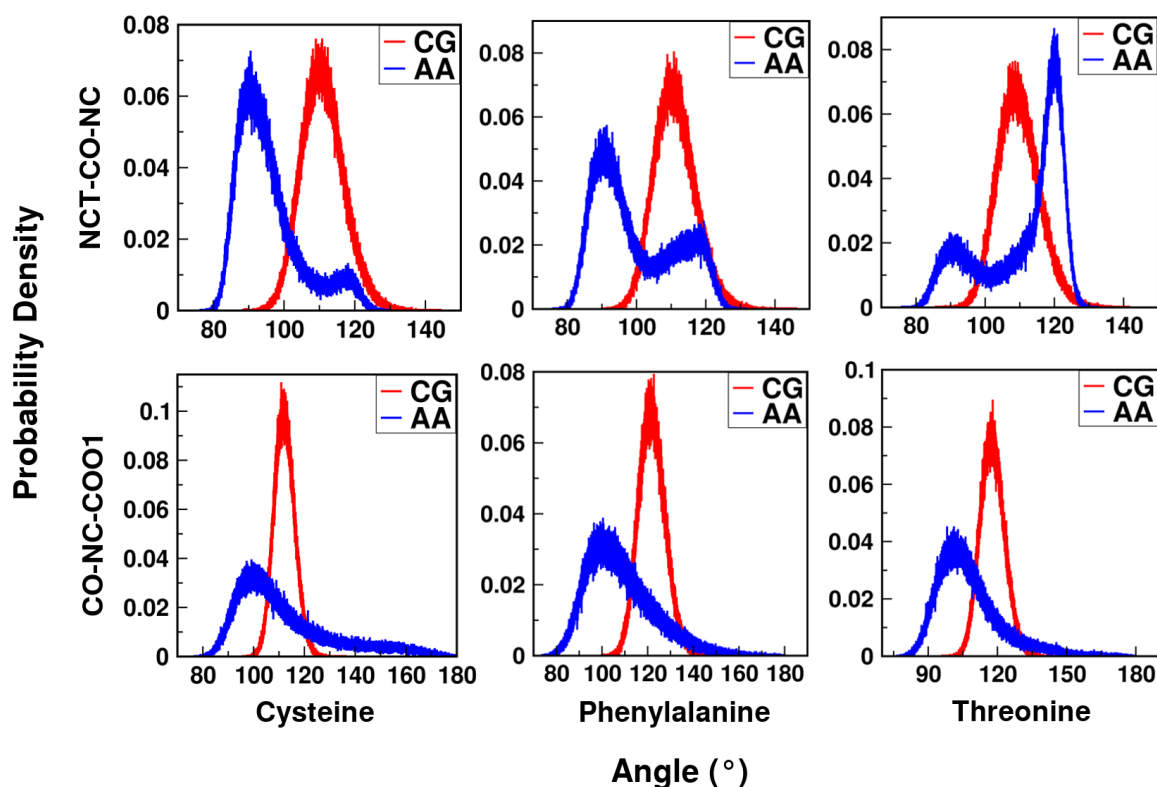
Figure 35: Comparison of backbone angle distributions for the angles NCT-CO-NC (top row) and CO-NC-COO1 (bottom row) from the cysteine (first column), phenylalanine (second column), and threonine (third column) dipeptides.
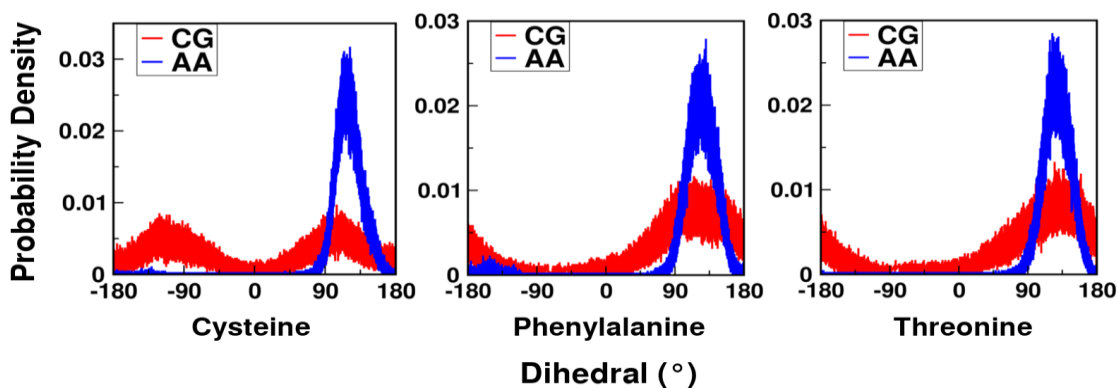
Figure 36: Comparison of backbone NCT-CO-NC-COO1 distribution from the cysteine (left), phenylalanine (center), and threonine (right) dipeptides.

The fourth issue of note is that some shifting of bonds and angles and some narrowing of

angle distributions were noticed in the CG distributions as compared with the AA distributions.

Upon investigation, it became apparent that the primary cause of this was from non-bonded

interactions of beads within a molecule that were within two or three bonds of each other, since we

had only excluded interactions between beads directly bonded to each other ("exclude 1-2" in the

configuration file). When excluding interactions between beads that were within two bonds of each

other ("exclude 1-3"), the shifting and narrowing largely disappeared to allow much better overlap

between the CG and AA distributions. This was also true of using "exclude scaled1-4" in the

configuration file. Figures 37, 38, and 39 show examples of this from the leucine dipeptide; the

bonded and non-bonded parameters used were the same for both CG models, but they do not

represent the final parameters chosen for the CG model. Problematically, to use "exclude 1-3" or

"exclude scaled1-4" often leads to greater instability in simulations and thus requires a smaller time

step. Since, then, we were still able to maintain good overlap in many cases—as can be seen in the

figures in the appendix—and only very little overlap in some, we decided to continue the

parameterization with "exclude 1-2" at least to the point of attempting simulations of PA self-

assembly using our CG models to obtain some indication of the amino acid bead and water bead

interactions and how they might need to be adjusted. Later model development focusing on

improving stability in the case of "exclude 1-3" or "exclude scaled1-4" could yield overall

improvements in the model while enabling useful speed of simulations. Because the all-atom

distributions have already been obtained, it would also be quite simple to generate the appropriate

parameters for the case of "exclude 1-3" or "exclude scaled1-4." The non-bonded parameters

would not need to be changed except for in two cases: 1) the CG bead representing the peptide

backbone carbonyl, as that was parameterized based on a 3-bead mapping of diethylketone, and 2)

the C2E2 bead for isoleucine as that was parameterized based on the four-bead mapping scheme

shown in Figure 14(b). Some of the distributions (e.g. bond and angle distributions within ring

structures such as that of the phenylalanine side chain) are not likely to differ based on the

exclusion criteria.
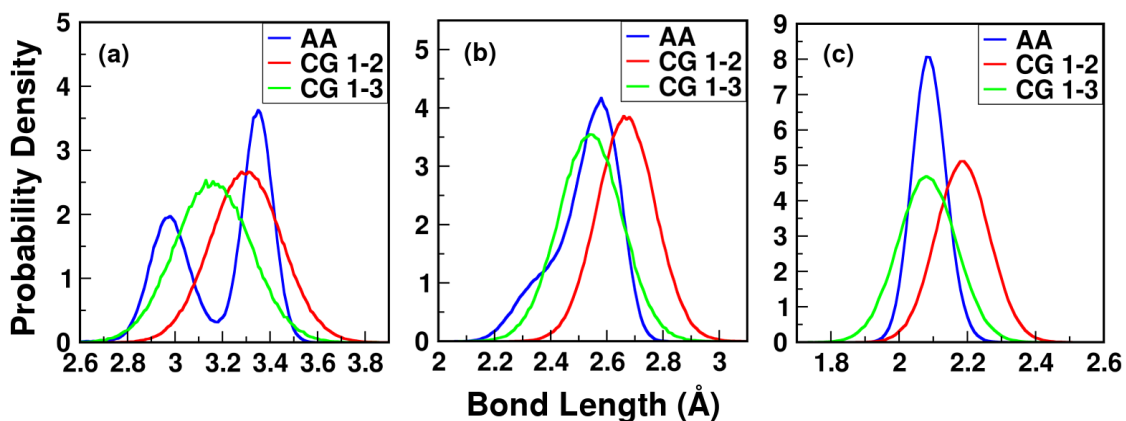
Figure 37: Comparisons of bond distributions, from the leucine dipeptide, between all-atom mapped (AA), coarse-grained using "exclude 1-2," (CG 1-2) and coarse-grained using "exclude 1-3" (CG 1-3). The same bonded and non-bonded parameters were used for both CG distributions. Distributions are of (a) C41-NCT, (b) NCT-CO, and (c) CO-NC.
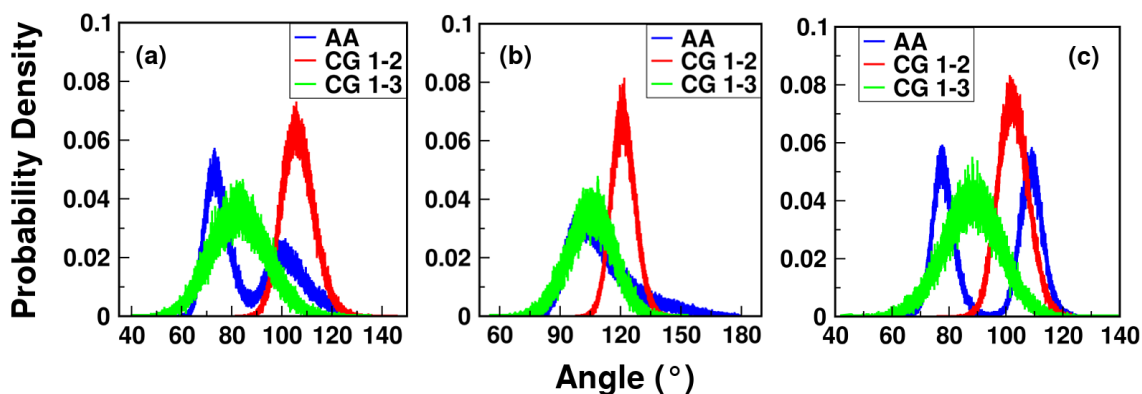
Figure 38: Comparisons of angle distributions, from the leucine dipeptide, between all-atom mapped (AA), coarse-grained using "exclude 1-2," (CG 1-2) and coarse-grained using "exclude 1-3" (CG 1-3). The same bonded and non-bonded parameters were used for both CG distributions. Distributions are of (a) C41-NCT-CO, (b) CO-NC-COO1, and (c) C41-NC-COO1.
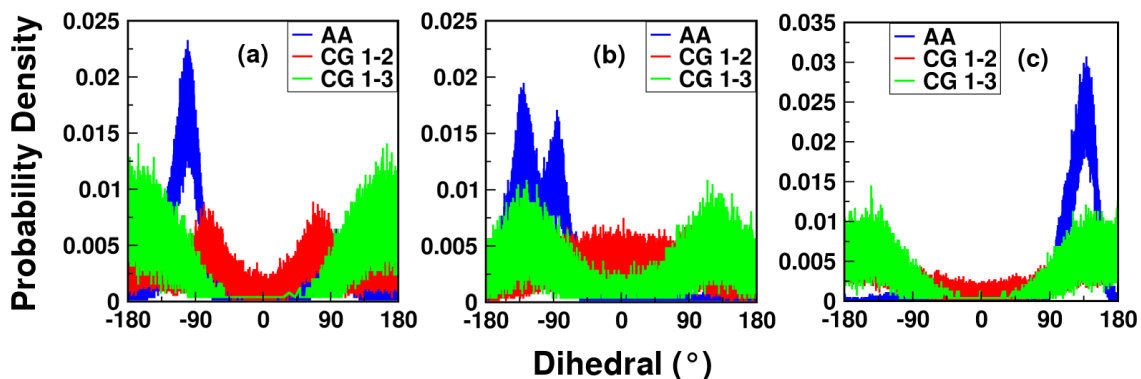
Figure 39: Comparisons of dihedral distributions, from the leucine dipeptide, between all-atom mapped (AA), coarse-grained using "exclude 1-2," (CG 1-2) and coarse-grained using "exclude 1-3" (CG 1-3). The same bonded and non-bonded parameters were used for both CG distributions. Distributions are of (a) C41-NCT-CO-NC, (b) NCT-CO-NC-C41, and (c) NCT-CO-NC-COO1.

50

To test these hypotheses, the non-bonded parameters for the diethylketone beads were optimized under the "exclude 1-3" constraint. The optimized non-bonded parameters for the carbonyl bead were then plugged into glycine, histidine, alanine, and leucine CG dipeptide simulations to optimize the bonded parameters for those dipeptides under the "exclude 1-3" constraint. The densities of those simulations with optimized bonded and non-bonded parameters were then compared with the AA simulation densities of those dipeptides. The optimized non-bonded parameter values for the CO bead are given in Table 4. Table 5 shows the results of the CO bead parameterization with "exclude 1-2" (from a 100 ns simulation) versus "exclude 1-3" (from a 10 ns simulation; if the simulation were extended to 100ns, the results ought to be similar). Table 6 compares the density results for AA versus CG "exclude 1-2" versus CG "exclude 1-3" for the test dipeptides GlyGly, HisHis, AlaAla, and LeuLeu. Figure 40 compares the RDFs of the AA CG-mapped, CG "exclude 1-2," and CG "exclude 1-3" optimized CO bead in diethylketone (Figure 40(a)) and the leucine dipeptide (Figure 40(b)). Figures 41 and 42 provide examples of optimized bond and angle distributions, respectively, of the leucine dipeptide for the cases of CG "exclude 1-2" and CG "exclude 1-3" in comparison with the AA CG-mapped distributions. These distributions were improved in most cases. The poorness of the match in Figure 41(c) for CG 1-3 may be partly due to keeping the force constants low to enhance the stability. If the stability for CG 1-3 were able to be improved (this possibility will be discussed in Section 6 regarding future work), then possibly the force constant for that bond could be increased, leading to a better match.

Table 4: Optimized CO bead non-bonded parameters for CG "exclude 1-3" (CG 1-3) and CG "exclude 1-2" (CG 1-2)

	CG 1-3	CG 1-2
$r_{min}/2$	2.0967	2.4361
epsilon	-0.9472	-0.6690

Table 5: Comparisons of density, heat of vaporization, and surface tension results and errors for optimized CG with "exclude 1-3" parameterization (CG 1-3) and optimized CG with "exclude 1-2" parameterization (CG 1-2) diethylketone simulations with the experimental values (Exp.)

	results			% errors	
	Exp.	**CG 1-3**	**CG 1-2**	**CG 1-3**	**CG 1-2**
density	0.8089	0.7908	0.7843	2.24	3.04
H^{vap}	9.2065	8.8903	8.4050	3.43	8.71
Surface tension	24.74	25.89	26.57	4.66	7.38

Table 6: Density results for select dipeptide and tripeptide simulations: AA vs CG "exclude 1-3" (CG 1-3) vs CG "exclude 1-2" (CG 1-2). The CO bead as parameterized using CG "exclude 1-3" was used for all the CG 1-3 results.

	results			% errors	
Peptide	**AA**	**CG 1-3**	**CG 1-2**	**CG 1-3**	**CG 1-2**
AlaAlaAla	1.17043	1.35137	1.1685	15.46	0.16
GlyGly	1.32142	1.32188	1.16645	0.03	11.73
HisHis	1.33100	1.38957	1.2497	4.40	6.11
LeuLeu	1.01118	1.08326	0.990916	7.13	2.00
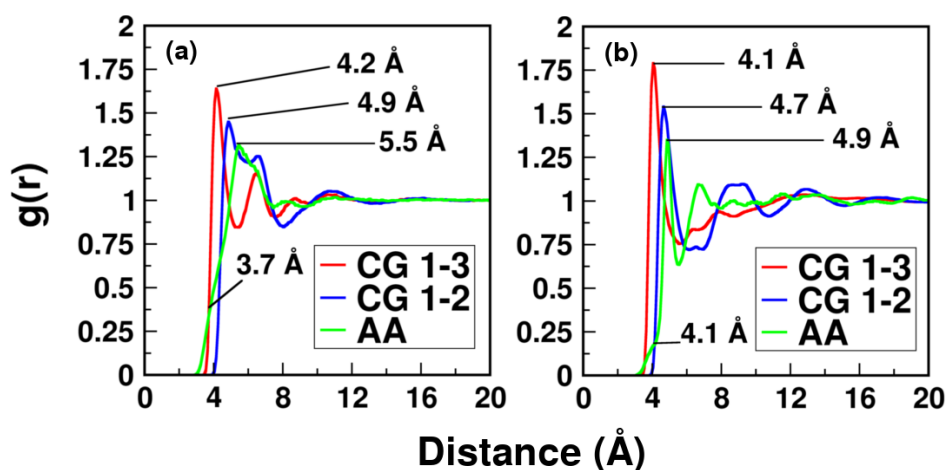
Figure 40: RDFs of the CO bead in (a) diethylketone and (b) the leucine dipeptide for the AA CG-mapped bead (AA), CG as parameterized with "exclude 1-2" (CG 1-2), and CG as parameterized with "exclude 1-3" (CG 1-3).
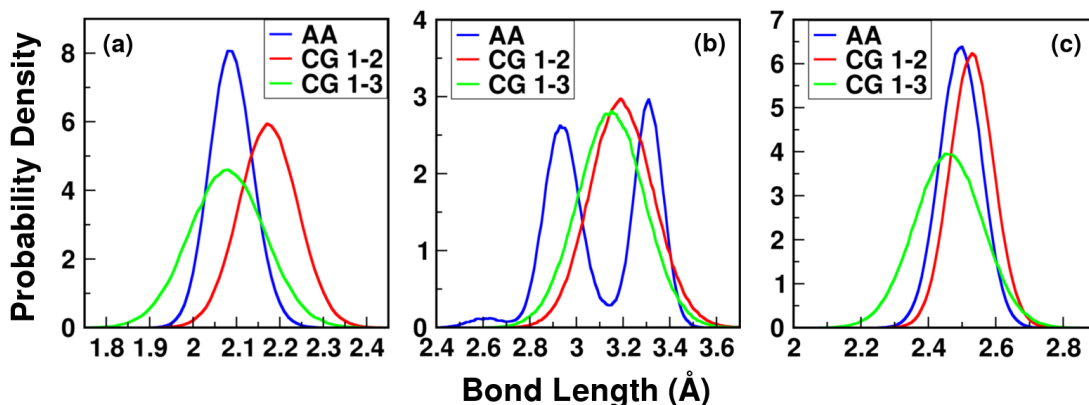
Figure 41: Distributions from the leucine dipeptide for the bonds (a) NCT-CO, (b) CO-NC, and (c) NC-C41 comparing the AA CG-mapped distribution (AA) with the CG distribution from parameterization with "exclude 1-2" (CG 1-2) and with the CG distribution from parameterization with "exclude 1-3" (CG 1-3).
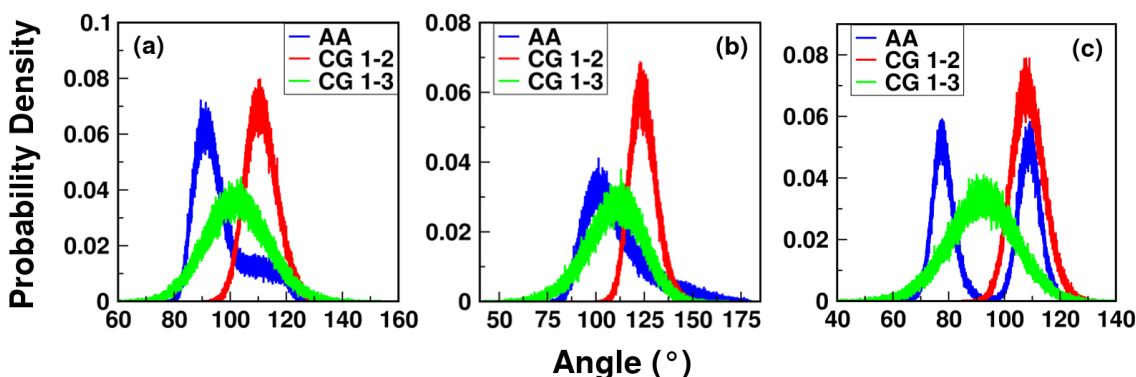
Figure 42: Distributions from the leucine dipeptide for the angles (a) NCT-CO-NC, (b) CO-NC-COO1, and (c) C41-NC-COO1 comparing the AA CG-mapped distribution (AA) with the CG distribution from parameterization with "exclude 1-2" (CG 1-2) and with the CG distribution from parameterization with "exclude 1-3" (CG 1-3).

Because of the challenges experienced with respect to "exclude 1-2" versus "exclude 1-3" or "exclude scaled1-4," alternatives to using diethylketone to parameterize the CO bead may lead to better outcomes. The most likely candidates include dimethylformamide and potentially N-methylacetamide. Dimethylformamide has the advantage of a carbonyl bonded to an amide group, which is similar to the situation in a peptide backbone, although the amide of dimethylformamide is bonded to two carbons and no hydrogens rather than one carbon and one hydrogen as is the case in a peptide backbone; thus there is no potential for hydrogen bonding in dimethylformamide. N-methylacetamide may be a possibility if the methyl group bonded to the carbonyl were to be modeled as a separate bead. However, the simulations would require a much smaller time step (e.g.

1 or 2 fs) because of the smallness of the methyl bead. The advantage to using N-methylacetamide, however, is that the backbone NC and backbone CO beads could be parameterized simultaneously, and the parameterization would include the effects of those beads in proximity to each other that is essentially the same as what occurs in a peptide backbone.

5. PA Self-Assembly Studies

5.1 Initial Self-Assembly Studies

As a first step in testing the amino acid models, we determined to attempt to simulate the self-assembly of a peptide amphiphile (PA). The PA chosen for this study was c16-AHL$_3$K$_3$-CO$_2$H shown in Figure 43; the CG model of the PA is shown in Figure 44. This is the PA that was simulated and synthesized in the study by Deshmukh, et al.,[4] examining the role of water in PA self-assembly. A combination of the Martini and CHARMM (with TIP3P water model) force fields were used for that study; the initial 150 ns of the simulation were with the CHARMM AA model which was mapped to the CG Martini model for the following ~16 µs, and then at the end of the 16 µs, the system was back-mapped to the CHARMM model for a final 150 ns to show atomic-level details. The simulation included 300 PAs for a simulated duration of 16 µs at 1 bar and 340 K. The PAs self-assembled into micelles in less than 1 µs which assembled into fibers within approximately 1.5-5 µs with some ongoing fiber breakage and formation. The simulated micelles were comprised of approximately 50 PAs, and the simulated fibers were approximately 50 Å in diameter. These results served as a guide in our amino acid model development.
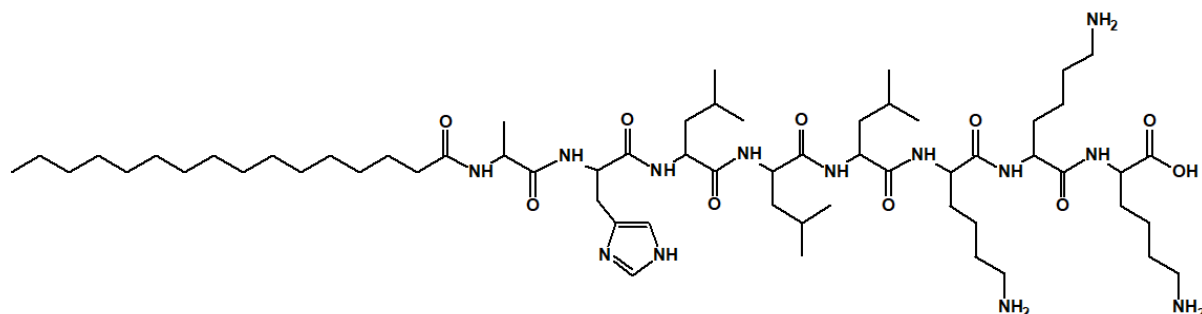
Figure 43: Structure of c16-AHL$_3$K$_3$-CO$_2$H

In order to keep the required simulation time to a minimum while still enabling the possibility of PA fiber formation, we chose to model 150 PAs in an initial box size of 108 cubic Å with 16,000 of the 1-site CG waters.[33] These numbers are consistent with those used in other PA self-assembly studies.[24,70] Periodicity of the MD simulations enabled the possibility of fiber formation from PAs on one side of the box interacting with the periodic images of PAs on the other side of the box. Simulations were conducted with NAMD and a pressure setpoint of 1 bar and a temperature setpoint of 340 K using the Nosé-Hoover Langevin piston barostat and the Langevin thermostat for pressure and temperature control. Intramolecular interactions were excluded for beads bonded to each other ("exclude 1-2"). A 12 Å cutoff was used with a switching distance of 9 Å and a pairlist distance of 16 Å. Simulations were minimized for 10,000 steps and then conducted for a total of 15 μs with a 5 fs timestep.
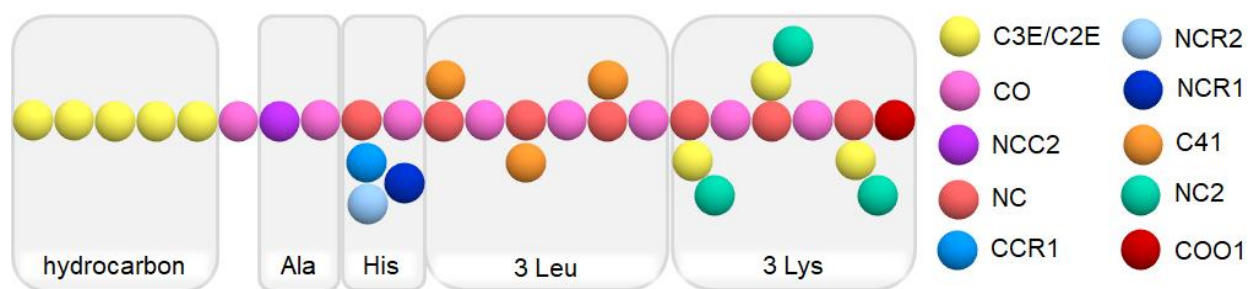
Figure 44: CG structure of c16-AHL$_3$K$_3$-CO$_2$H showing bead types

To begin with, we ran a PA simulation using the Lorentz-Berthelot (LB) combining rules for all the beads excepting the C3E and C2E beads which were given the interaction parameters with

water that had previously been developed for them.[48] As can be seen in Figure 44, the hydrocarbon tail was thus tuned with respect to water while the peptide portion primarily was not. The LB combining rules relate to the LJ functional form as shown in equation 21 in which ε is the depth of the potential well, σ is the distance at which the potential between particles is zero, and r is the distance between particles. The Lorentz rule is given in equation 22, and the Berthelot rule is given in equation 23. After running the simulation for a few microseconds, however, it became apparent that defined, individual fibers were not forming. Within just a few nanoseconds, the PAs clumped together into micelle-like groups as shown in the 5 ns snapshot for the 100% simulation in Figure 46, in which the percentage reflects that the LB combining rules were not scaled for this simulation. Those micelle-like groups quickly bundled together to form windowpane structures which persisted throughout the 15 μs simulation, shown in Figure 46. Continually throughout the simulation, the assemblies broke apart (lost contiguity especially with respect to the hydrocarbon interiors) and regrouped. The 15 μs snapshot for the 100% simulation shows the PA assemblies in the middle of regrouping.

$$V_{LJ} = 4\varepsilon \left[\frac{\sigma^{12}}{r} - \frac{\sigma^6}{r} \right] \tag{21}$$

$$\sigma_{ij} = \frac{\sigma_{ii} + \sigma_{jj}}{2} \tag{22}$$

$$\varepsilon_{ij} = \sqrt{\varepsilon_{ii}\varepsilon_{jj}} \tag{23}$$

Because the 100% simulation was not behaving with the kind of character that experimental and MD studies indicate is typical for PA fiber self-assemblies, as described in Section 3.2, we decided to run two more simulations: one in which the epsilon interaction values were scaled to 85% of the combining rule values, and another in which the epsilon interaction values were scaled to 115% of the combining rule values. In neither case was the sigma interaction value scaled. This was because it is the epsilon value that controls the depth of the LJ potential well and

describes the strength of attraction between two beads. In the case of the 115% simulation, the PAs showed clumping into micelle-like groups with the hydrocarbon tails bundling together. However, after more than a microsecond in simulation time, the PA peptide segments remained rather diffuse and showed no tendency to form fibers, as shown in Figure 45. Thus the strength of attraction of the peptide segment for the water beads was too strong. It also is unlikely that fibers would have formed, given the close proximity of the PAs to each other; if the peptide segments had preferred associating together rather than remaining somewhat diffuse in the water, the microsecond of simulation time provided ample opportunities to show such behavior. The 85% simulation, on the other hand, did show fiber formation. In this case, however, the fibers materialized immediately with no real micelle stage. Figure 46 shows that at just 5 ns, the PAs of the 85% simulation were already in a rough fiber formation. By the time 300 ns had been simulated, there were already neat fibers. These fibers remained stable throughout the remainder of the simulation. The measured diameter was approximately 58 Å at 15 μs.
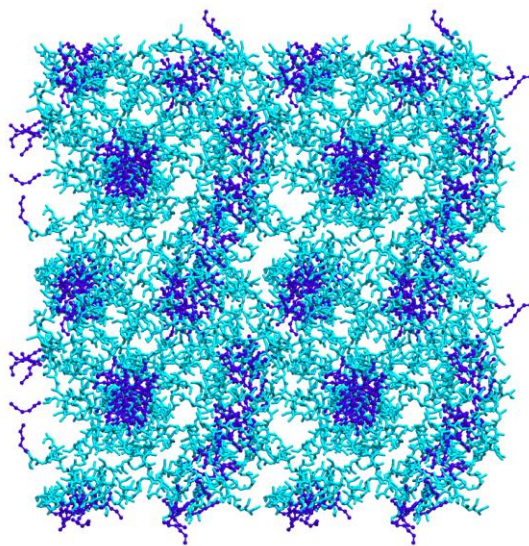
115% - 1.2 μs

Figure 45: Simulation results after 1.2 μs with amino acid bead epsilon interaction values with water scaled to 115% of the combining rule values.

Because of the behavior exhibited with both the 100% and 85% simulations, we decided to run another simulation in which the epsilon interaction values were scaled to 90% of the

combining rule values, and another in which the epsilon interaction values were scaled to 95% of the combining rule values. The 90% simulation results were very similar to the results of the 85% simulation, with no real micelle formation and essentially immediate fiber formation. The fiber diameter was measured at approximately 58 Å at 15 μs in this case as well. The 95% simulation behaved very similarly to the 100% simulation.

The formation of fibers with an increase in hydrophobicity makes sense considering that past research has shown the importance of hydrophobicity in fiber formation.[27,28] However, the immediacy of formation and lack of a micelle stage suggest that the model may require further refinement, possibly by tuning the interaction parameters of different amino acid beads with each other. To help determine whether the scaled epsilon values were on the right track for the individual beads, hydration free energies were calculated for the beads, side chain groupings, or analogues, as described in the following paragraphs, for comparison with experimental values.

5.2 Hydration Free Energies

In order to provide insight into the realistic strength or weakness of the interactions between water and the amino acid beads utilized in the $c16$-AHL_3K_3-CO_2H PA, the potential of mean force (PMF) was calculated for water-solute systems in which the solute was representative of a bead or an analogue in the PA. In NAMD, it is relatively easy to calculate the PMF of a system using the adaptive biasing force (ABF) method. To use this method to determine the Gibbs hydration free energy of a molecule, a 52 cubic Å box of 2000 CG waters was first equilibrated. Then a single solute molecule was placed approximately 20 Å outside the box in one dimension. The ABF will act along the reaction coordinate which, in our case, was defined as the projection in the z-direction of the distance between the center of mass of the water box and that of the solute molecule. The free energy along this coordinate can be considered as a potential that can be calculated from the average force acting along that coordinate, with the average force being the negative gradient of the potential (as in equation 3). Thus integrating the average force yields the potential. The

instantaneous force that acts along the reaction coordinate is actually the sum of the average force—which only depends on the value of the coordinate—and a random force that has a zero average and which can be approximated as diffusive. The running time average of the instantaneous force is calculated during a simulation which provides an estimate of the derivative of the free energy at each point on the path that the solute molecule travels. At the same time that the average force is recorded, an equal and opposite external biasing force—the ABF—is applied so as to cancel that average force. This biasing force will stabilize at values close to zero when the average force is at an equilibrium value. At that point, there is an approximately flat potential of mean force which enables the solute molecule to easily move along and explore the coordinate when otherwise free energy barriers would have prohibited such exploration as the solute molecule would have gotten trapped in some states for lengths of time that could certainly exceed the duration of the simulation.[71]

In NAMD, the ABF method is implemented using the colvars package.[72,73] Two windows were used: a window of 22 Å to 40 Å and another of 5 Å to 22 Å; the water box was centered at 0 Å. Two windows were chosen in this manner to allow for sufficient exploration of the reaction coordinate for 50ns simulations. For window 2, the initial configuration file was taken from the window 1 configuration shortly after the simulation had begun and at a point at which the solute molecule had moved into the interface between the water and vacuum. The 50ns simulation for each window was used to obtain the PMFs. CG amino acid side chain structures and analogues were used: N-butylamine for alanine, N-butylamine for lysine, 4-methylimidazole for histidine, isobutane for leucine, formaldehyde for the backbone carbonyl, and N-methylacetamide for the backbone NC bead. The 1-site CG water model was used for the water. The combining rules were used for the amino acid bead and water bead interaction parameters, except for the C2 and C3 hydrocarbon beads for which the already-developed interaction parameters[48] were used. Simulations were run at 298 K and 1 bar for comparison with experimental values at the same

conditions. The results are shown in Figure 47 and Table 7. The differences in the N-butylamine results occur because of the different mapping schemes. These free energy calculations are a ballpark approximation to help indicate whether the amino acid bead interaction parameters with water are approximately correct, or in which direction they should move (higher for greater attraction to water, or lower for less attraction to water). The results indicate that the attraction of the amino acid beads for water is too strong for the leucine side chain bead, pretty decent for the CO bead (represented by formaldehyde), and not strong enough for the remainder. Considering these results in light of the PA simulation results, it seems likely that while the individual beads may experience stronger attractions to water than the combining rules provide for, there are also stronger interactions between different amino acid beads to each other than the combining rules provide. In order to achieve the hydrophobicity necessary for PA self-assembly, it may thus be necessary to consider the strength of the attractions between different amino acid beads.

Table 7: Experimental and CG MD values of hydration free energies, ΔG_{hyd}, for select molecules.

Molecule	ΔG_{hyd}^{exp}[a] (kcal/mol)	ΔG_{hyd}^{CG} (kcal/mol)
N-butylamine (Ala)	-9.52	-3.0
N-butylamine (Lys)	-9.52	-0.4
4-methylimidazole (His)	-10.27	-7.0
isobutane (Leu)	2.28	-1.0
formaldehyde (CO)	-2.77	-1.5
N-methylacetamide (NC)	-10.07	-5.0

[a] ΔG_{hyd}^{exp} values from refs [74–76].
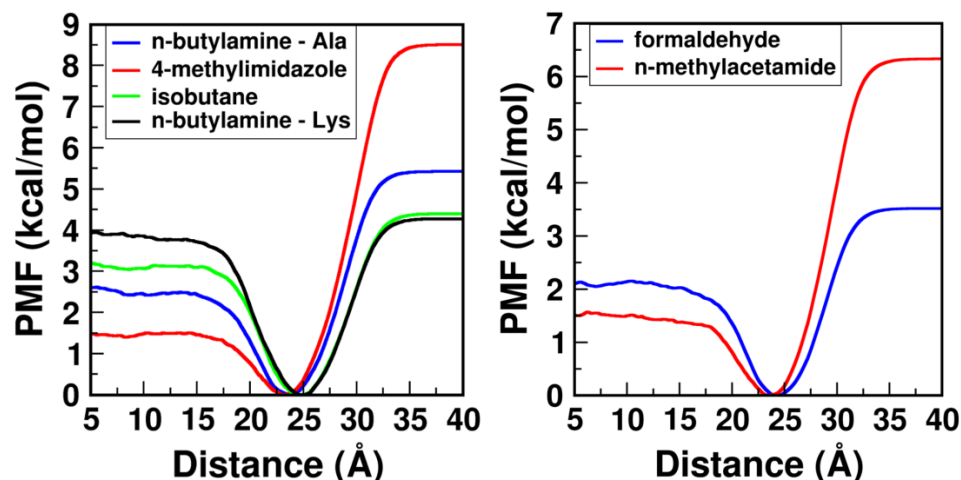
Figure 47: Free energy results for side chain and backbone CG structures and analogues

5.3 Final Self-Assembly Study

Because of the results from the hydration free energy studies, one more PA simulation was conducted. For this simulation, denoted "varied," in Figure 46, all the amino acid bead interaction parameters were scaled up to 115% of the combining rule values except for the CCR1 histidine bead which was neither scaled up nor down, the backbone carbonyl (CO) bead which was scaled up to 105% of the combining rule value, and the leucine side chain bead (C41) which was scaled down to 70% of the combining rule value in order to better approximate the epsilon interaction parameter of a hydrocarbon bead with water. The epsilon value of CCR1 was not scaled since it represents two sp[2] hybridized carbons and no nitrogen or carbon atoms, and the CO bead was only scaled up slightly since the hydration free energy results appeared to indicate that the combining rule was approximately correct for the carbonyl.

This simulation in which scaling was varied has the advantage of showing micelle formation, but as seen in the 15 μs snapshot in Figure 46, the PAs remained as micelles and never formed fibers. Again, given the proximity of the PA micelles to each other, it seems unlikely that continuing the simulation for a longer period of time would have resulted in fiber formation. Additionally, research published using the Martini model to simulate the c16-AHL$_3$K$_3$-CO$_2$H PA showed fiber formation within the first few microseconds,[4] although it is possible that such a time
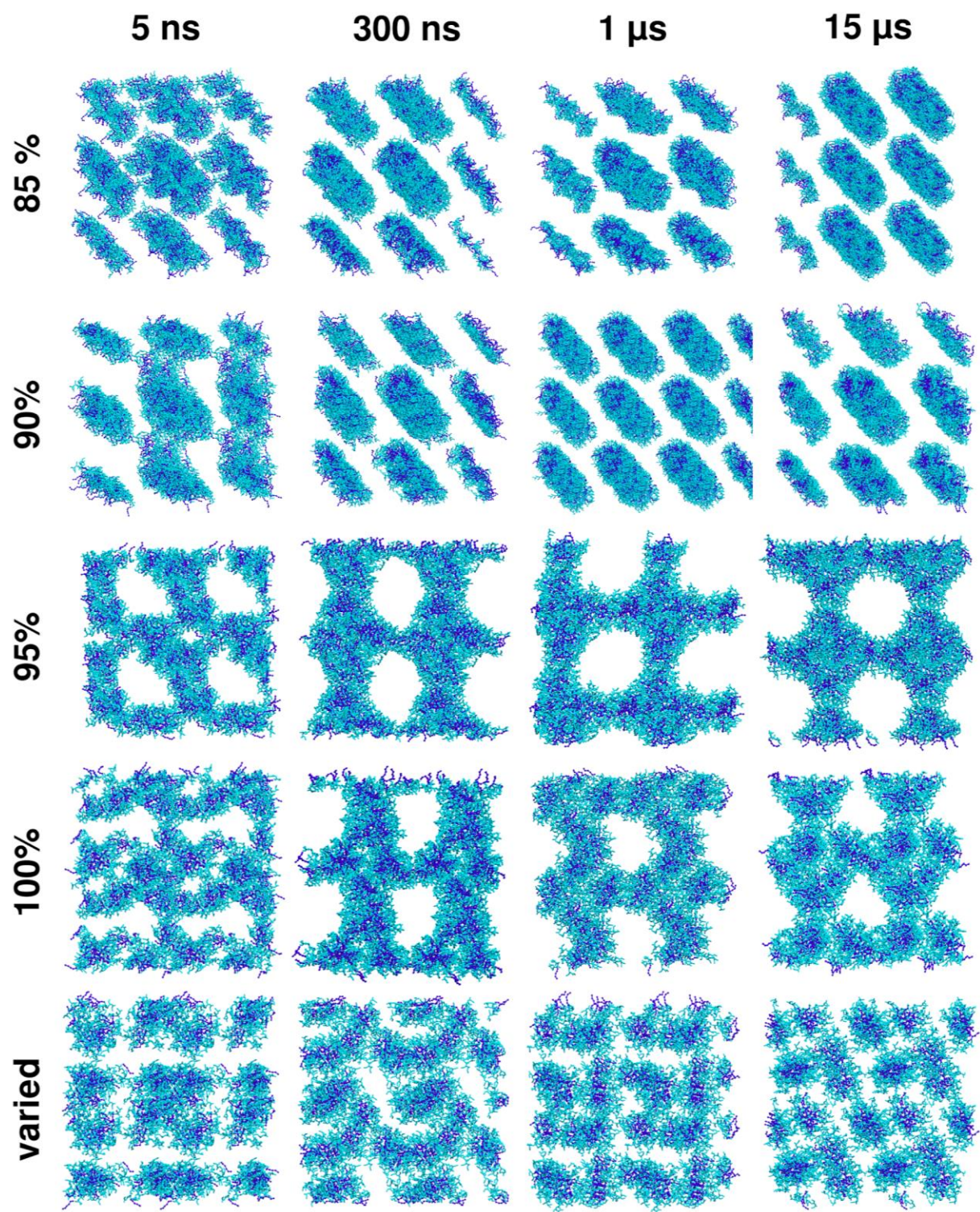
Figure 46: Snapshots from simulations of PAs (blue) in water (not shown). Darker blue - hydrocarbon tails; lighter blue - peptide segments. Percentages reflect percent of combining rule epsilon values while the "varied" simulation possessed primarily 115% of combining rule epsilon values with two bead types scaled to 70% and 105% and a third bead type left with its combining rule value.

frame is not realistic. Considering that the 115% simulation shows clumping only of the hydrocarbon tails while the "varied" simulation shows more defined micelle formation, it becomes clear that it is important to consider the individual chemistry of each bead based on the atoms and functional group(s) the bead represents. This is especially clear considering the hydration free energy study results and when considering that the hydrocarbon leucine side chain bead (C41) certainly would interact with water differently than the beads which possess atoms with the capability of forming hydrogen bonds.

6. Conclusions & Future Work

At the beginning of the project of which this manuscript provides a record, it was desired to 1) develop mapping schemes for the amino acids and peptide backbone with intermediate resolution (a preference for 2:1 and 3:1 beads), 2) optimize non-bonded parameters for the beads based on experimental thermodynamic values of analogues, 3) define appropriate bonded parameters for the amino acids and peptides based on all-atom dipeptide simulation bond, angle, and dihedral distributions, and 4) use the non-bonded and bonded parameters thus developed for c16-AHL$_3$K$_3$-CO$_2$H PA simulation studies with the goal of witnessing fiber formation as a way to test the model and provide an indication of the direction needed for future model development. The mapping scheme was defined without too much trouble; only in two cases (tryptophan and tyrosine) were we required to modify the mapping scheme for viability. The optimized non-bonded parameters that were developed provided, for the most part, good agreement with experimental results. Parameterization of methanol, ethanol, m-cresol, and diethylketone (with "exclude 1-2") partially fell short in this regard and may require more work and different tactics to attempt to improve. The bonded parameters developed resulted in CG bond, angle, and dihedral distributions that are shown throughout the Appendix. In some cases the matches were excellent, mostly they were reasonable, and in some cases there was little overlap between the CG and AA CG-mapped

distributions. The reason for incidents of poor overlap, as discussed, was identified primarily as a result of the 1-2 exclusion criteria provided in the CG configuration file. This could easily be remedied, but with the catch that the result would be an overall much less stable model with the implication that it may not be feasible in some cases to conduct simulations—unless, of course, some other fix was found for stabilizing the model. With respect to the PA simulations, the results showed that use of the LB combining rules fails to capture the behavior expected based on experimental and MD research. When the overall hydrophobicity was increased, fibers formed; yet scaling down the interactions between the amino acid beads and water seemed, for the most part, the wrong direction to take based on the hydration free energies the beads exhibited in comparison with experimental values. Most likely, an overall increase in the hydrophobicity combined with sensitivity to the particular chemistry each bead represents would be required for more physically realistic models. This may be better accomplished by increasing the strength of attraction of some of the amino acid beads types for each other than by decreasing the strength of attraction of some of the amino acid bead types for the water bead. The hydration free energies may be useful in direction the scaling of the interaction parameters of the amino acid beads with the water bead. Obtaining guidance for tuning the interaction parameters of the amino acid beads with each other will be discussed below.

While this work represents great progress made in creating a model for amino acids and peptides, there are a number of challenges and limitations that this base model faces and which could be targeted for improvement. As previously noted at various points in this manuscript, these challenges and limitations include 1) lack of charges to represent charged amino acids; 2) the amino acid bead interactions have not been fine-tuned, and currently they rely on the LB combining rules, which may be inadequate in some cases; 3) simulations can be run at a 5 fs time step, but they are somewhat unstable and will frequently crash after only a few million steps; 4) no explicit modeling of hydrogen bonding and no attempts to parameterize for secondary structure

conformations; and 5) our model shows that the bond and angle distributions do not always match the AA bond and angle distributions very well, with shifting and narrowing being observed in many cases. Potential solutions to these challenges and limitations will be addressed individually in the paragraphs below.

(1) The amino acids exist as zwitterions at neutral pH, with a positive charge on the N-terminus and a negative charge on the C-terminus. There are also various pKa values for many of the amino acid side chains, such as histidine and glutamic acid. While in laboratory experiments a buffer is sometimes added to neutralize an amino acid sequence, the reality in many cases is going to be that most or all of the amino acids that are capable of possessing a charge are going to possess a charge in the experimental conditions. For this reason, it is important for a protein force field to contain parameters for the amino acids in these charged states, as is the case with the AA force field CHARMM and the CG Martini and SDK force fields, for example. The charges in the SDK model were developed by parameterization against guanidium salt solutions at various concentrations, and the model does not distinguish lysine from arginine or aspartic acid from glutamic acid.[77] The Martini model appears to have used AA simulations of the charged residues for parameterizing the charged beads. I suggest that for the model described in this manuscript, AA simulations of charged residues or analogues using the CHARMM force field be used to parameterize the beads that represent any atoms that could exist in a charged state.

(2) Tuning the interaction parameters of the amino acid beads with each other could be challenging. In the Martini model, amino acid beads were classified into four main types (polar, nonpolar, apolar, and charged) with sub-classifications based on hydrogen bonding capabilities or degree of polarity. Amino acid beads were assigned a classification based on water/oil partitioning coefficients of analogues of the side chains. The beads were then assigned interaction parameters with other types in accordance with the interaction matrix given in the original Martini publication.[21] Something similar could be undertaken for the models developed in this manuscript.

65

(3) As long as the model contains beads that contains only two heavy atoms (a 2:1 mapping scheme), that in and of itself will partially restrict the ability of the model to run smoothly without crashing at larger time steps, since the mass of a bead with a 2:1 mapping scheme is typically lower which influences the stability.[49] Without artificially inflating the mass of a bead, an option that could potentially result in improved stability would be to fix to a given length the bonds of ring structures (as in the side chains of histidine, phenylalanine, tyrosine, and tryptophan). These bonds tend to have very small ranges over which they vibrate, as shown in Figure 33, instead of using a somewhat high force constant for those bonds, replacing the harmonic bonds and angles of the ring with a set of constraints could be a potential work-around. This is what was done in the Martini model.[44] Additionally, a careful analysis of the angles and dihedrals could reveal situations in which heavy fluctuations are likely to happen (when three of the four beads in a dihedral have a very wide angle, close to 180 degrees). In that case, the options of removing any dihedrals deemed unnecessary or using special forms of dihedrals could be evaluated and implemented.[78]

(4) Tuning the amino acid bead interaction parameters with each other would be a way to capture the potential of hydrogen bond formation between particular amino acid beads. This may be especially important for the backbone CO and NC beads since it is the ability of the carbonyl oxygen to hydrogen bond with the hydrogen of the backbone nitrogen that leads to secondary structures like α helices and β sheets.

(5) The shifts and narrowing of bond and angle distributions was shown to result from using "exclude 1-2" while the distributions were much improved when using "exclude 1-3" and "exclude scaled1-4." While the latter two tend to lead to greater instability and the requirement of a smaller time step than when applying "exclude 1-2," it could be an option to use "exclude 1-3" or "exclude scaled1-4" if any of the potential methods for stabilization—as addressed in (3) above— are effective.

Another possibility for future work is to conduct CG MD simulations for some of the analogues at different temperatures to test how well they reproduce experimental values (density, heat of vaporization, and surface tension) at the same temperatures. Depending on how well the models were able to reproduce the experimental values, re-parameterization could be considered to try to obtain a better fit across a temperature range if desired. Another option would be to define particular parameters for particular temperatures for overall accuracy for a wide range of conditions.

With such efforts directed towards improving the amino acid models, the work described in this manuscript has the potential to result in CG models ideal for studying PA self-assembly that, in combination with the hydrocarbon[47] and water[33] models previously developed by the group, may lead to more physically accurate simulations than other protein models currently available. Already, this work has shown great promise in this respect. And because of the potential for PAs to be useful in especially biomedical applications, such efforts may be quite rewarding.

References

(1)   Kokkoli, E.; Mardilovich, A.; Wedekind, A.; Rexeisen, E. L.; Garg, A.; Craig, J. A. Self-Assembly and Applications of Biomimetic and Bioactive Peptide-Amphiphiles. *Soft Matter* **2006**, *2* (12), 1015–1024. https://doi.org/10.1039/b608929a.

(2)   Cui, H.; Webber, M. J.; Stupp, S. I. Self-Assembly of Peptide Amphiphiles: From Molecules to Nanostructures to Biomaterials. *Biopolymers* **2010**, *94* (1), 1–18. https://doi.org/10.1002/bip.21328.

(3)   Eskandari, S.; Guerin, T.; Toth, I.; Stephenson, R. J. Recent Advances in Self-Assembled Peptides: Implications for Targeted Drug Delivery and Vaccine Engineering. *Adv. Drug Deliv. Rev.* **2017**, *110–111*, 169–187. https://doi.org/10.1016/J.ADDR.2016.06.013.

(4)   Deshmukh, S. A.; Solomon, L. A.; Kamath, G.; Christopher Fry, H.; S Sankaranarayanan, S. K. Water Ordering Controls the Dynamic Equilibrium of Micelle–Fibre Formation in Self-Assembly of Peptide Amphiphiles. *Nat. Commun.* **2016**, *7*, 1–11. https://doi.org/10.1038/ncomms12367.

(5)   Brini, E.; Algaer, E. A.; Ganguly, P.; Li, C.; Rodríguez-Ropero, F.; Van Der Vegt, N. F. A. Systematic Coarse-Graining Methods for Soft Matter Simulations-a Review. *Soft Matter* **2013**, *9* (7), 2108–2119. https://doi.org/10.1039/c2sm27201f.

(6)   Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chem. Rev.* **2016**, *116* (14), 7898–7936. https://doi.org/10.1021/acs.chemrev.6b00163.

(7)   Ingólfsson, H. I.; Lopez, C. A.; Uusitalo, J. J.; de Jong, D. H.; Gopal, S. M.; Periole, X.; Marrink, S. J. The Power of Coarse Graining in Biomolecular Simulations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4* (3), 225–248. https://doi.org/10.1002/wcms.1169.

(8)   Berg JM, Tymoczko JL, S. L. *Biochemistry*, 5th ed.; W. H. Freeman: New York, 2002.

(9)   *Biology 2e*, 2nd ed.; OpenStax CNX, 2018.

(10)   Boekhoven, J.; Zha, R. H.; Tantakitti, F.; Zhuang, E.; Zandi, R.; Newcomb, C. J.; Stupp, S. I. Alginate–Peptide Amphiphile Core–Shell Microparticles as a Targeted Drug Delivery System. *RSC Adv.* **2015**, *5* (12), 8753–8756. https://doi.org/10.1039/C4RA16593D.

(11)   Bulut, S.; Erkal, T. S.; Toksoz, S.; Tekinay, A. B.; Tekinay, T.; Guler, M. O. Slow Release and Delivery of Antisense Oligonucleotide Drug by Self-Assembled Peptide Amphiphile Nanofibers. *Biomacromolecules* **2011**, *12* (8), 3007–3014. https://doi.org/10.1021/bm200641e.

(12)   Silva, G. A.; Czeisler, C.; Niece, K. L.; Beniash, E.; Harrington, D. A.; Kessler, J. A.; Stupp, S. I. Selective Differentiation of Neural Progenitor Cells by High-Epitope Density Nanofibers. *Science (80-.).* **2004**, *303* (5662), 1352–1355. https://doi.org/10.1126/SCIENCE.1093783.

(13)   Tysseling-Mattiace, V. M.; Sahni, V.; Niece, K. L.; Birch, D.; Czeisler, C.; Fehlings, M. G.; Stupp, S. I.; Kessler, J. A. Self-Assembling Nanofibers Inhibit Glial Scar Formation and Promote Axon Elongation after Spinal Cord Injury. *J. Neurosci.* **2008**, *28* (14), 3814–3823. https://doi.org/10.1523/JNEUROSCI.0143-08.2008.

(14) Qian, Y.; Kaur, K.; Foster, J. C.; Matson, J. B. Supramolecular Tuning of H_2S Release from Aromatic Peptide Amphiphile Gels: Effect of Core Unit Substituents. *Biomacromolecules* **2019**, *20* (2), 1077–1086. https://doi.org/10.1021/acs.biomac.8b01732.

(15) Matsui, H.; MacCuspie, R. Metalloporphyrin Nanotube Fabrication Using Peptide Nanotubes as Templates. *Nano Lett.* **2001**, *1* (12), 671–675. https://doi.org/10.1021/NL0156080.

(16) Fry, H. C.; Garcia, J. M.; Medina, M. J.; Ricoy, U. M.; Gosztola, D. J.; Nikiforov, M. P.; Palmer, L. C.; Stupp, S. I. Self-Assembly of Highly Ordered Peptide Amphiphile Metalloporphyrin Arrays. *J. Am. Chem. Soc.* **2012**, *134* (36), 14646–14649. https://doi.org/10.1021/ja304674d.

(17) Garifullin, R.; Erkal, T. S.; Tekin, S.; Ortaç, B.; Gürek, A. G.; Ahsen, V.; Yaglioglu, H. G.; Elmali, A.; Guler, M. O. Encapsulation of a Zinc Phthalocyanine Derivative in Self-Assembled Peptide Nanofibers. *J. Mater. Chem.* **2012**, *22* (6), 2553–2559. https://doi.org/10.1039/C1JM14181C.

(18) Huang J, M. A. J. CHARMM36 All-Atom Additive Protein Force Field: Validation Based on Comparison to NMR Data. *J. Comput. Chem.* **2013**, *34* (25), 2135–2145. https://doi.org/10.1002/jcc.23354.

(19) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117* (19), 5179–5197. https://doi.org/10.1021/ja00124a002.

(20) Robertson, M. J.; Tirado-Rives, J.; Jorgensen, W. L. Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. *J. Chem. Theory Comput.* **2015**, *11* (7), 3499–3509. https://doi.org/10.1021/acs.jctc.5b00356.

(21) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Peter, D.; Marrink, S.-J. The MARTINI Coarse-Grained Force Field: Extension to Proteins. **2008**, 819–834. https://doi.org/10.1021/ct700324x.

(22) Cheon, M.; Chang, I.; Hall, C. K. Extending the PRIME Model for Protein Aggregation to All 20 Amino Acids. *Proteins Struct. Funct. Bioinforma.* **2010**, *78* (14), 2950–2960. https://doi.org/10.1002/prot.22817.

(23) Devane, R.; Shinoda, W.; Moore, P. B.; Klein, M. L.; Pennsyl, V. Transferable Coarse Grain Nonbonded Interaction Model for Amino Acids. *J. Chem. Theory Comput.* **2009**, *5*, 2115–2124.

(24) Lee, O. S.; Cho, V.; Schatz, G. C. Modeling the Self-Assembly of Peptide Amphiphiles into Fibers Using Coarse-Grained Molecular Dynamics. *Nano Lett.* **2012**, *12* (9), 4907–4913. https://doi.org/10.1021/nl302487m.

(25) Shimada, T.; Sakamoto, N.; Motokawa, R.; Koizumi, S.; Tirrell, M. Self-Assembly Process of Peptide Amphiphile Worm-Like Micelles. *J. Phys. Chem. B* **2012**, *116* (1), 240–243. https://doi.org/10.1021/jp209105z.

(26) Yu, T.; Schatz, G. C. Free-Energy Landscape for Peptide Amphiphile Self-Assembly: Stepwise versus Continuous Assembly Mechanisms. *J. Phys. Chem. B* **2013**, *117* (45), 14059–14064. https://doi.org/10.1021/jp409305e.

(27) Paramonov, S. E. .; Jun, H.-W.; Hartgerink, J. D. Self-Assembly of Peptide–Amphiphile Nanofibers: The Roles of Hydrogen Bonding and Amphiphilic Packing. *J. Am. Chem. Soc.* **2006**, *128* (22), 7291–7298. https://doi.org/10.1021/JA060573X.

(28) Velichko, Y. S. .; Stupp, Samuel I.; Olvera de laCruz, M. Molecular Simulation Study of Peptide Amphiphile Self-Assembly. *J. Phys. Chem. B* **2008**, *112* (8), 2326–2334. https://doi.org/10.1021/JP074420N.

(29) Lee, O.-S.; Stupp, S. I.; Schatz, G. C. Atomistic Molecular Dynamics Simulations of Peptide Amphiphile Self-Assembly into Cylindrical Nanofibers. *J. Am. Chem. Soc* **2011**, *133*, 3677–3683. https://doi.org/10.1021/ja110966y.

(30) Ortony, J. H.; Newcomb, C. J.; Matson, J. B.; Palmer, L. C.; Doan, P. E.; Hoffman, B. M.; Stupp, S. I. Internal Dynamics of a Supramolecular Nanofibre. *Nat. Mater.* **2014**, *13* (8), 812–816. https://doi.org/10.1038/nmat3979.

(31) Frederix, P. W. J. M.; Patmanidis, I.; Marrink, S. J. Molecular Simulations of Self-Assembling Bio-Inspired Supramolecular Systems and Their Connection to Experiments. *Chem. Soc. Rev.* **2018**, *47* (10), 3470–3489. https://doi.org/10.1039/c8cs00040a.

(32) Manandhar, A.; Kang, M.; Chakraborty, K.; Tang, P. K.; Loverde, S. M. Molecular Simulations of Peptide Amphiphiles. *Org. Biomol. Chem.* **2017**, *15* (38), 7993–8005. https://doi.org/10.1039/C7OB01290J.

(33) Bejagam, K. K.; Singh, S.; An, Y.; Berry, C.; Deshmukh, S. A. PSO-Assisted Development of New Transferable Coarse-Grained Water Models. *J. Phys. Chem. B* **2018**, *122* (6), 1958–1971. https://doi.org/10.1021/acs.jpcb.7b10542.

(34) Davtyan, A.; Schafer, N. P.; Zheng, W.; Clementi, C.; Wolynes, P. G.; Papoian, G. A. AWSEM-MD: Protein Structure Prediction Using Coarse-Grained Physical Potentials and Bioinformatically Based Local Structure Biasing. *J. Phys. Chem. B* **2012**, *116* (29), 8494–8503. https://doi.org/10.1021/jp212541y.

(35) Bereau, T.; Deserno, M. Generic Coarse-Grained Model for Protein Folding and Aggregation. *J. Chem. Phys.* **2009**, *130* (23). https://doi.org/10.1063/1.3152842.

(36) Kolinski, A. Protein Modeling and Structure Prediction with a Reduced Representation. *Acta Biochim. Pol.* **2004**, *51* (2), 349–371. https://doi.org/035001349.

(37) Hills, R. D.; Lu, L.; Voth, G. A. Multiscale Coarse-Graining of the Protein Energy Landscape. *PLoS Comput. Biol.* **2010**, *6* (6), 1–15. https://doi.org/10.1371/journal.pcbi.1000827.

(38) Sterpone, F.; Melchionna, S.; Tuffery, P.; Pasquali, S.; Mousseau, N.; Cragnolini, T.; Chebaro, Y.; St-Pierre, J. F.; Kalimeri, M.; Barducci, A.; et al. The OPEP Protein Model: From Single Molecules, Amyloid Formation, Crowding and Hydrodynamics to DNA/RNA Systems. *Chem. Soc. Rev.* **2014**, *43* (13), 4871–4893. https://doi.org/10.1039/c4cs00048j.

(39) Pasi, M.; Lavery, R.; Ceres, N. PaLaCe: A Coarse-Grain Protein Model for Studying Mechanical Properties. *J. Chem. Theory Comput.* **2013**, *9* (1), 785–793. https://doi.org/10.1021/ct3007925.

(40)   Kar, P.; Gopal, S. M.; Cheng, Y. M.; Predeus, A.; Feig, M. PRIMO: A Transferable Coarse-Grained Force Field for Proteins. *J. Chem. Theory Comput.* **2013**, *9* (8), 3769–3788. https://doi.org/10.1021/ct400230y.

(41)   Das, R.; Baker, D. Macromolecular Modeling with Rosetta. *Annu. Rev. Biochem.* **2008**, *77* (1), 363–382. https://doi.org/10.1146/annurev.biochem.77.062906.171838.

(42)   Liwo, A.; Baranowski, M.; Czaplewski, C.; Gołaś, E.; He, Y.; Jagieła, D.; Krupa, P.; Maciejczyk, M.; Makowski, M.; Mozolewska, M. A.; et al. A Unified Coarse-Grained Model of Biological Macromolecules Based on Mean-Field Multipole-Multipole Interactions. *J. Mol. Model.* **2014**, *20* (8), 2306. https://doi.org/10.1007/s00894-014-2306-5.

(43)   Basdevant, N.; Borgis, D.; Ha-Duong, T. Modeling Protein–Protein Recognition in Solution Using the Coarse-Grained Force Field SCORPION. *J. Chem. Theory Comput.* **2013**, *9* (1), 803–813. https://doi.org/10.1021/ct300943w.

(44)   Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; De Vries, A. H. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* **2007**, *111* (27), 7812–7824. https://doi.org/10.1021/jp071097f.

(45)   Uusitalo, Jaakko J.; Ingólfsson, Helgi I.; Akhshi, Parisa; Tieleman, D. Peter; Marrink, S. J. Martini Coarse-Grained Force Field: Extension to DNA. *J. Chem. Theory Comput.* **2015**, *11* (8), 3932–3945. https://doi.org/10.1016/j.bpj.2017.05.043.

(46)   Kang, M.; Cui, H.; Loverde, S. M. Coarse-Grained Molecular Dynamics Studies of the Structure and Stability of Peptide-Based Drug Amphiphile Filaments. *Soft Matter* **2017**, *13* (42), 7721–7730. https://doi.org/10.1039/C7SM00943G.

(47)   An, Y.; Bejagam, K. K.; Deshmukh, S. A. Development of New Transferable Coarse-Grained Models of Hydrocarbons. *J. Phys. Chem. B* **2018**, *122* (28), 7143–7153. https://doi.org/10.1021/acs.jpcb.8b03822.

(48)   An, Y.; Bejagam, K. K.; Deshmukh, S. A. Development of Transferable Nonbonded Interactions between Coarse-Grained Hydrocarbon and Water Models. *J. Phys. Chem. B* **2019**, *123* (4), 909–921. https://doi.org/10.1021/acs.jpcb.8b07990.

(49)   Choe, J. I.; Kim, B. Determination of Proper Time Step for Molecular Dynamics Simulation. *Bull. Korean Chem. Soc.* **2000**, *21* (4), 419–424.

(50)   Fitch, C. A.; Platzer, G.; Okon, M.; Garcia-Moreno, B. E.; McIntosh, L. P. Arginine: Its PKa Value Revisited. *Protein Sci.* **2015**, *24* (5), 752–761. https://doi.org/10.1002/pro.2647.

(51)   Frisch, M. J. et al. Gaussian 09 Revision D.01. Gaussian Inc.: Wallingford, CT 2009.

(52)   Richard, A. M. and M. W. DSSTox (Distributed Structure-Searchable Toxicity Database Network) Project And Website. Presented at Board of Scientific Counselors 2005, RTP, NC, June 19 - 20, 2006.

(53)   Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, B. E. PubChem 2019 Update: Improved Access to Chemical Data. Nucleic Acids Res. 2019. https://doi.org/10.1093/nar/gky1033.

(54)   Avogadro: An Open-Source Molecular Builder and Visualization Tool.

(55) L. Martínez, R. Andrade, E. G. Birgin, J. M. M. Packmol: A Package for Building Initial Configurations for Molecular Dynamics Simulations. *J. Comput. Chem.* **2009**, *30* (13), 2157–2164.

(56) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26* (16), 1781–1802. https://doi.org/10.1002/jcc.20289.

(57) Humphrey, W., Dalke, A. and Schulten, K. VMD - Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38.

(58) Toton, D.; Lorenz, C. D.; Rompotis, N.; Martsinovich, N.; Kantorovich, L. Temperature Control in Molecular Dynamic Simulations of Non-Equilibrium Processes. *J. Phys. Condens. Matter* **2010**, *22* (7), 074205. https://doi.org/10.1088/0953-8984/22/7/074205.

(59) Di Pierro, M.; Elber, R.; Leimkuhler, B. A Stochastic Algorithm for the Isobaric–Isothermal Ensemble with Ewald Summations for All Long Range Forces. *J. Chem. Theory Comput.* **2015**, *11* (12), 5624–5637. https://doi.org/10.1021/acs.jctc.5b00648.

(60) Chandler, D. *Introduction to Modern Statistical Mechanics*; Oxford University Press: New York, 1987.

(61) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. **2016**.

(62) Wang, J.; Hou, T. Application of Molecular Dynamics Simulations in Molecular Property Prediction . 1 . Density and Heat of Vaporization. *J. Chem. Theory Comput.* **2011**, *7*, 2151–2165. https://doi.org/10.1021/ct200142z.

(63) Yaws, C. L. *Yaws' Handbook of Thermodynamic and Physical Properties of Chemical Compounds*; Knovel, 2003.

(64) Shukla, R. K.; Kumar, A.; Srivastava, U.; Awasthi, N.; Pandey, J. D. Critical Evaluation of Surface Tension of Binary Liquid Mixtures from Associated and Nonassociated Processes at Various Temperatures: An Experimental and Theoretical Study. *Can. J. Phys.* **2013**, *91* (3), 211–220.

(65) Diky, Vladimir; Muzny, Chris D.; Smolyanitsky, Alexander Y.; Bazyleva, Ala; Chirico, Robert D.; Magee, Joe W.; Paulechka, Yauheni; Kazakov, Andrei F.; Townsend, Scott A.; Lemmon, Eric W.; Frenkel, Michael D.; Kroenlein, K. G. NIST Standard Reference Database 103b. Thermodynamic Data for Pure Compounds. 2016.

(66) U.S. Coast Guard Department of Transportation. *CHRIS - Hazardous Chemical Data. Volume II.*; U.S. Government Printing Office: Washington, D.C., 1984.

(67) Horstmann, S.; Fischer, K.; Gmehling, J. Vapor-Liquid-Liquid Equilibria, Azeotropic, and Excess Enthalpy Data for the Binary System n-Undecane + Propionamide and Pure-Component Vapor Pressure and Density Data for Propionamide. *J. Chem. Eng. Data* **2004**, *49* (6), 1494–1498. https://doi.org/10.1021/je034255e.

(68) *CRC Handbook of Chemistry and Physics*, 94th ed.; Haynes, William M.; Lide, David R.; Bruno, T. J., Ed.; CRC Press: Boca Raton, 2013.

(69) Wolfarth, Ch; Wolfarth, B. *Surface Tension of Pure Liquids and Binary Liquid Mixtures. Landolt-Börnstein - Group IV Physical Chemistry (Numerical Data and Functional Relationships in Science and Technology), Vol 28*; Lechner, M. D., Ed.; Springer: Berlin, 1997.

(70) Lee, O. S.; Liu, Y.; Schatz, G. C. Molecular Dynamics Simulation of β-Sheet Formation in Self-Assembled Peptide Amphiphile Fibers. *J. Nanoparticle Res.* **2012**, *14* (8). https://doi.org/10.1007/s11051-012-0936-z.

(71) Comer, J.; Gumbart, J. C.; Hénin, J.; Lelièvre, T.; Pohorille, A.; Chipot, C. The Adaptive Biasing Force Method: Everything You Always Wanted To Know but Were Afraid To Ask. *J. Phys. Chem. B* **2015**, *119* (3), 1129–1151. https://doi.org/10.1021/jp506633n.

(72) Darve, E.; Rodríguez-Gómez, D.; Pohorille, A. Adaptive Biasing Force Method for Scalar and Vector Free Energy Calculations. *J. Chem. Phys.* **2008**, *128* (14), 144120. https://doi.org/10.1063/1.2829861.

(73) Fiorin, G.; Klein, M. L.; Hénin, J. Using Collective Variables to Drive Molecular Dynamics Simulations. *Mol. Phys.* **2013**, *111* (22–23), 3345–3362. https://doi.org/10.1080/00268976.2013.813594.

(74) Wolfenden, R.; Andersson, L.; Cullis, P. M.; Southgate, C. C. B. Affinities of Amino Acid Side Chains for Solvent Water. *Biochemistry* **1981**, *20* (4), 849–855. https://doi.org/10.1021/bi00507a030.

(75) In, Y.; Chai, H. H.; No, K. T. A Partition Coefficient Calculation Method with the SFED Model. *J. Chem. Inf. Model.* **2005**, *45* (2), 254–263. https://doi.org/10.1021/ci0498564.

(76) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. The SGB/NP Hydration Free Energy Model Based on the Surface Generalized Born Solvent Reaction Field and Novel Nonpolar Hydration Free Energy Estimators. *J. Comput. Chem.* **2002**, *23* (5), 517–529. https://doi.org/10.1002/jcc.10045.

(77) DeVane, R.; Shinoda, W.; Moore, P. B.; Klein, M. L. Transferable Coarse Grain Nonbonded Interaction Model for Amino Acids. *J. Chem. Theory Comput.* **2009**, *5* (8), 2115–2124. https://doi.org/10.1021/ct800441u.

(78) Bulacu, M.; Goga, N.; Zhao, W.; Rossi, G.; Monticelli, L.; Periole, X.; Tieleman, D. P.; Marrink, S. J. Improved Angle Potentials for Coarse-Grained Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2013**, *9* (8), 3282–3292. https://doi.org/10.1021/ct400219n.

Appendix A: Mapping Schemes

Table A.1: Amino Acid Mapping Schemes
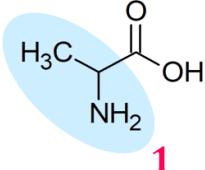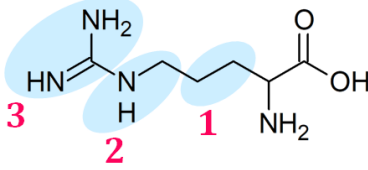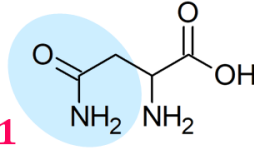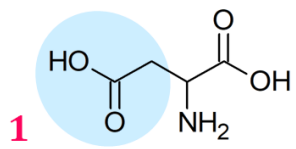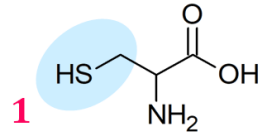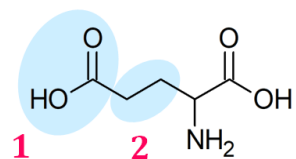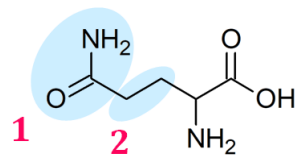
Name	Structure	Bead Name	Bead Type
Alanine		1) ANT (AB within backbone)	1) NCC1 (NCC2 within backbone)
Arginine		1) RS1 2) RS2 3) RS3	1) RS1 2) RS2 3) RS3
Asparagine		1) NS	1) CON2
Aspartic Acid		1) DS	1) COO2
Cysteine		1) CS	1) SC
Glutamic Acid		1) ES2 2) ES1	1) COO1 2) C2M
Glutamine		1) QS2 2) QS1	1) CON1 2) C2M

Table A.1, continued: Amino Acid Mapping Schemes
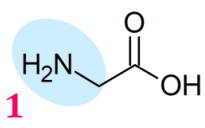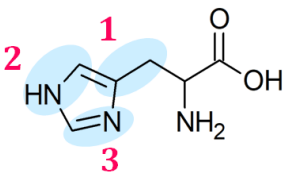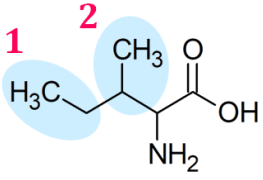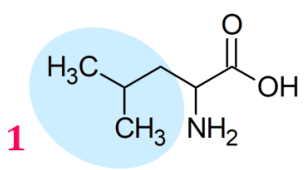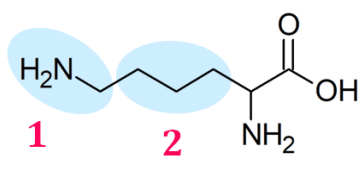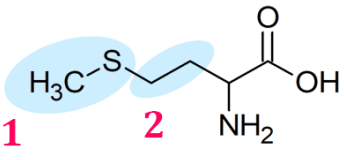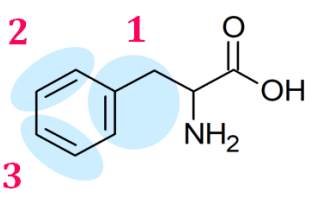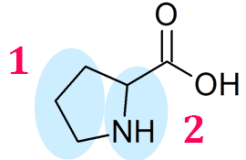
Glycine		1) GNT (GNC within backbone)	1) GNT (GNC within backbone)
Histidine		1) HS1 2) HS2 3) HS3	1) CCR1 2) NCR1 3) NCR2
Isoleucine		1) IS2 2) IS1	1) C2E 2) C2E2
Leucine		1) LS	1) C41
Lysine		1) KS2 2) KS1	1) NC2 2) C3M
Methionine		1) MS2 2) MS1	1) CS 2) C2M
Phenylalanine		1) FS1 2) FS21 3) FS22	1) TL2F 2) BZF1 3) BZF2
Proline		1) PS 2) PN	1) PS 2) PNC

Table A.1, continued: Amino Acid Mapping Schemes
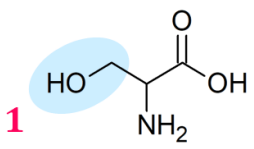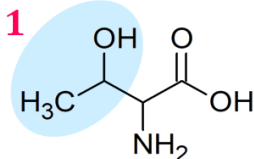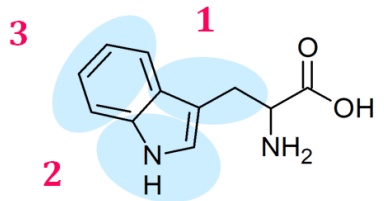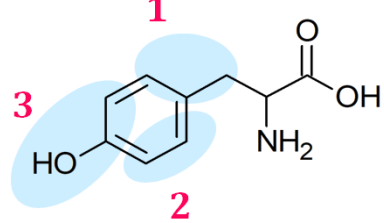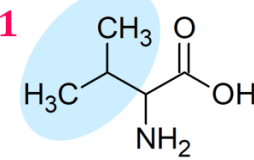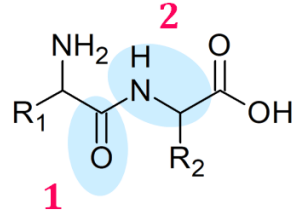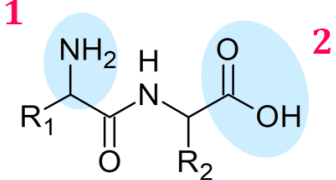
Serine		1) SS	1) COH1
Threonine		1) TS	1) CCOH
Tryptophan		1) WS1 2) WS2 3) WS3	1) TL2W 2) PL1 3) TOL4
Tyrosine		1) YS1 2) YS2 3) YS3	1) TL2Y 2) BZY 3) COH3
Valine		1) VS	1) C3E2
Backbone		1) BCO 2) BNC	1) CO 2) NC
Termini		1) NT 2) CT	1) NCT 2) COO1

Table A.2: Analogue Mapping Schemes
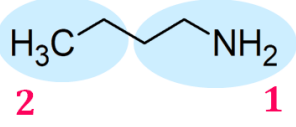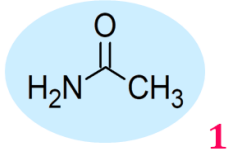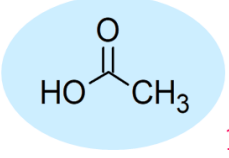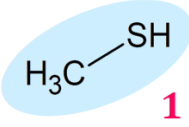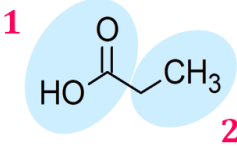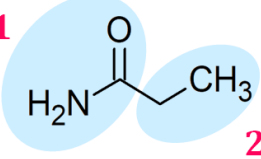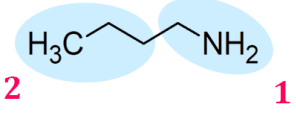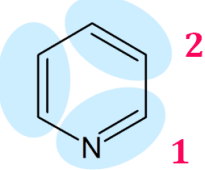
Amino Acid	Name	Structure	Bead Type
alanine	n-butylamine	H_3C — ... — NH_2 (2, 1)	1) NCC1 2) C2E
asparagine	acetamide	H_2N—C(=O)—CH_3 (1)	1) CON2
aspartic acid	acetic acid	HO—C(=O)—CH_3 (1)	1) COO2
cysteine	methanethiol	H_3C—SH (1)	1) SC
glutamic acid	propionic acid	HO—C(=O)—CH_3 (1, 2)	1) COO1 2) C2E
glutamine	propionamide	H_2N—C(=O)—CH_3 (1, 2)	1) CON1 2) C2M
glycine	butylamine	H_3C — ... — NH_2 (2, 1)	1) NC2 2) C3M
histidine	pyridine	pyridine ring (2, 2, 1)	1) NCR2 2) BZ

histidine	pyrrole		1) NCR1 2) PL3
histidine	toluene		1) CCR1 2) BZ 3) TOL3
isoleucine	3,4-dimethylhexane		1) C2E 2) C2E2
leucine	2,5-dimethylhexane		1) C41
lysine	butylamine		1) NC2 2) C3E
methionine	ethyl methyl sulfide		1) CS 2) C2M
phenylalanine	toluene		1) TL2F 2) BZ (BZF1 and BZF2)
serine	methanol		1) COH1

threonine	ethanol	H₃C—OH **1**	1) CCOH
tryptophan	pyrrole	pyrrole structure, **1**, **2**	1) PL1 2) PL2
tryptophan	toluene	toluene structure, **1**, **2**	1) TOL4 2) TOL2
tyrosine	m-cresol	m-cresol structure, **1**, **2**, **3**	1) COH3 2) BZY 3) TL2Y
valine	2,3-dimethylbutane	2,3-dimethylbutane structure, **1**, **1**	1) C3E2
backbone	methylacetamide	methylacetamide structure, **1**, **2**	1) NC 2) CCO
backbone	diethylketone	diethylketone structure, **1**, **2**, **1**	1) CC 2) CO

Table A.2, continued: Analogue Mapping Schemes

N-terminus	butylamine		1) NCT 2) C3E
C-terminus	propionic acid		1) COO1 2) C2E

Appendix B: CG Analogue Parameter Development Results

Table B.1: Comparison of experimental and CG model densities

Amino Acid	Analogue	Experimental Density (g/cm³)[a]	Model Average (g/cm³)	% Error	Standard Deviation
Ala	N-butylamine	0.741	0.728	1.78	7.93E-05
Asn	acetamide	0.999	1.002	0.27	9.01E-05
Asp	acetic acid	1.043	1.042	0.11	1.20E-04
Cys	methanethiol	0.892	0.887	0.56	5.70E-05
Gln	propionamide	0.96	0.964	0.43	3.36E-05
His	pyridine	0.979	0.967	1.20	5.64E-05
His	pyrrole	0.965	0.944	2.22	7.97E-05
His	toluene	0.865	0.856	1.07	5.94E-05
Ile	3,4-dimethylhexane	0.716	0.703	1.81	4.02E-05
Leu	2,5-dimethylhexane	0.694	0.687	0.96	4.49E-05
Lys	N-butylamine	0.741	0.731	1.40	3.10E-05
Met	ethyl methyl sulfide	0.832	0.834	0.28	4.74E-05
Phe	toluene	0.865	0.862	0.40	3.36E-05
Ser	methanol	0.787	0.786	0.07	1.27E-04
Thr	ethanol	0.787	0.789	0.25	1.30E-04
Trp	pyrrole	0.965	0.957	0.83	3.86E-05
Trp	toluene	0.865	0.864	0.14	8.75E-05
Tyr	m-cresol	1.03	1.036	0.53	5.59E-05
Val	2,3-dimethylbutane	0.658	0.652	0.85	4.17E-05
Backbone	N-methylacetamide	0.9405	0.94152	0.11	5.73E-05
Backbone	diethylketone	0.8089	0.7843	3.04	6.50E-05

[a] Experimental data from refs [63–67]

Table B.2: Comparison of experimental and CG model heats of vaporization

Amino Acid	Analogue	Experimental H^{vap} (kcal/mol)[a]	Model Average (kcal/mol)	% Error	Standard Deviation
Ala	N-butylamine	8.537	7.398	13.34	2.98E-03
Asn	acetamide	*NA*	--	--	--
Asp	acetic acid	5.583	5.481	1.83	7.60E-04
Cys	methanethiol	5.872	5.544	5.59	4.75E-04
Gln	propionamide	*NA*	--	--	--
His	pyridine	9.61	8.72	9.28	4.58E-03
His	pyrrole	10.78	8.89	17.55	2.95E-03

Table B.2, continued: Comparison of experimental and CG model heats of vaporization

His	toluene	9.0846	8.5631	5.74	5.84E-03
Ile	3,4-dimethylhexane	9.314	9.570	2.75	6.41E-03
Leu	2,5-dimethylhexane	9.046	8.398	7.16	2.41E-03
Lys	N-butylamine	8.537	7.092	16.93	3.33E-03
Met	ethyl methyl sulfide	7.6123	7.0342	7.59	1.46E-03
Phe	toluene	9.0846	8.4578	6.90	4.59E-03
Ser	methanol	8.946	4.312	51.80	4.68E-04
Thr	ethanol	10.115	4.900	51.56	9.41E-04
Trp	pyrrole	10.78	9.31	13.60	1.95E-03
Trp	toluene	9.0846	7.9952	11.99	1.77E-03
Tyr	m-cresol	14.749	10.417	29.37	3.13E-03
Val	2,3-dimethylbutane	6.96	6.47	7.05	2.04E-03
Backbone	N-methylacetamide	*NA*	--	--	--
Backbone	diethylketone	9.2065	8.4050	8.71	4.73E-03
a Experimental data from refs [63,65,68]					

Table B.3: Comparison of experimental and CG model surface tensions

Amino Acid	Analogue	Experimental Surface Tension (mN/m)[a]	Model Average (mN/m)	% Error	Standard Deviation
Ala	N-butylamine	23.44	23.787	1.48	0.159
Asn	acetamide	38.96	38.963	0.01	0.157
Asp	acetic acid	27.1	28.172	3.96	0.223
Cys	methanethiol	31	31.510	1.65	0.133
Gln	propionamide	31.31	31.259	0.16	0.150
His	pyridine	36.72	37.502	2.13	0.223
His	pyrrole	36.32	38.040	4.74	0.143
His	toluene	27.93	28.043	0.40	0.064
Ile	3,4-dimethylhexane	21.21	22.578	6.45	0.104
Leu	2,5-dimethylhexane	19.28	20.420	5.91	0.220
Lys	N-butylamine	23.44	23.720	1.20	0.092
Met	ethyl methyl sulfide	24.42	25.651	5.04	0.114
Phe	toluene	27.93	29.197	4.54	0.177
Ser	methanol	22.07	22.655	2.65	0.121
Thr	ethanol	21.97	22.833	3.93	0.037
Trp	pyrrole	36.32	35.902	1.15	0.147
Trp	toluene	27.93	28.335	1.45	0.062
Tyr	m-cresol	35.76	37.381	4.53	0.226
Val	2,3-dimethylbutane	16.87	17.073	1.20	0.113

Table B.3, continued: Comparison of experimental and CG model surface tensions

Backbone	N-methylacetamide	32.53	32.933	1.24	0.160
Backbone	diethylketone	24.74	26.565	7.38	0.160
a Experimental data from refs [63–66,68,69]					

Appendix C: Peptides and Analogues, Distributions and RDFs

C1: Alanine

Ala-Ala-Ala Tripeptide
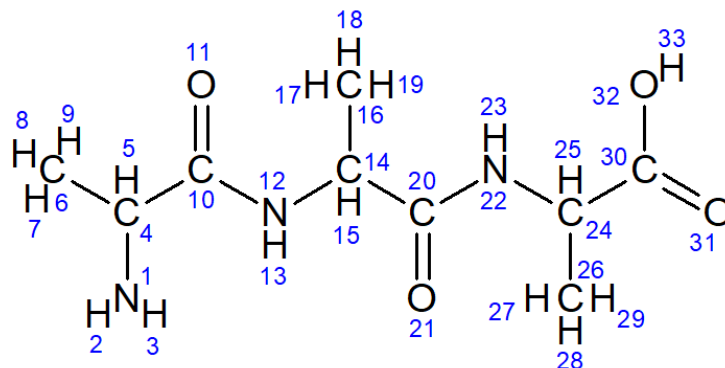
Figure C1.1: All-atom Alanine tripeptide with atom numbering
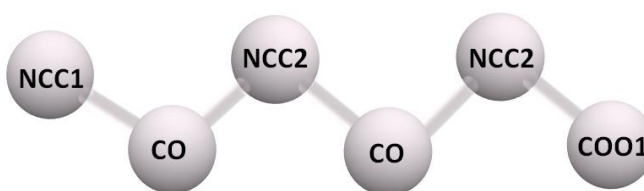
Figure C1.2: CG model of Alanine tripeptide showing bead types
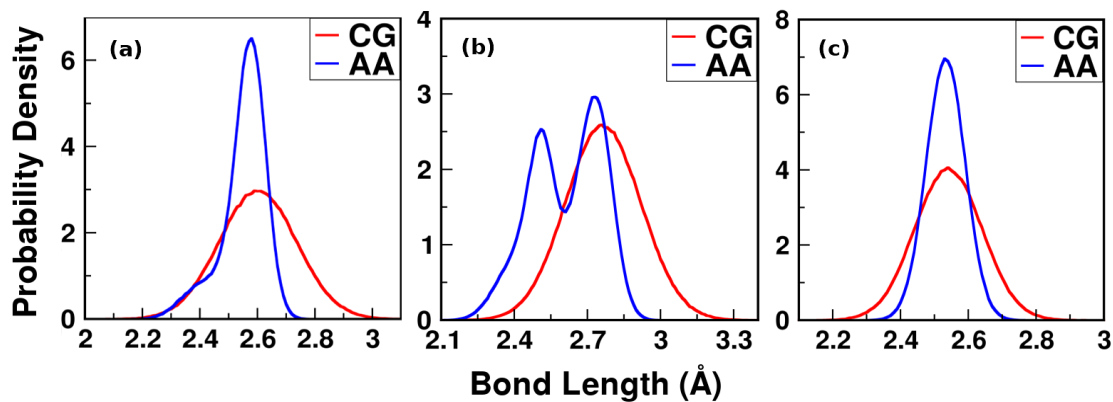
Figure C1.3: Distributions of the bonds (a) NCC1-CO, (b) CO-NCC2, and (c) NCC2-COO1
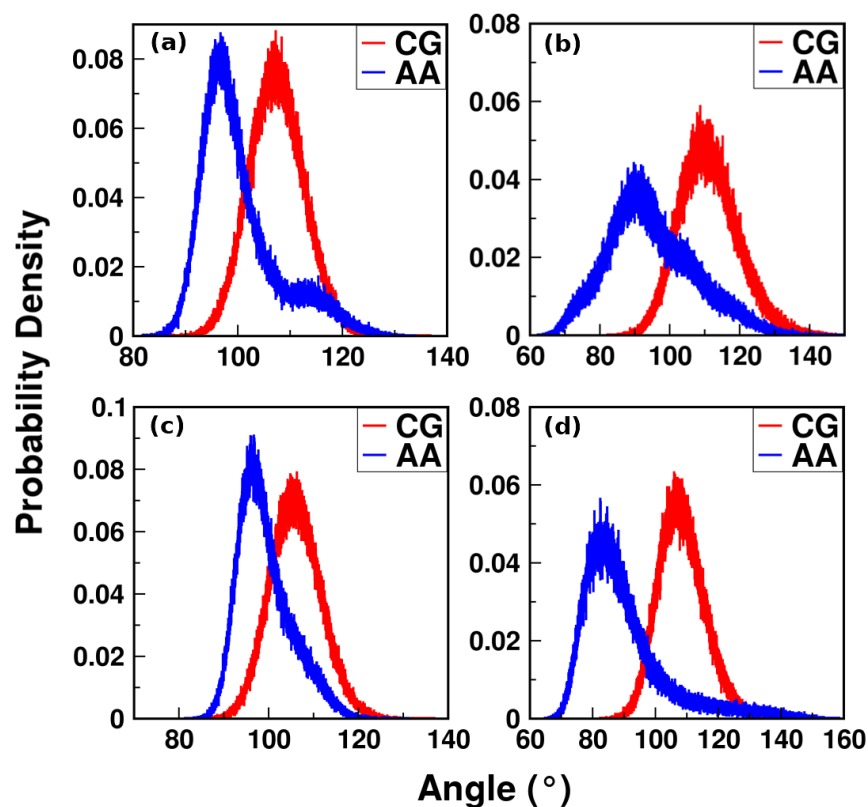
**Figure C1.4:** Distributions of angles (a) NCC1-CO-NCC2, (b) CO-NCC2-CO, (c) NCC2-CO-NCC2, and (d) CO-NCC2-COO1
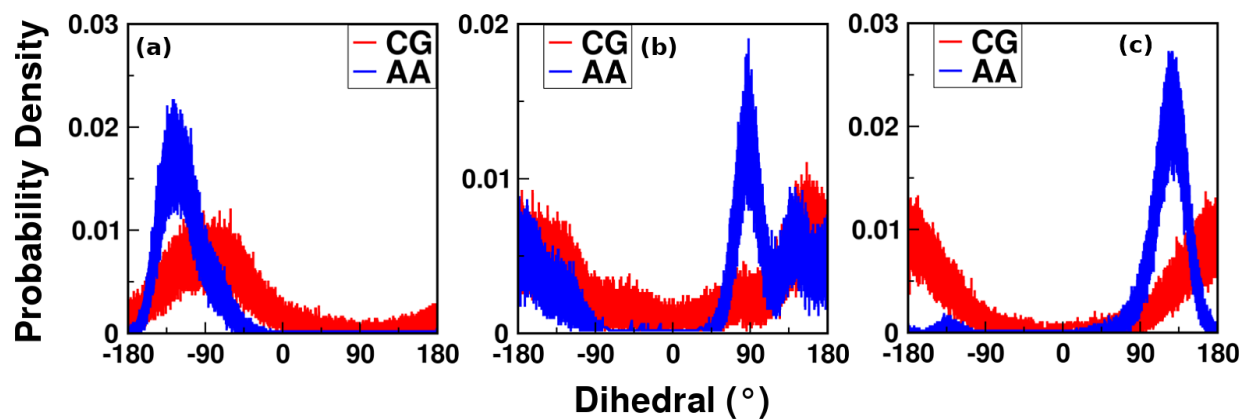
Figure C1.5: Distributions of dihedrals (a) NCC1-CO-NCC2-CO (b) CO-NCC2-CO-NCC2, and (c) NCC2-CO-NCC2-COO1

85
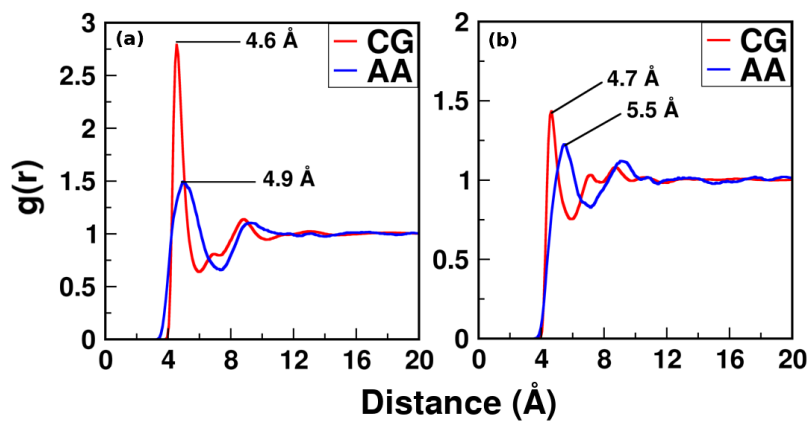
Figure C1.6: RDFs of (a) NCC1 and (b) NCC2

Analogue: N-butylamine
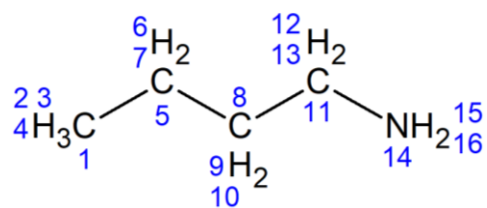
Figure C1.7: All-atom N-butylamine with atom numbering

Figure C1.8: CG model of N-butylamine showing bead types
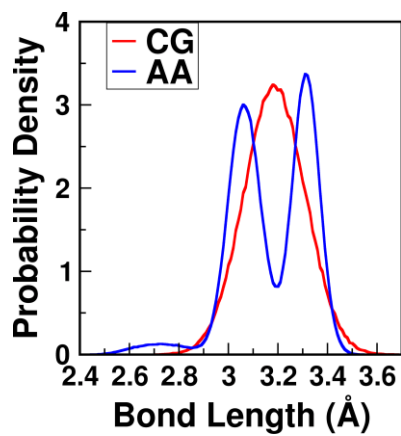
Figure C1.9: Distribution of bond NCC1-C2E
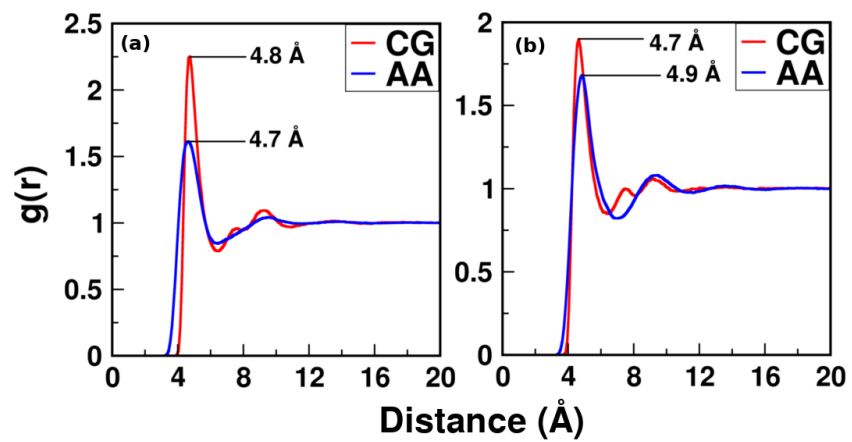
Figure C1.10: RDFs of (a) NCC1 and (b) C2E

C2: Arginine

Table C2.1: Arginine side chain atom charges. Hydrogens HB3 and HH22 are not in the neutral side chain; *NA* signifies "Not Available" in consequence. Atom names are CHARMM atom names.

Atom Name	Charges		
	CHARMM	**Gaussian**	**Assigned**
CB	-0.18	-0.518	-0.380
HB1	0.09	0.137	0.137
HB2	0.09	0.139	0.137
HB3	*NA*	*0.136*	*NA*
CG	-0.18	-0.091	-0.091
HG1	0.09	0.143	0.130
HG2	0.09	0.116	0.130
CD	0.20	-0.317	-0.317
HD1	0.09	0.162	0.151
HD2	0.09	0.140	0.151
NE	-0.70	-0.177	-0.177
HE	0.44	0.221	0.221
CZ	0.64	0.099	0.099
NH1	-0.80	-0.445	-0.445
HH11	0.46	0.186	0.186
HH22	*0.46*	*NA*	*NA*
NH2	-0.80	-0.380	-0.380
HH21	0.46	0.228	0.224
HH22	0.46	0.220	0.224
Total	**1**	**-0.001**	**0**

Ala-Arg Dipeptide
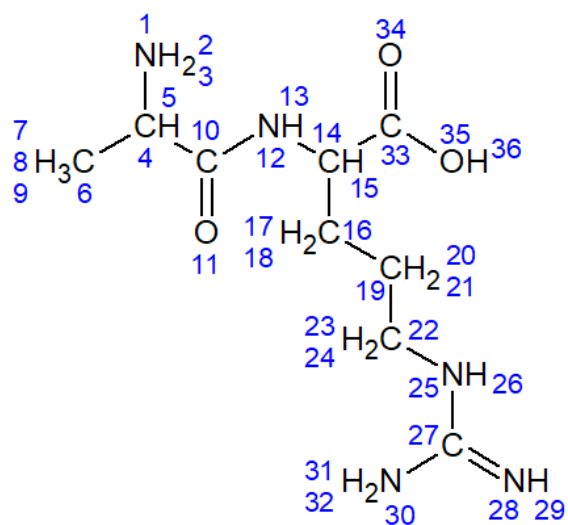
Figure C2.1: All-atom Alanyl-Arginine dipeptide with atom numbering
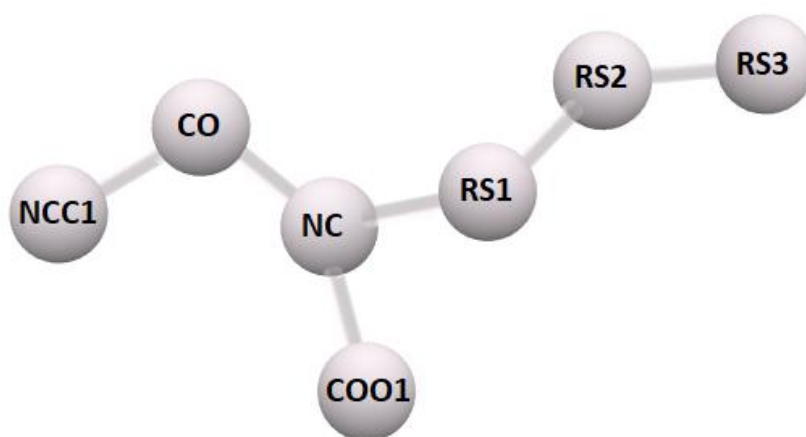
Figure C2.2: CG model of Alanyl-Arginine dipeptide showing bead types
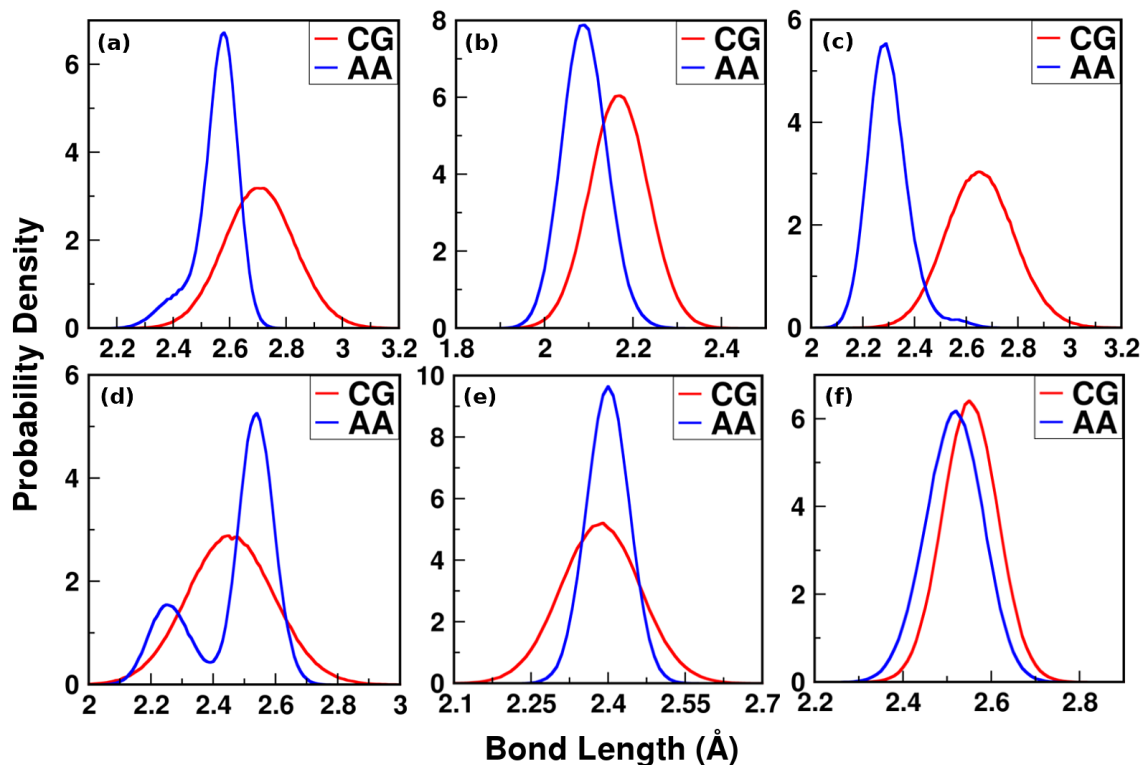
Figure C2.3: Distribution of bonds (a) NCC1-CO, (b) CO-NC, (c) NC-RS1, (d) RS1-RS2, (e) RS2-RS3, and (f) NC-COO1
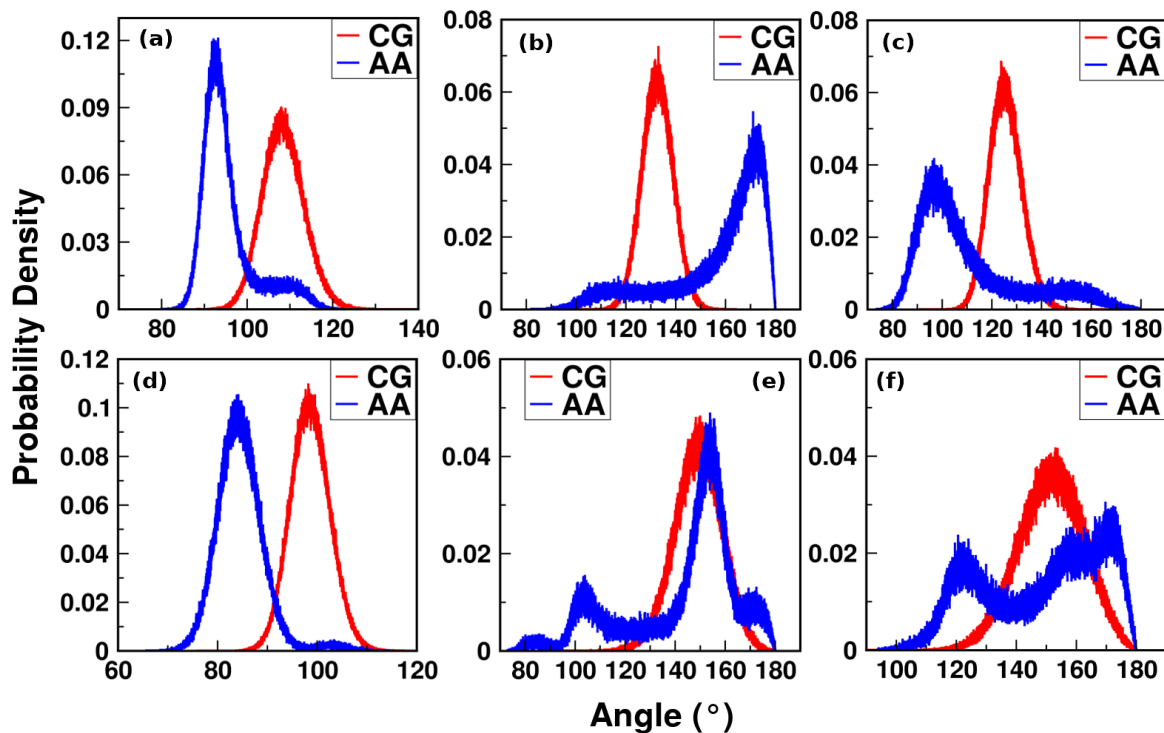
Figure C2.4: Distributions of angles (a) NCC1-CO-NC, (b) CO-NC-RS1, (c) CO-NC-COO1, (d) RS1-NC-COO1, (e) NC-RS1-RS2, and (f) RS1-RS2-RS3
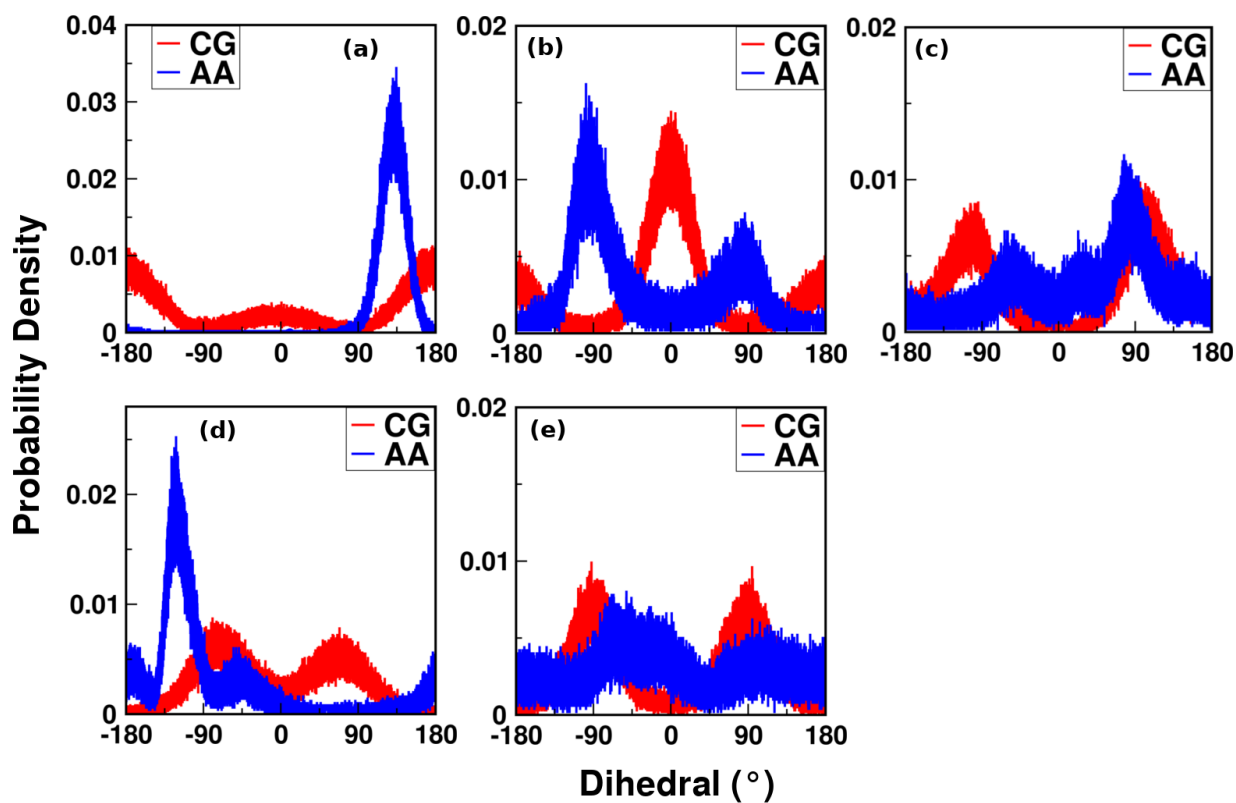
Figure C2.5: Distributions of dihedrals (a) NCC1-CO-NC-COO1, (b) NCC1-CO-NC-RS1, (c) CO-NC-RS1-RS2 (d) COO1-NC-RS1-RS2, and (e) NC-RS1-RS2-RS3
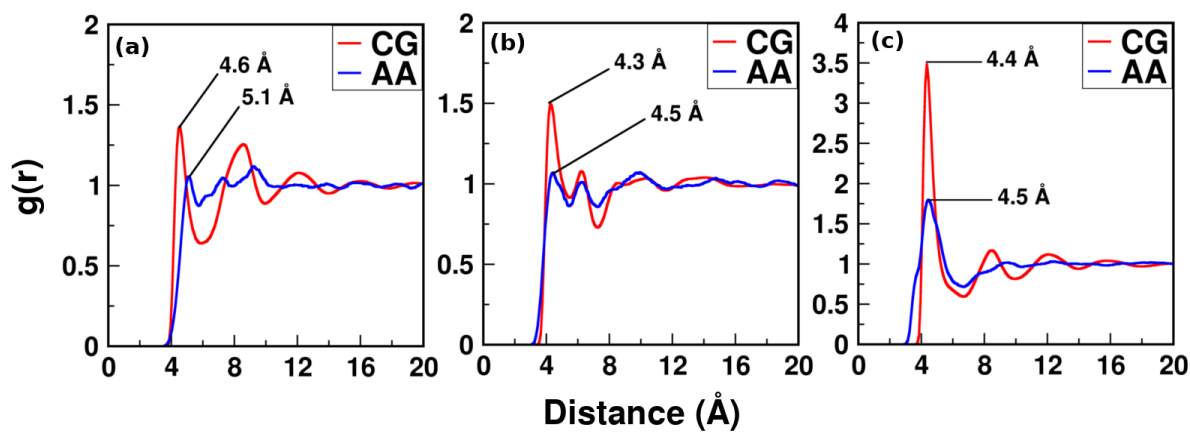
Figure C2.6: RDFs of (a) RS1, (b) RS2, and (c) RS3

C3: Asparagine

Asn-Asn Dipeptide
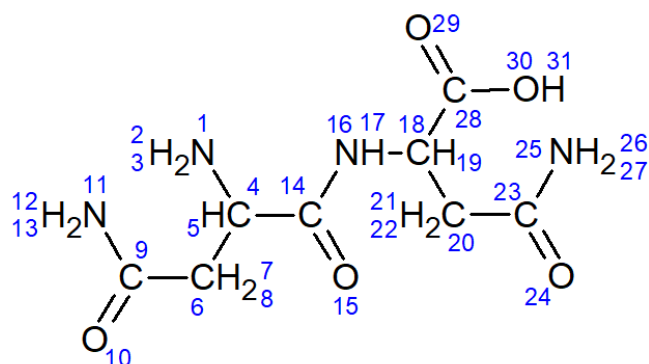
Figure C3.1: All-atom Asparagine dipeptide with atom numbering
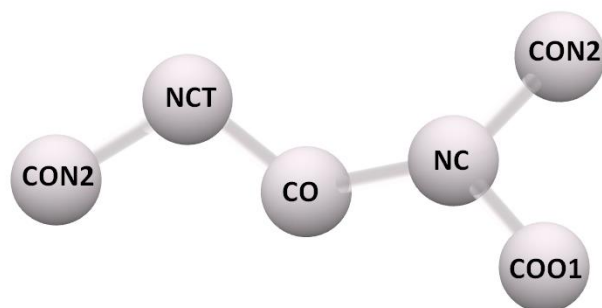
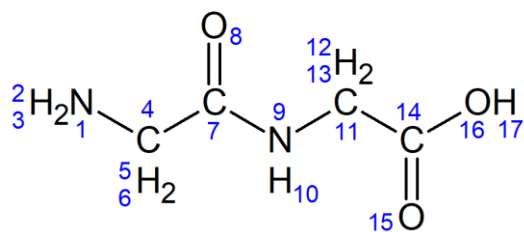**Figure C3.2:** CG model of Asparagine dipeptide showing bead types
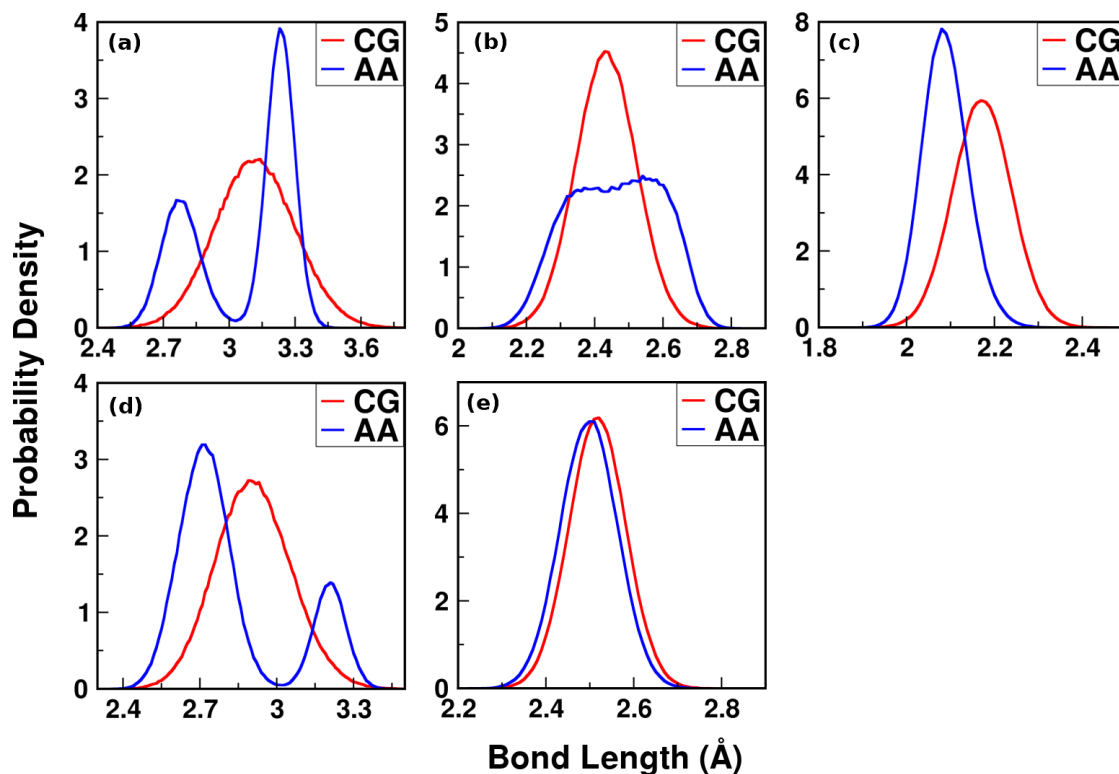
Figure C3.3: Distributions of bonds (a) NS-NCT, (b) NCT-CO, (c) CO-NC, (d) NC-CON2, and (e) NC-COO1
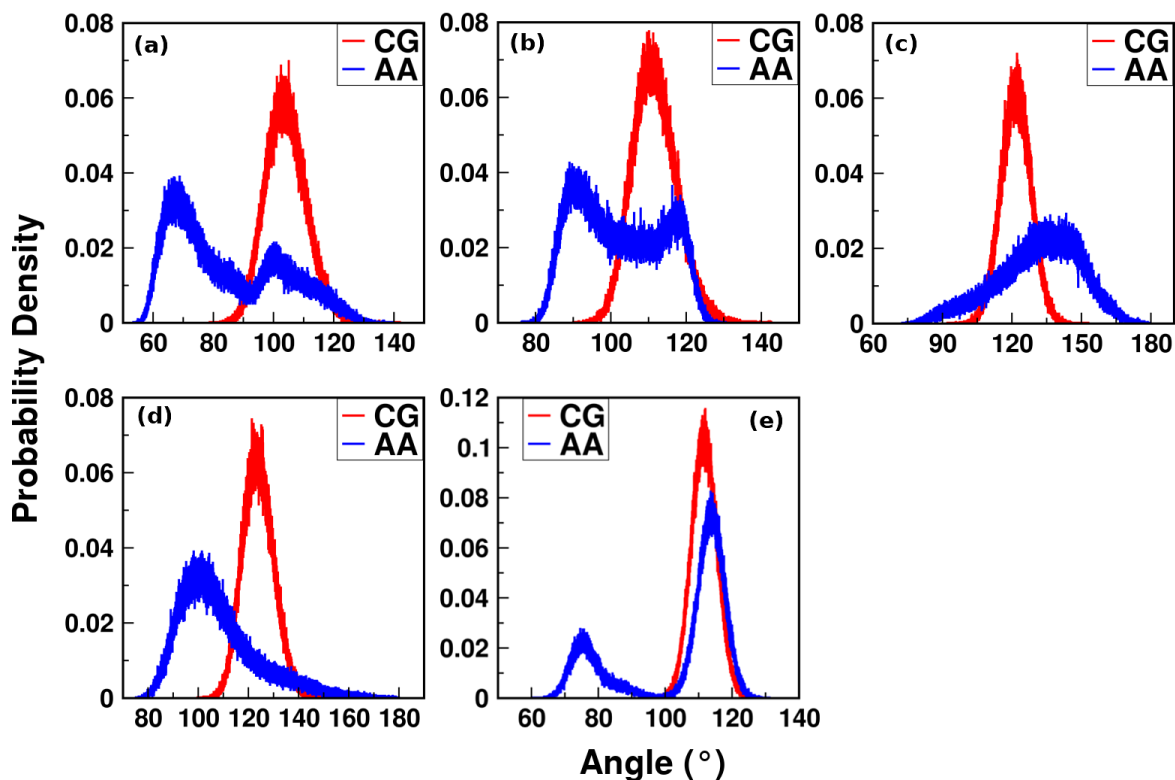
Figure C3.4: Distributions of angles (a) CON2-NT-CO, (b) NT-CO-NC, (c) CO-NC-CON2, (d) CO-NC-COO1, and (e) CON2-NC-COO1
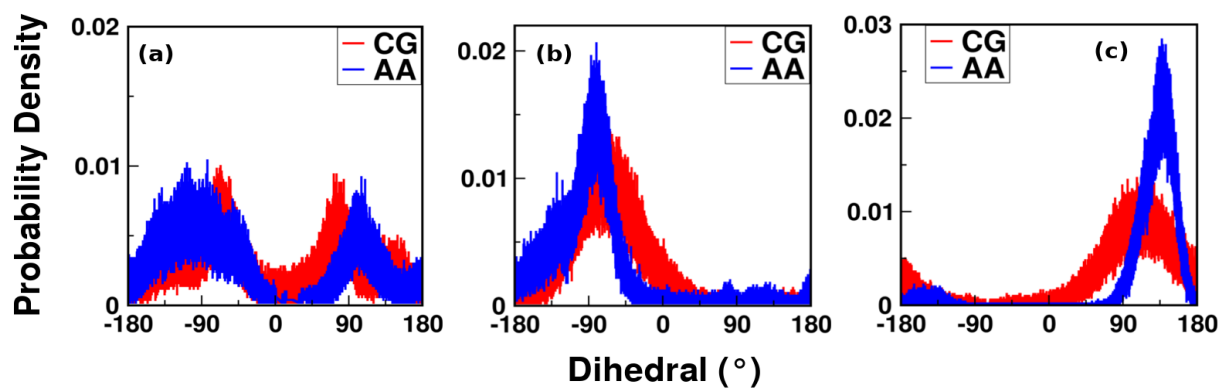
Figure C3.5: Distributions of dihedrals (a) CON2-NCT-CO-NC, (b) NCT-CO-NC-CON2, and (c) NCT-CO-NC-COO1
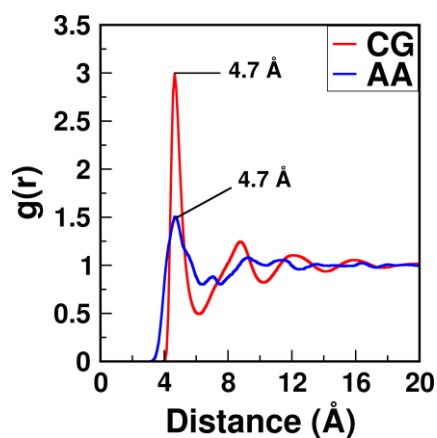
Figure C3.6: RDF of CON2

Analogue: Acetamide
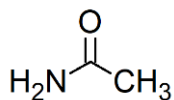
Figure C3.7: All-atom acetamide

Figure C3.8: CG model of acetamide showing bead type
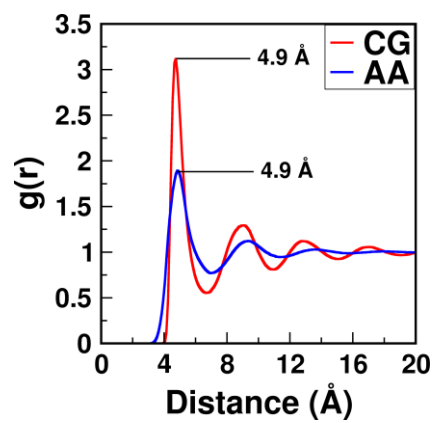
Figure C3.9: RDF of CON2

C4: Aspartic Acid

Asp-Asp Dipeptide
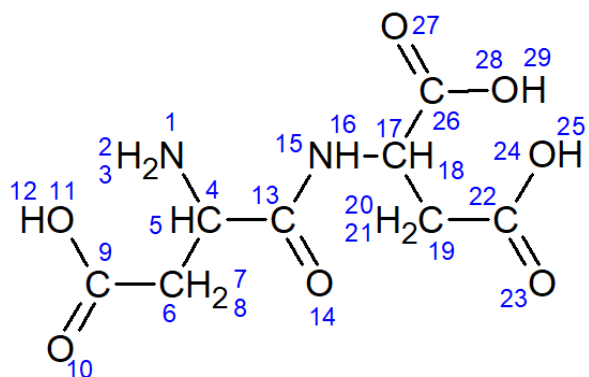
Figure C4.1: All-atom Aspartic Acid dipeptide with atom numbering
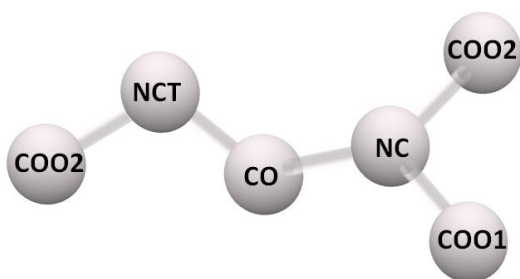
Figure C4.2: CG model of Aspartic Acid dipeptide showing bead types
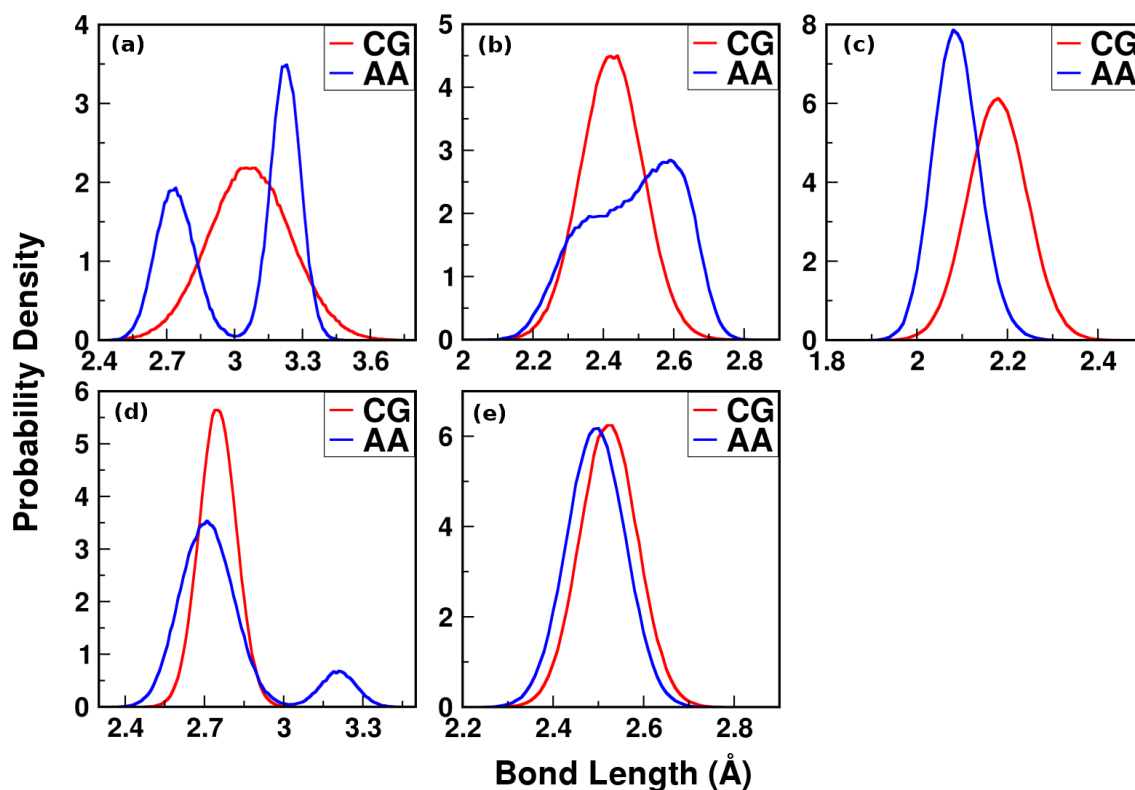
Figure C4.3: Distributions of bonds (a) COO2-NCT, (b) NCT-CO, (c) CO-NC, (d) NC-COO2, and (e) NC-COO1
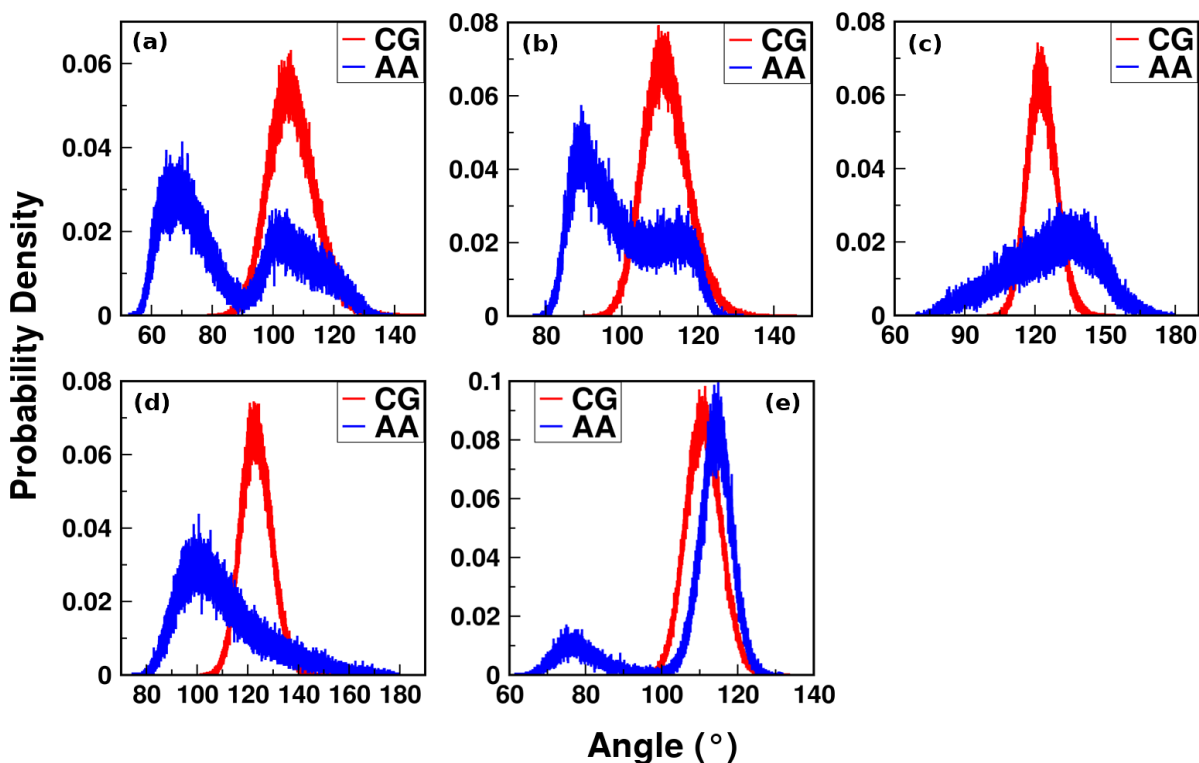
Figure 6.4: Distributions of angles (a) COO2-NCT-CO, (b) NCT-CO-NC, (c) CO-NC-COO2, (d) CO-NC-COO1, and (e) COO2-NC-COO1

97
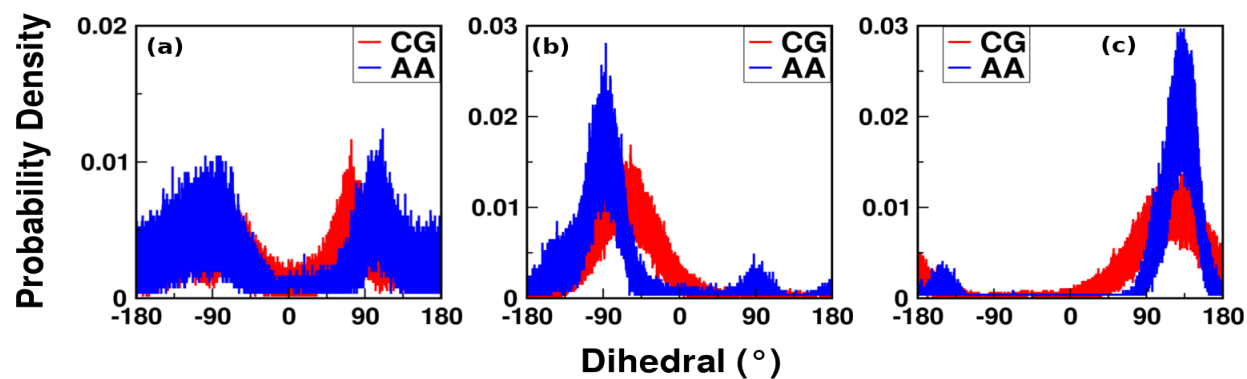
Figure C4.5: Distributions of dihedrals (a) COO2-NCT-CO-NC, (b) NCT-CO-NC-COO2, and (c) NCT-CO-NC-COO1
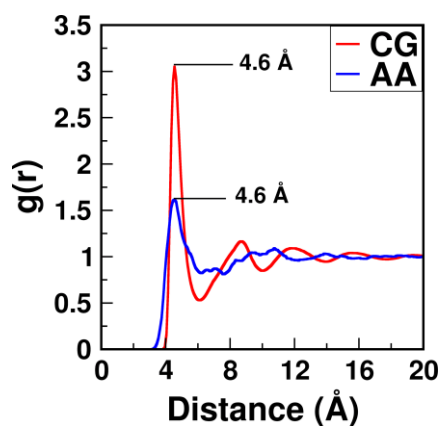
Figure C4.6: RDF of COO2

Analogue: Acetic Acid
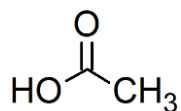
Figure C4.7: All-atom acetic acid

Figure C4.8: CG model of acetic acid showing bead type
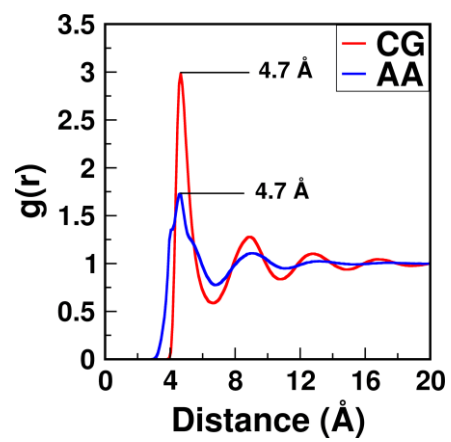
Figure C4.9: RDF of COO2

C5: Backbone & Termini

Backbone Carbonyl Analogue: Diethylketone
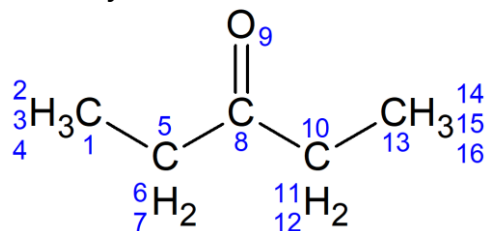
Figure C5.1: All-atom diethylketone with atom numbering
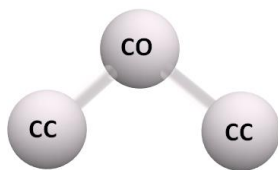
Figure C5.2: CG model of diethylketone showing bead types
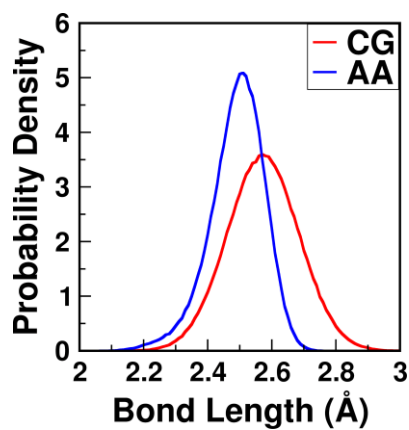
Figure C5.3: Distribution of bond CC-CO
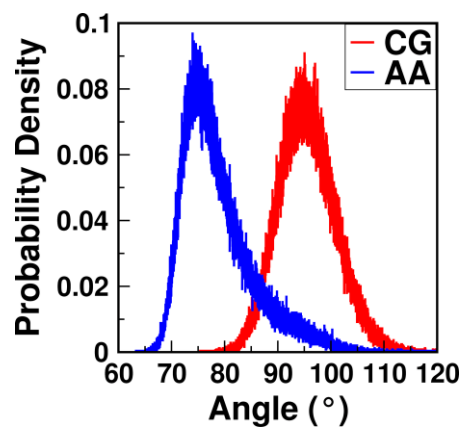
Figure C5.4: Distribution of angle CC-CO-CC
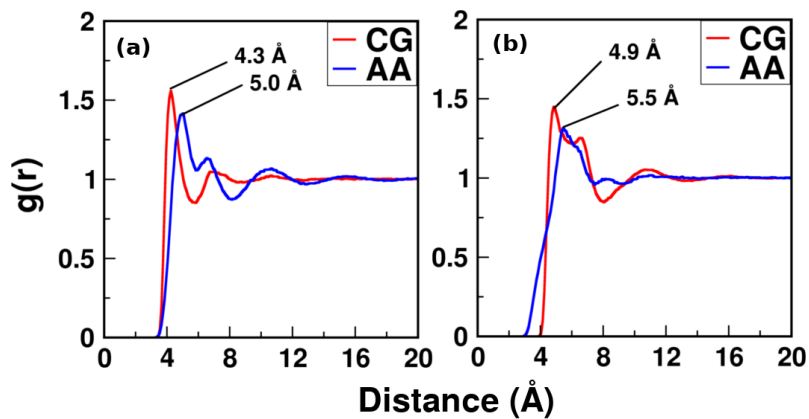
Figure C5.5: RDFs of (a) CC and (b) CO

Backbone N-Cα Analogue: Methylacetamide
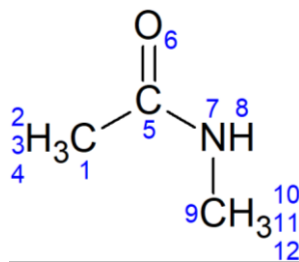
Figure C5.6: All-atom N-methylacetamide with atom numbering
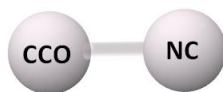
Figure C5.7: CG model of N-methylacetamide showing bead types
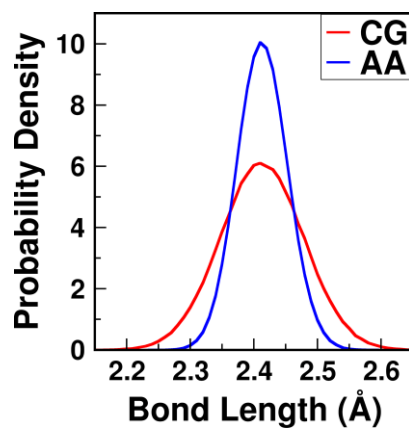
Figure C5.8: Distribution of bond NC-CCO
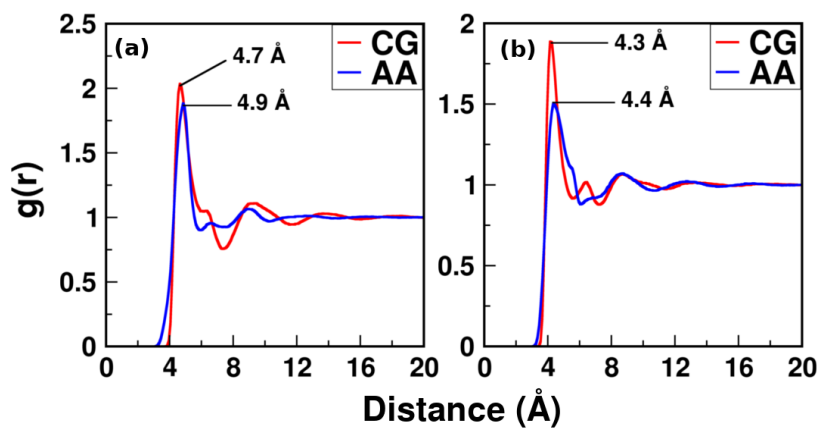
Figure C5.9: RDFs of (a) CCO and (b) NC

C6: Cysteine

Cys-Cys Dipeptide
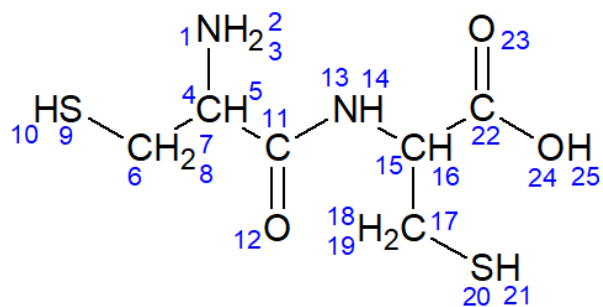
Figure C6.1: All-atom Cysteine dipeptide with atom numbering
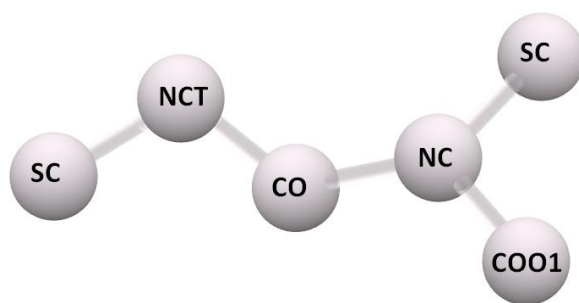
**Figure C6.2:** CG model of Cysteine dipeptide showing bead types
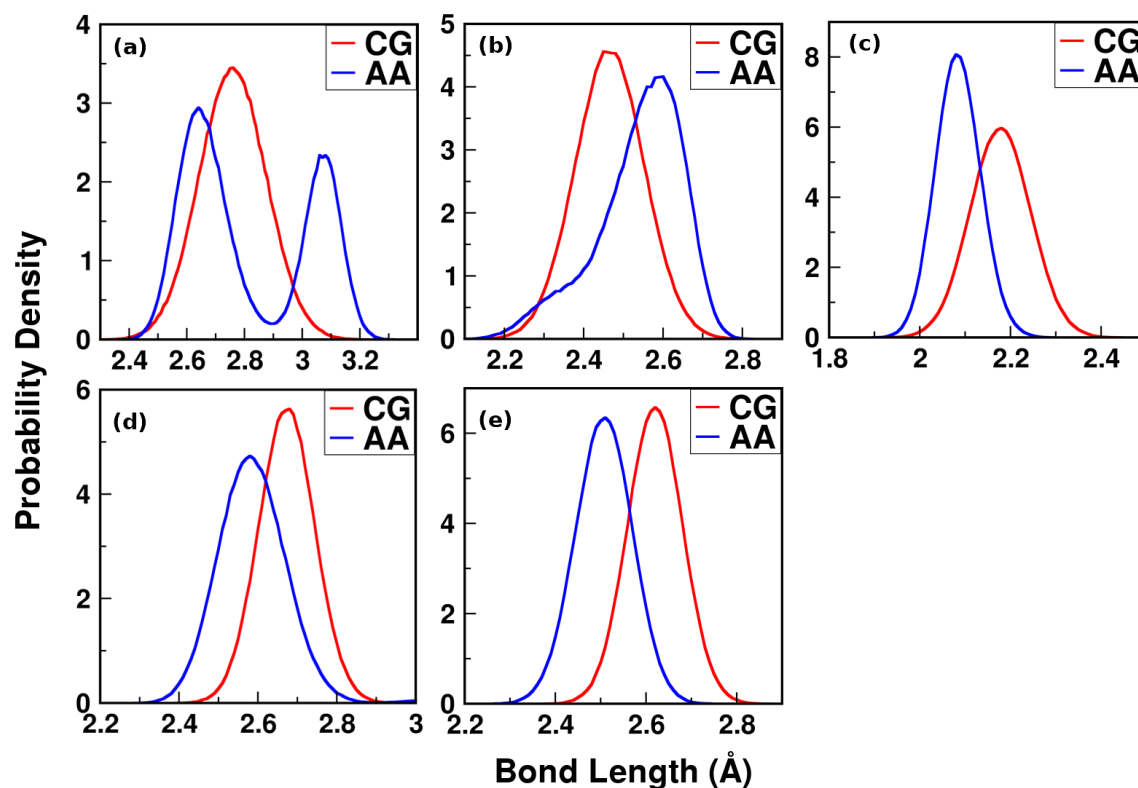
Figure C6.3: Distributions of bonds (a) SC-NCT, (b) NCT-CO, (c) CO-NC, (d) NC-SC, and (e) NC-COO1
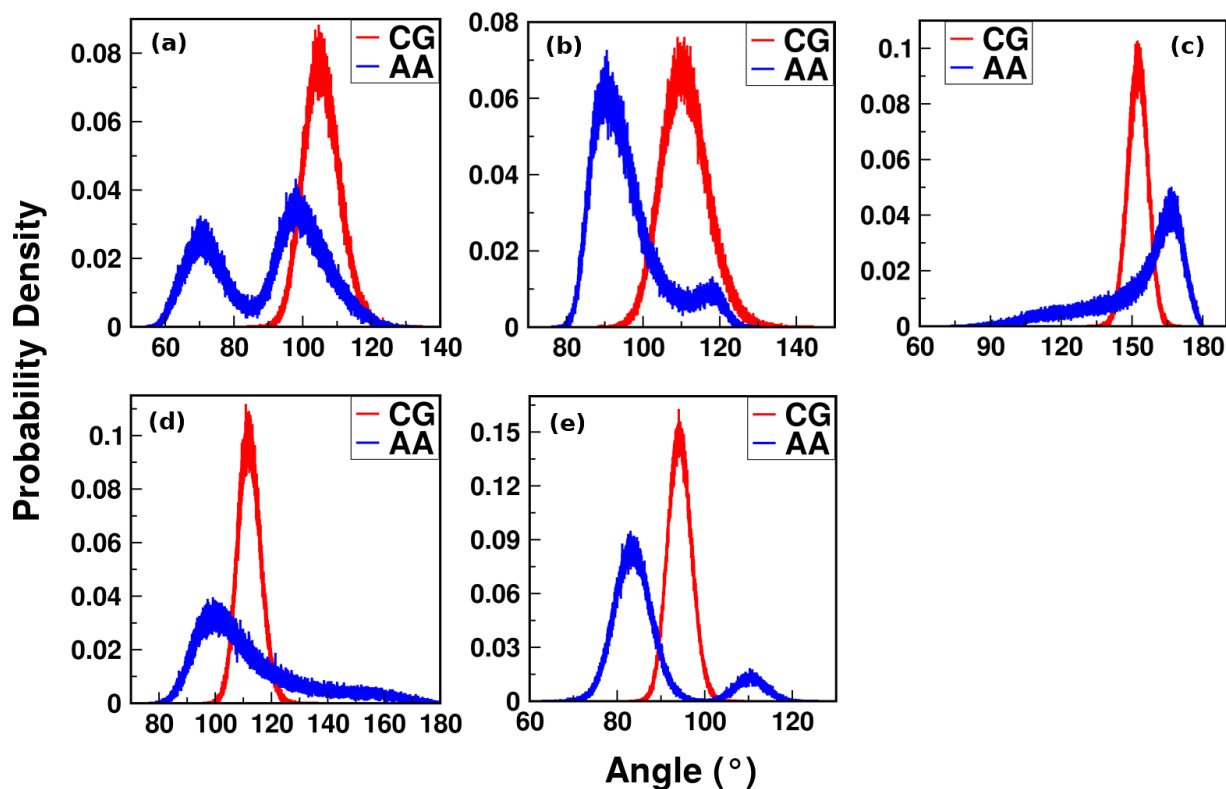
Figure C6.4: Distributions of angles (a) SC-NCT-CO, (b) NCT-CO-NC, (c) CO-NC-SC, (d) CO-NC-COO1, and (e) SC-NC-COO1
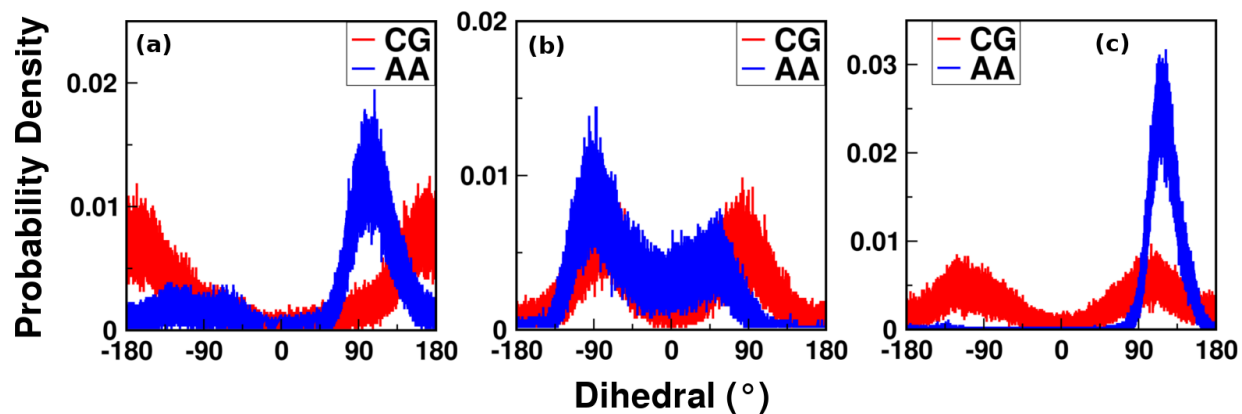
Figure C6.5: Distributions of dihedrals (a) SC-NCT-CO-NC, (b) NCT-CO-NC-SC, and (c) NCT-CO-NC-COO1
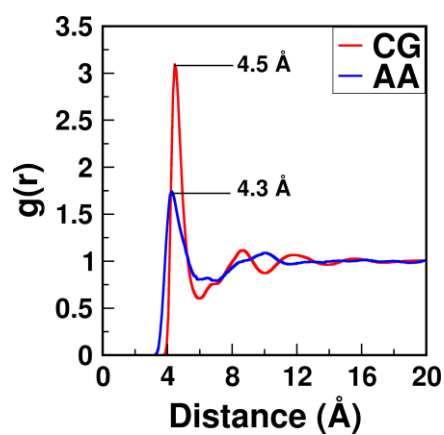
Figure C6.6: RDF of SC

Analogue: Methanethiol
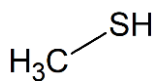
Figure C6.7: All-atom methanethiol

Figure C6.8: CG model of methanethiol showing bead type

105
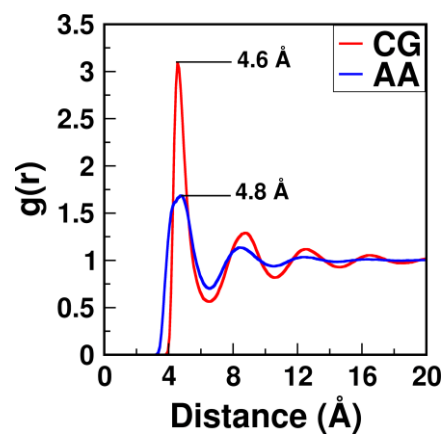
Figure C6.9: RDF of SC

C7: Glutamic Acid

Glu-Glu Dipeptide
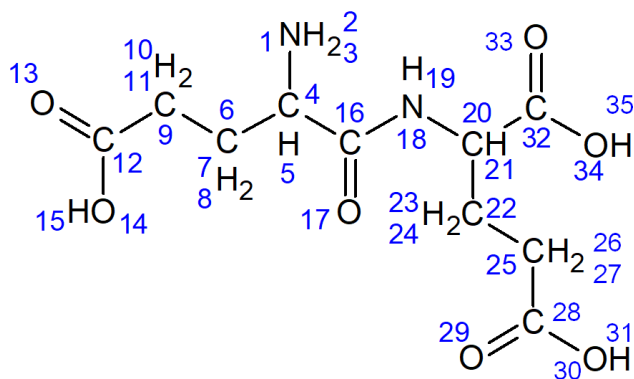
Figure C7.1: All-atom Glutamic Acid dipeptide with atom numbering
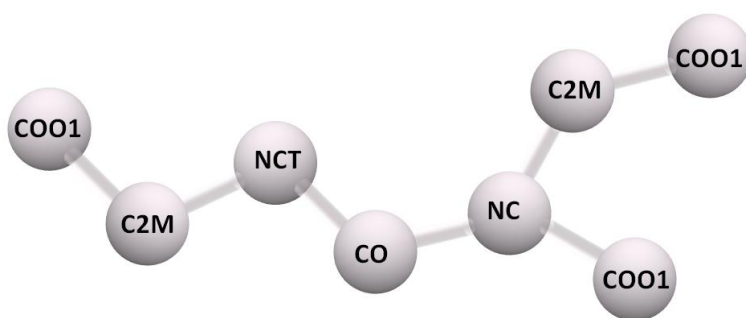
Figure C7.2: CG model of Glutamic Acid dipeptide showing bead types
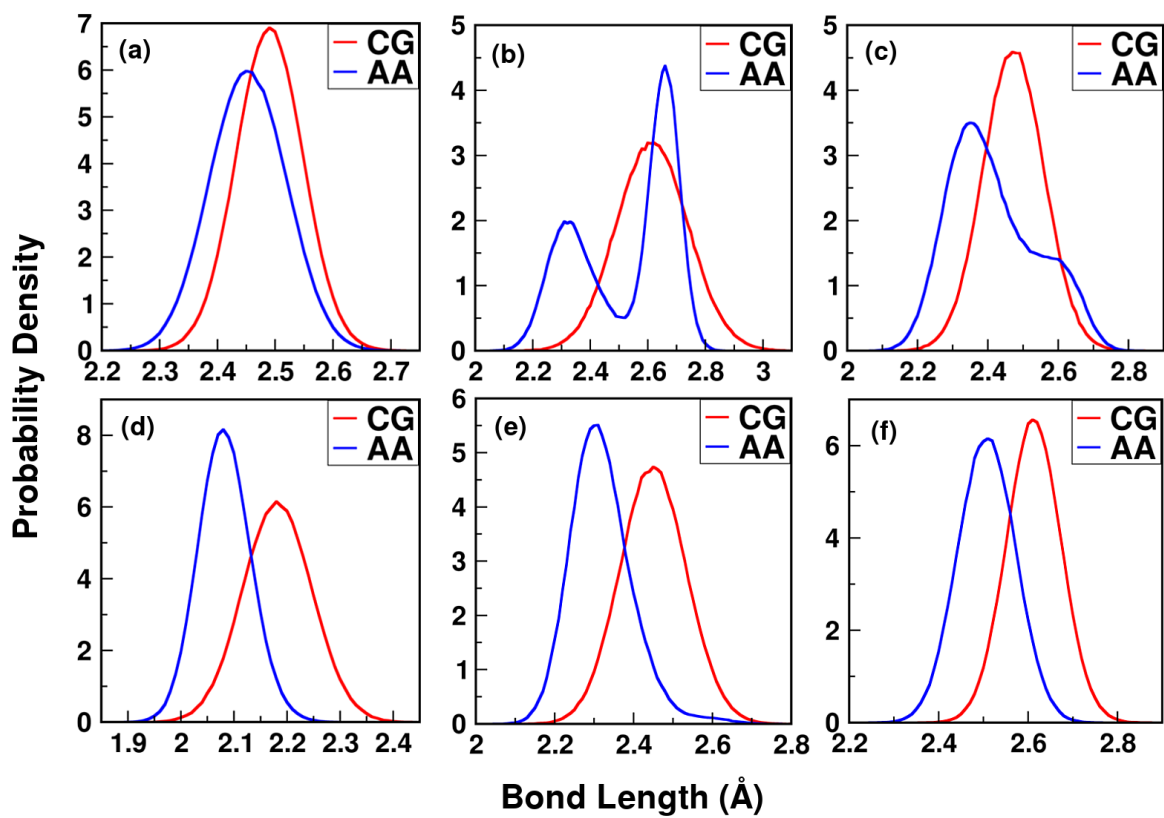
Figure C7.3: Distributions of bonds (a) COO1-C2M, (b) C2M-NCT, (c) NCT-CO, (d) CO-NC, (e) NC-C2M, and (f) NC-COO1
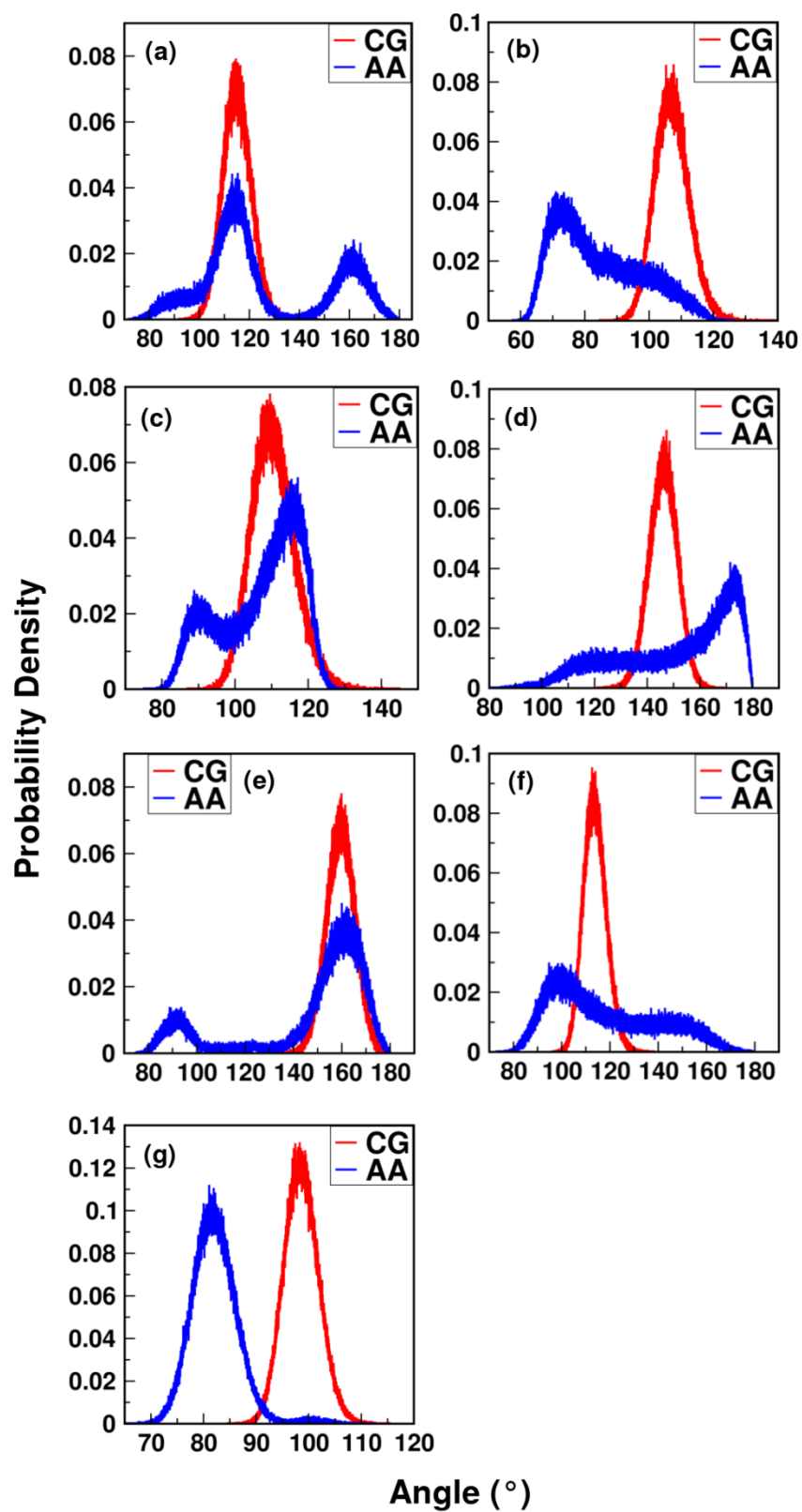
Figure C7.4: Distributions of angles (a) COO1-C2M-NCT, (b) C2M-NCT-CO, (c) NCT-CO-NC, (d) CO-NC-C2M, (e) NC-C2M-COO1, (f) CO-NC-COO1, and (g) C2M-NC-COO1
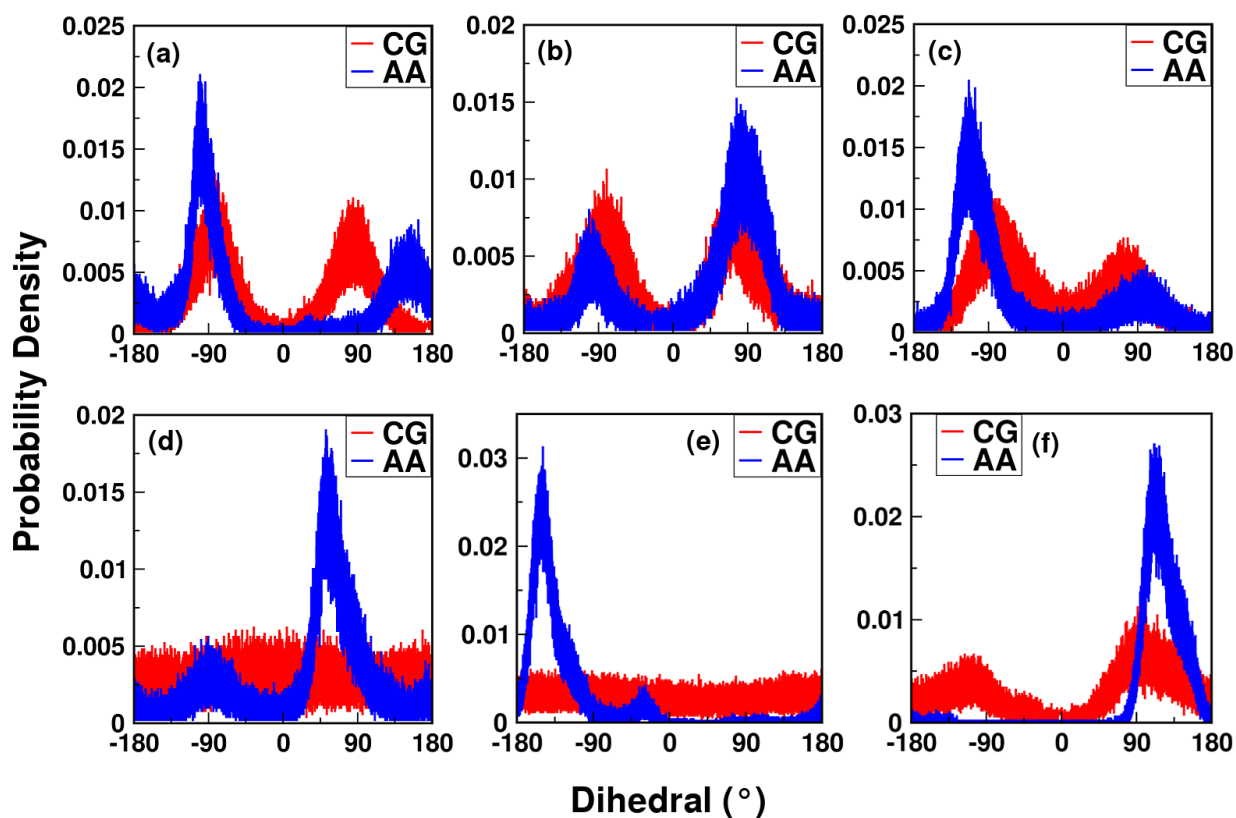
Figure C7.5: Distributions of dihedrals (a) COO1-C2M-NCT-CO, (b) C2M-NCT-CO-NC, (c) NCT-CO-NC-C2M, (d) CO-NC-C2M-COO1, (e) COO1-C2M-NC-COO1, and (f) NCT-CO-NC-COO1
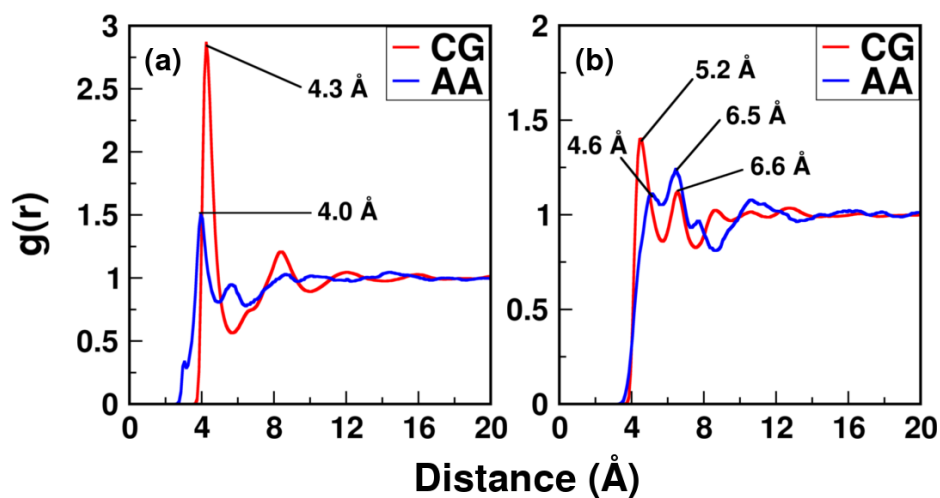
Figure C7.6: RDFs of (a) COO1 and (b) C2M

C8: Glutamine

Gln-Gln Dipeptide
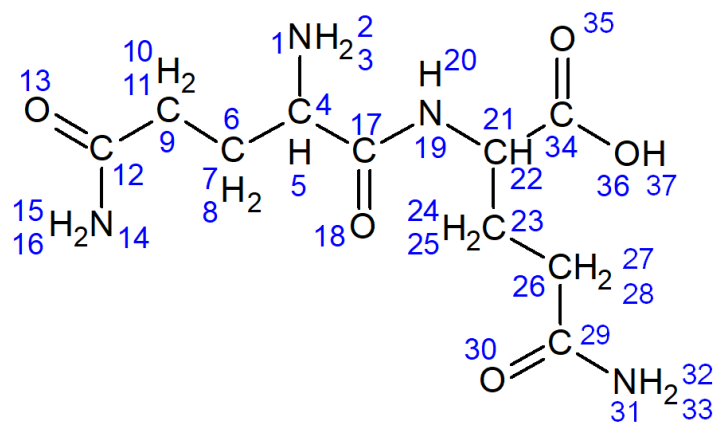
Figure C8.1: All-atom Glutamine dipeptide with atom numbering
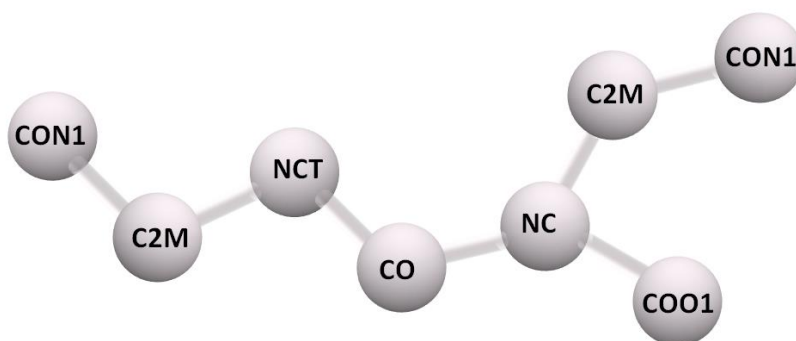
Figure C8.2: CG model of Glutamine dipeptide showing bead types
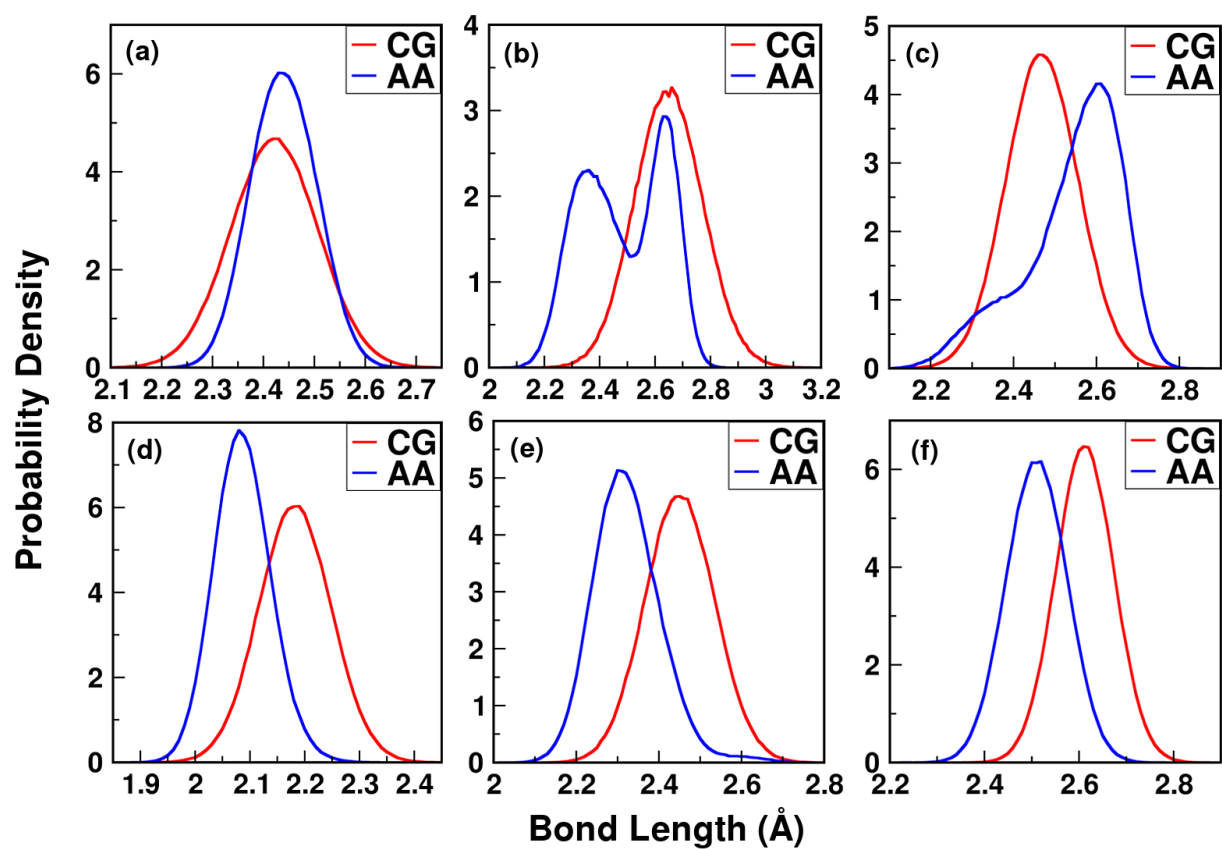
Figure C8.3: Distributions of bonds (a) CON1-C2M, (b) C2M-NCT, (c) NCT-CO, (d) CO-NC, (e) NC-C2M, and (f) NC-COO1
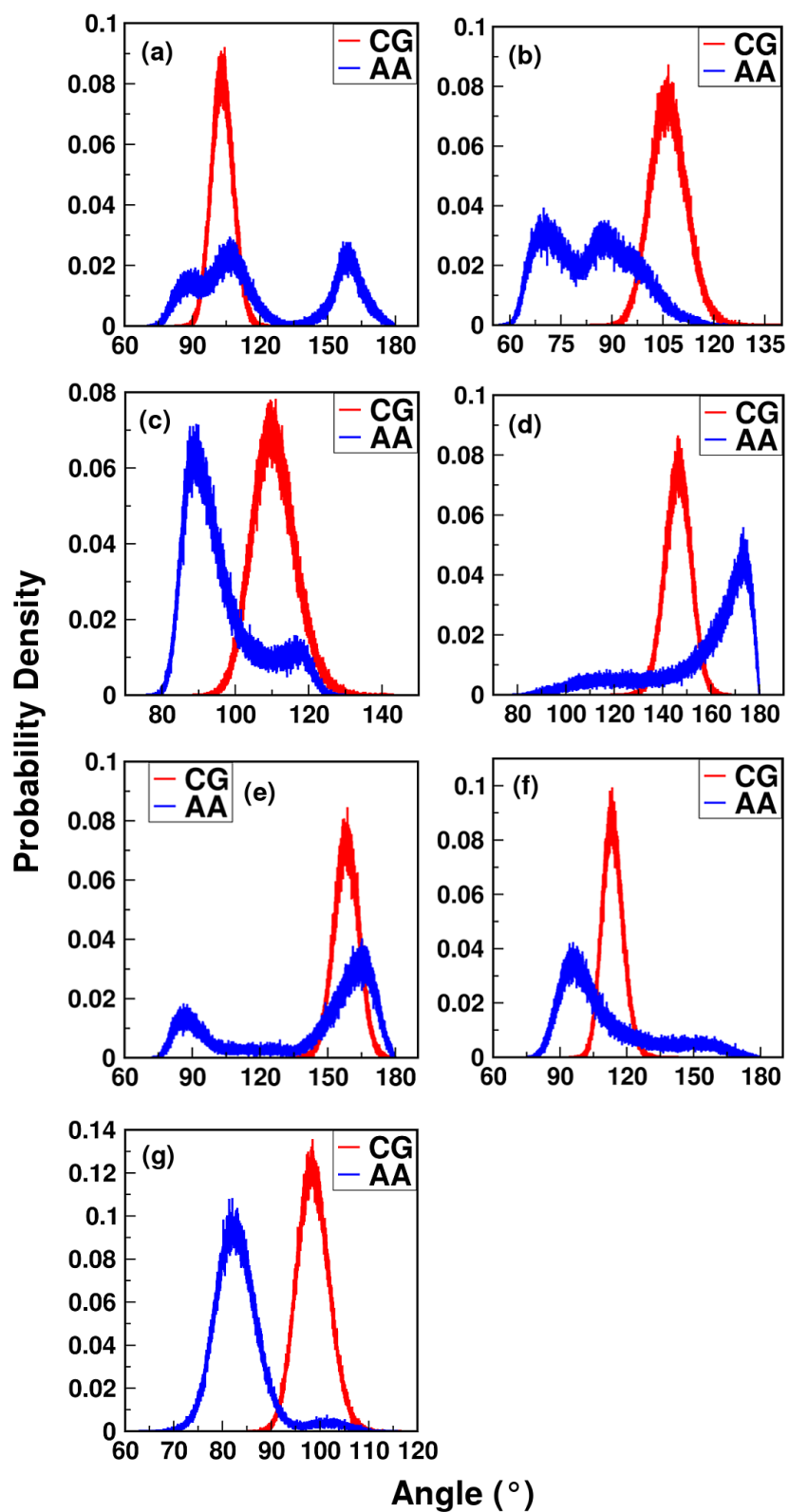
Figure C8.4: Distributions of angles (a) CON1-C2M-NCT, (b) C2M-NCT-CO, (c) NCT-CO-NC, (d) CO-NC-C2M, (e) NC-C2M-CON1, (f) CO-NC-COO1, and (g) C2M-NC-COO1
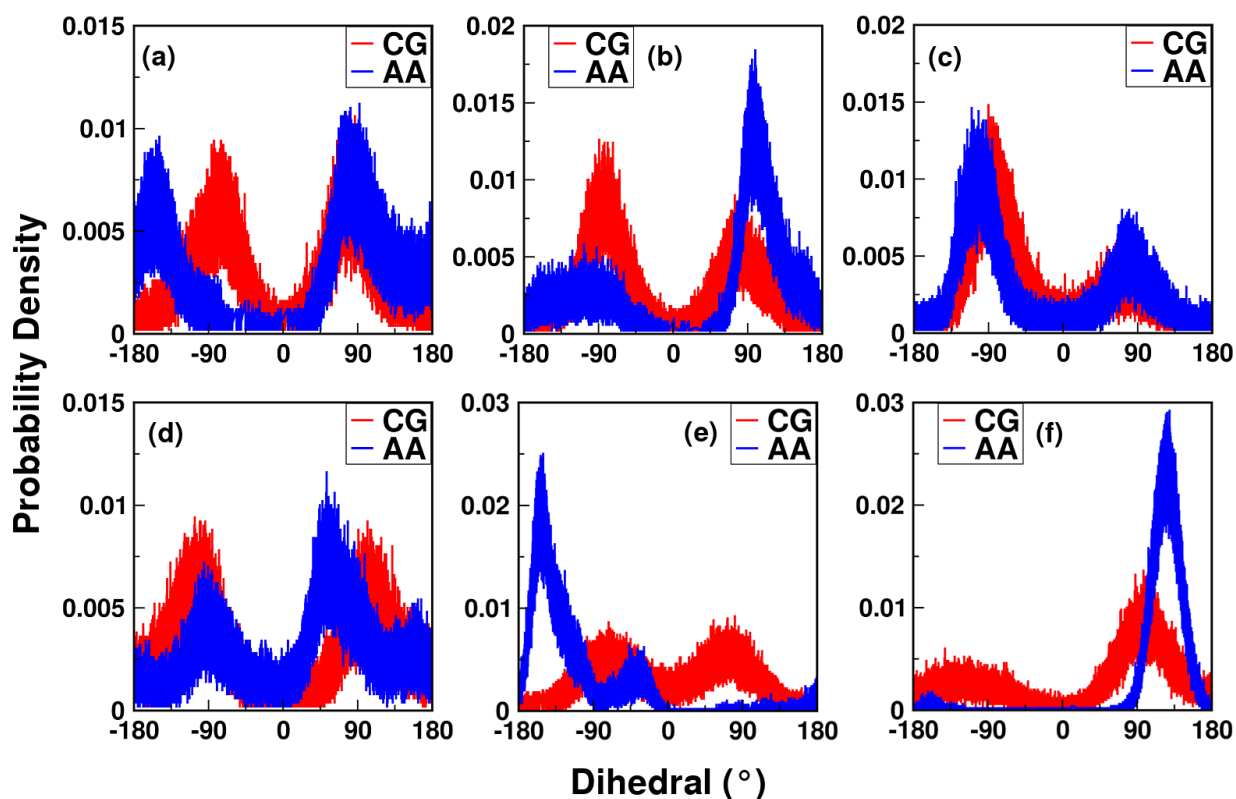
Figure C8.5: Distributions of dihedrals (a) CON1-C2M-NCT-CO, (b) C2M-NCT-CO-NC, (c) NCT-CO-NC-C2M, (d) CO-NC-C2M-CON1, (e) CON1-C2M-NC-COO1, and (f) NCT-CO-NC-COO1
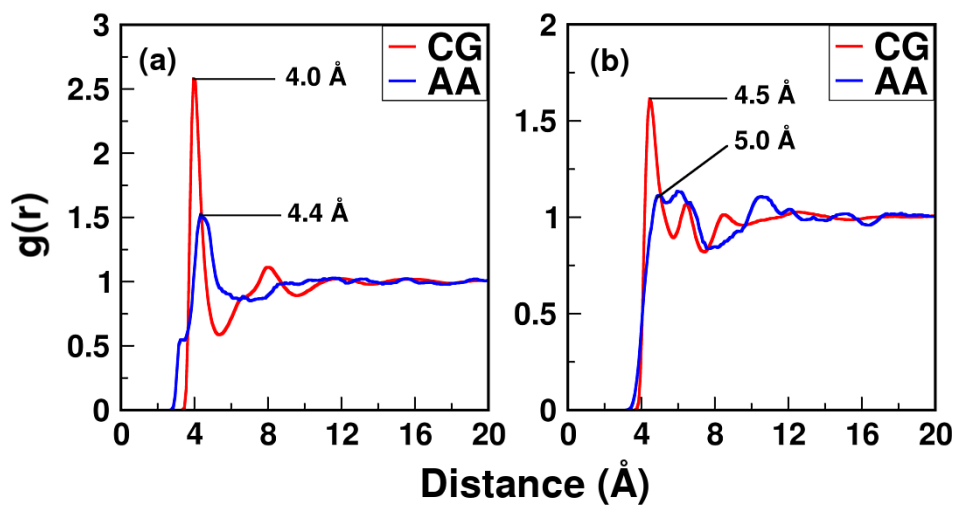
Figure C8.6: RDFs of (a) CON1 and (b) C2M

Analogue: Propionamide
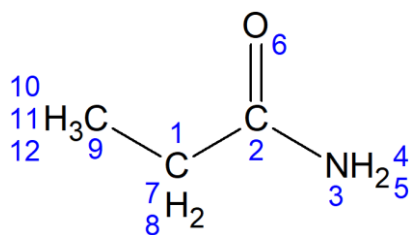
Figure C8.7: All-atom propionamide with atom numbering
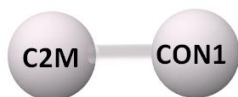
Figure C8.8: CG model of propionamide showing bead type
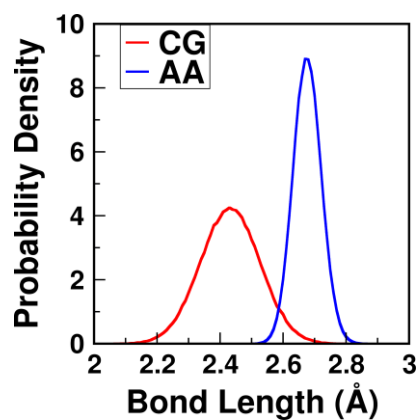
Figure C8.9: Distribution of bond CON1-C2M. The equilibrium bond length from the GlnGln dipeptide was used so as to provide a better match to the glutamine side chain.
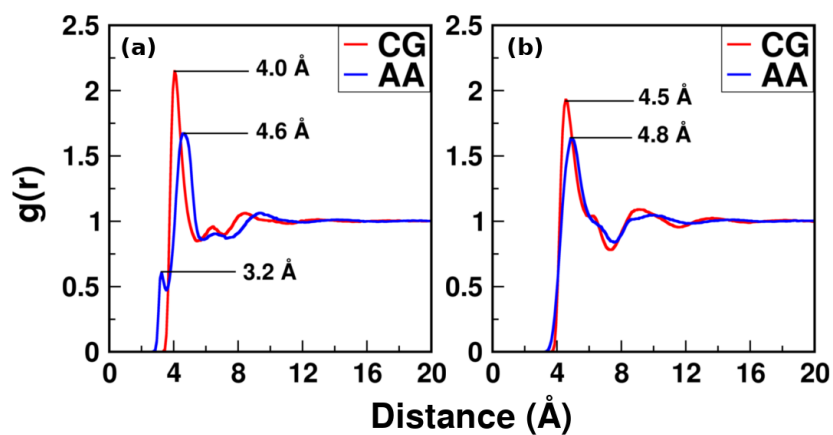
Figure C8.10: RDFs of (a) CON1 and (b) C2M

C9: Glycine

Gly-Gly Dipeptide

Figure C9.1: All-atom Glycine dipeptide with atom numbering
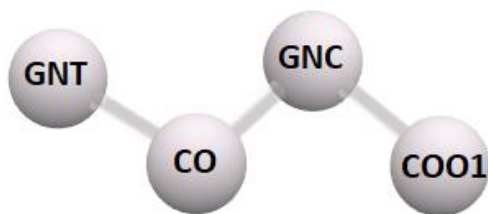
Figure C9.2: CG model of Glycine dipeptide showing bead types
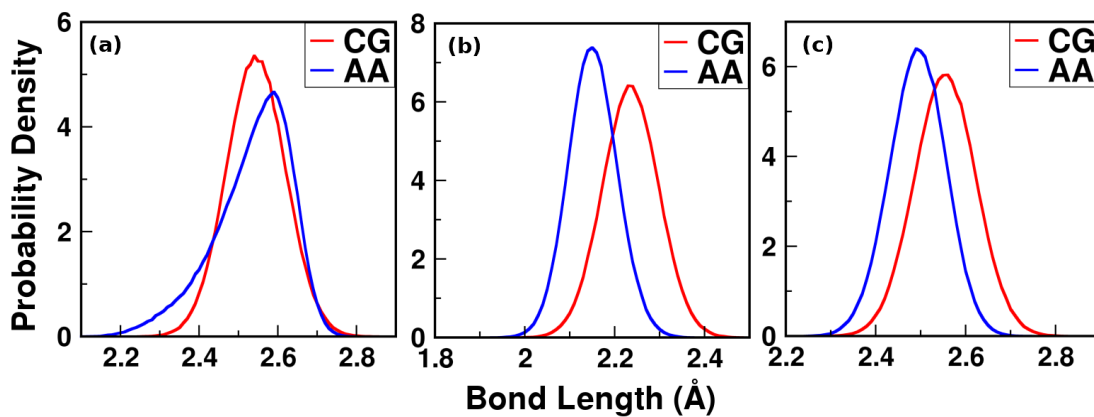
Figure C9.3: Distributions of bonds (a) GNT-CO, (b) CO-GNC, and (c) GNC-COO1
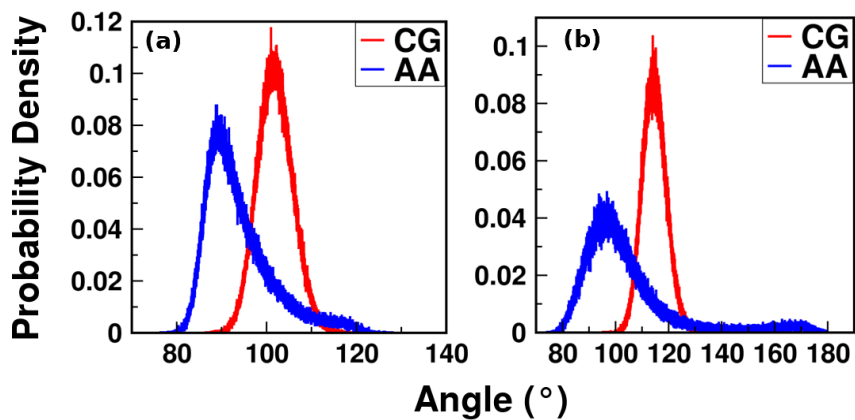
Figure C9.4: Distributions of angles (a) GNT-CO-GNC and (b) CO-GNC-COO1
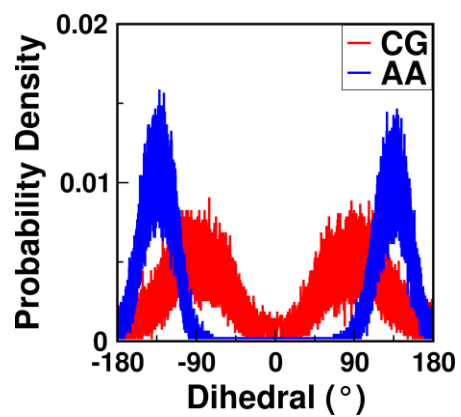
Figure C9.5: Distribution of dihedral GNT-CO-GNC-COO1
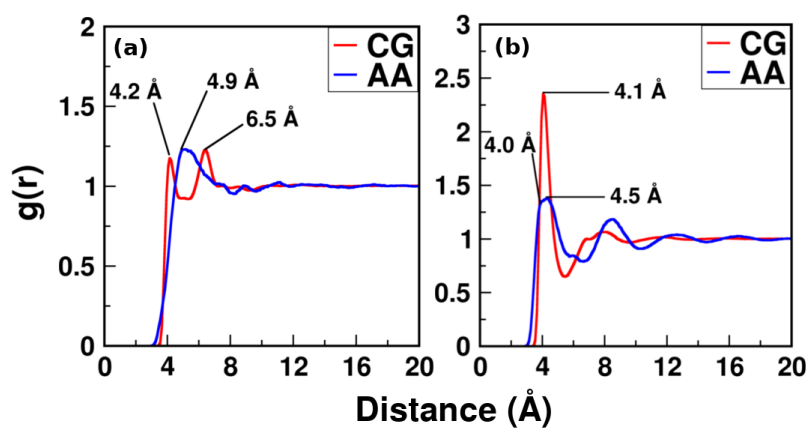
Figure C9.6: RDFs of (a) GNC and (b) GNT

C10: Histidine

His-His Dipeptide
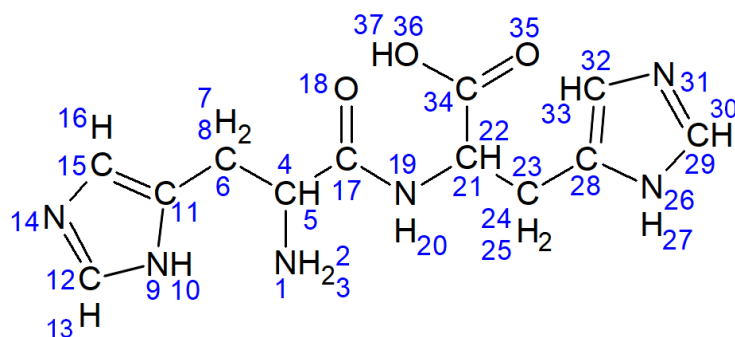
Figure C10.1: All-atom Histidine dipeptide with atom numbering
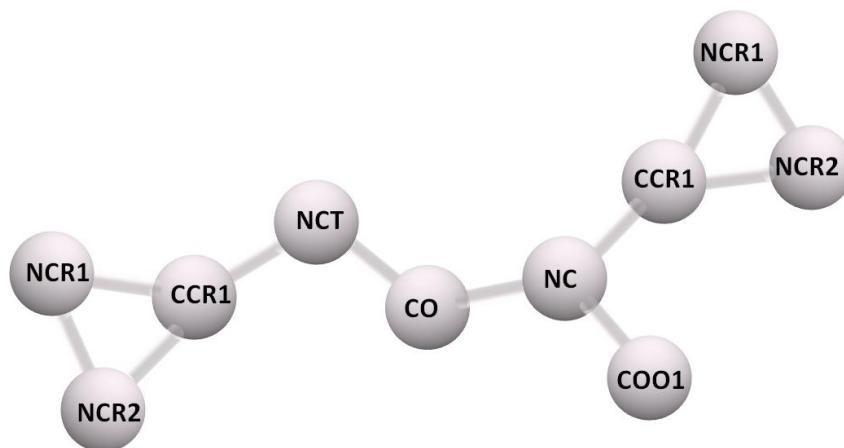
Figure C10.2: CG model of Histidine dipeptide showing bead types
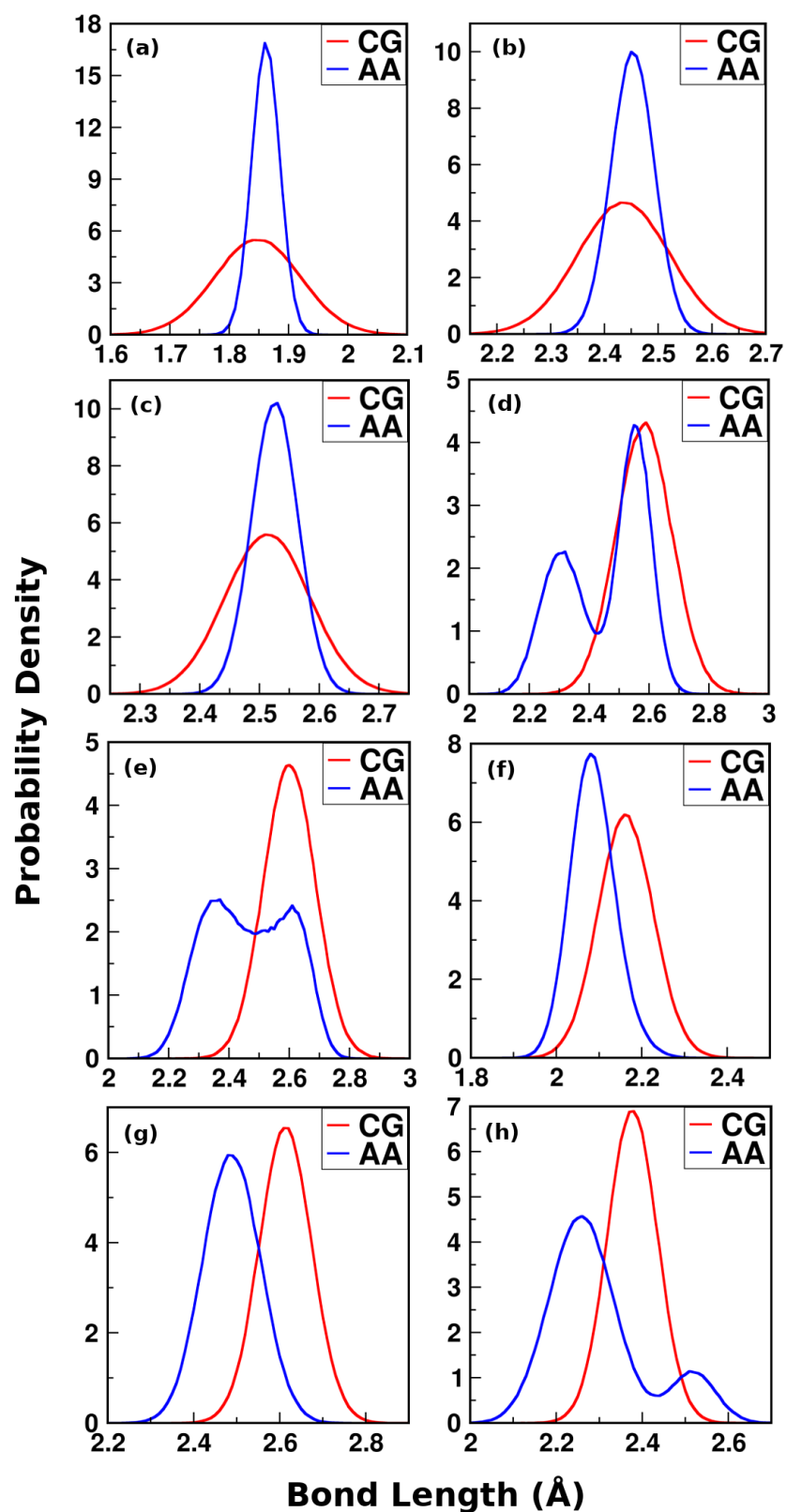
Figure C10.3: Distributions of bonds (a) NCR2-NCR1, (b) NCR1-CCR1, (c) NCR2-CCR1, (d) CCR1-NCT, (e) NCT-CO, (f) CO-NC, (g) NC-COO1, and (h) NC-CCR1
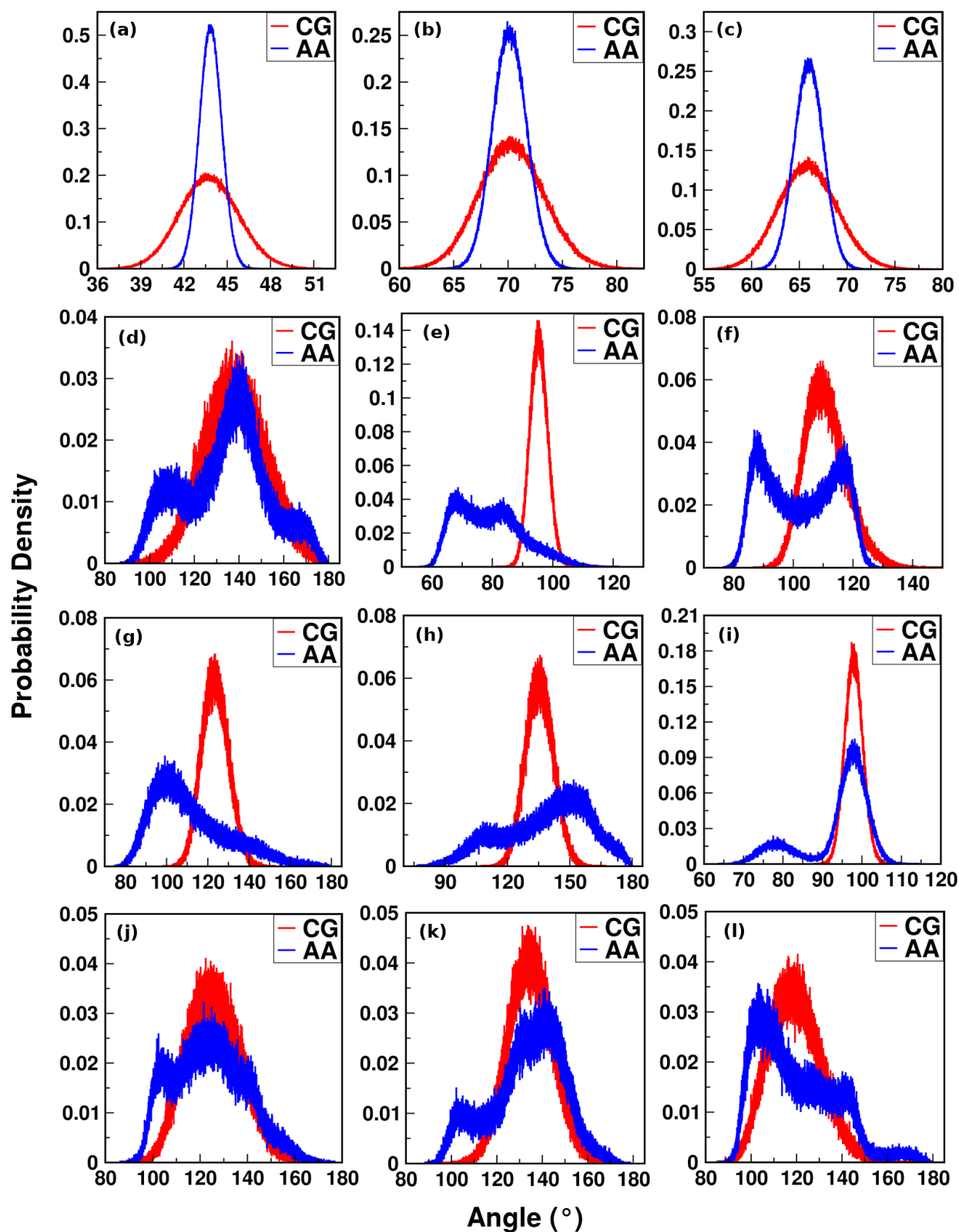
Figure C10.4: Distributions of angles (a) NCR2-CCR1-NCR1, (b) CCR1-NCR1-NCR2, (c) NCR1-NCR2-CCR1, (d) NCR1-CCR1-NCT, (e) CCR1-NCT-CO, (f) NCT-CO-NC, (g) CO-NC-COO1, (h) CO-NC-CCR1, (i) COO1-NC-CCR1, (j) NC-CCR1-NCR2, (k) NCR2-CCR1-NCT, and (l) NC-CCR1-NCR1
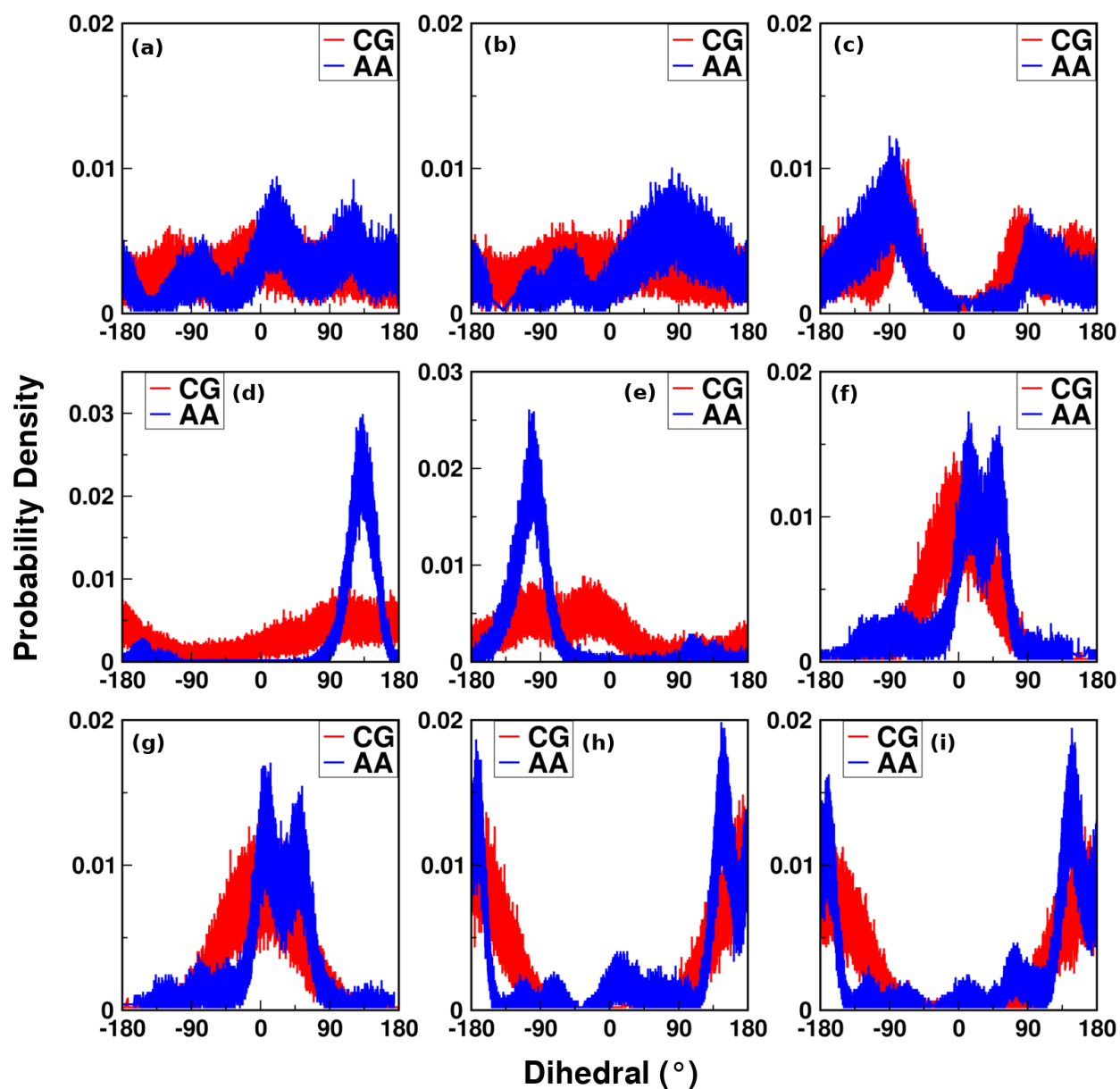
Figure C10.5: Distribution of dihedrals (a) NCR1-CCR1-NCT-CO, (b) NCR2-CCR1-NCT-CO, (c) CCR1-NCT-CO-NC, (d) NCT-CO-NC-COO1, (e) NCT-CO-NC-CCR1, (f) CO-NC-CCR1-NCR2, (g) CO-NC-CCR1-NCR1, (h) COO1-NC-CCR1-NCR2, and (i) COO1-NC-CCR1-NCR1
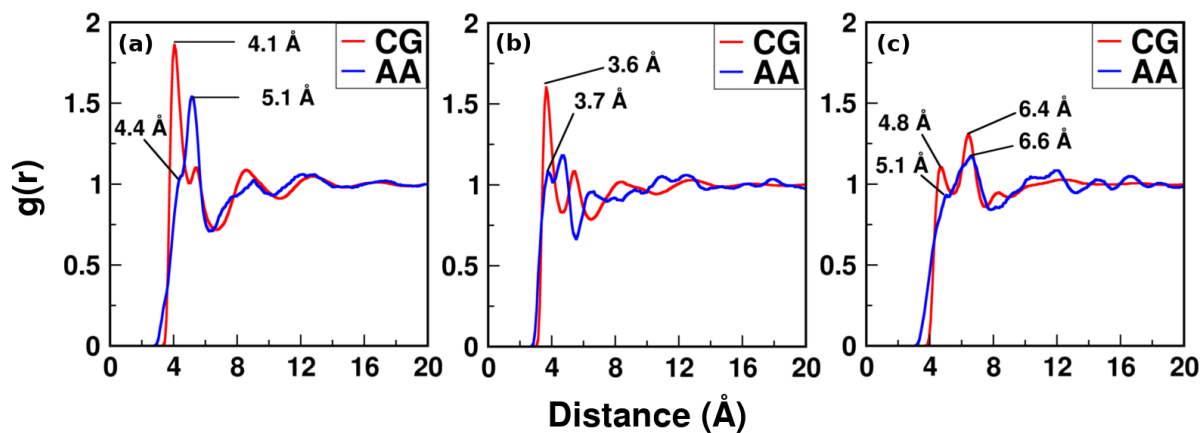
Figure C10.6: RDFs of (a) CCR1, (b) NCR1, and (c) NCR2

Analogue 1: Pyridine
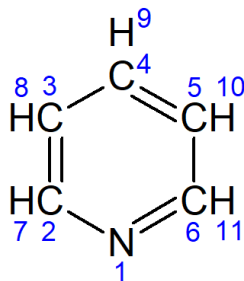
Figure C10.7: All-atom pyridine with atom numbering
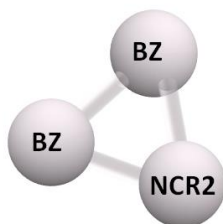
Figure C10.8: CG model of pyridine showing bead types
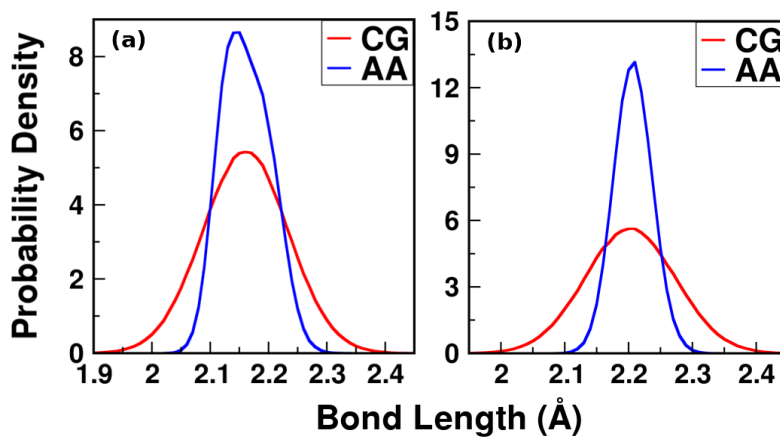
Figure C10.9: Distributions of bonds (a) NCR2-BZ and (b) BZ-BZ

122
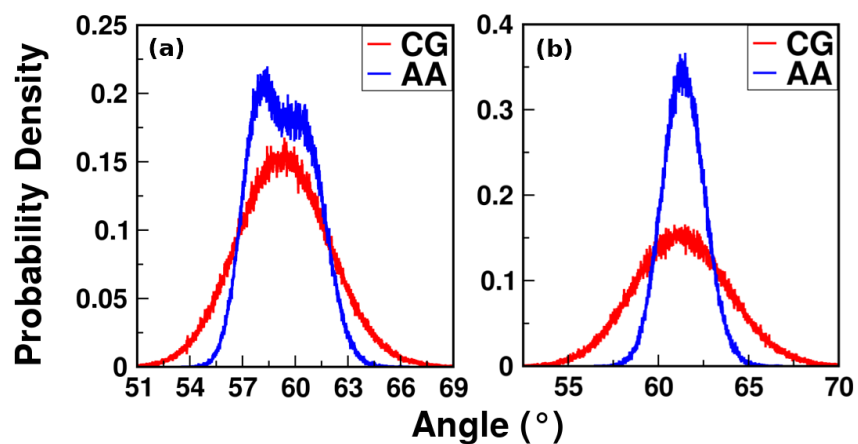
Figure C10.10: Distributions of angles (a) NCR2-BZ-BZ and (b) BZ-NCR2-BZ
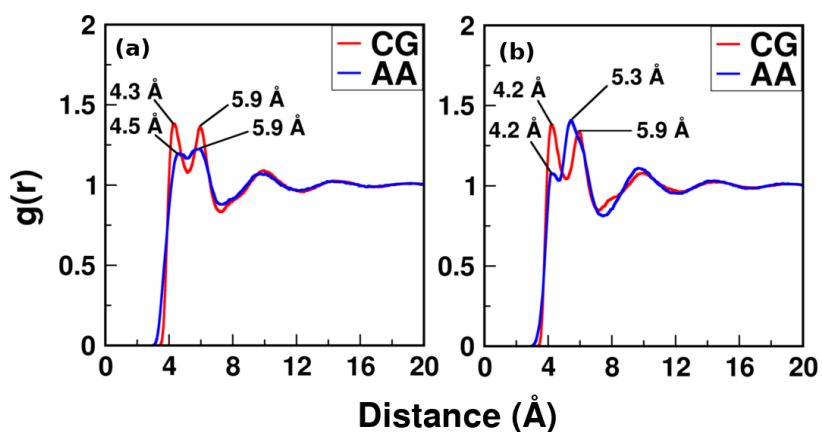
Figure C10.11: RDFs of (a) BZ and (b) NCR2

Analogue 2: Pyrrole
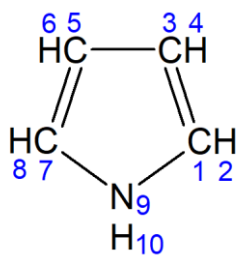
Figure C10.12: All-atom pyrrole with atom numbering
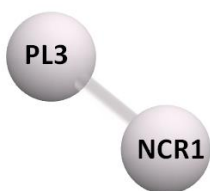
Figure C10.13: CG model of pyrrole showing bead types

123
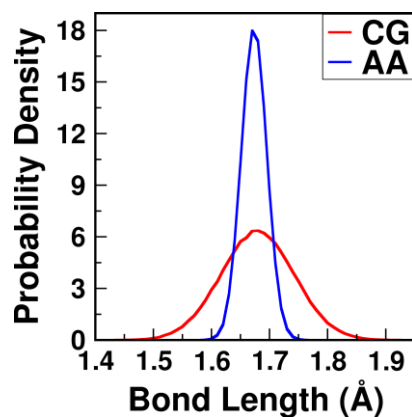
Figure C10.14: Distribution of bond NCR1-PL3
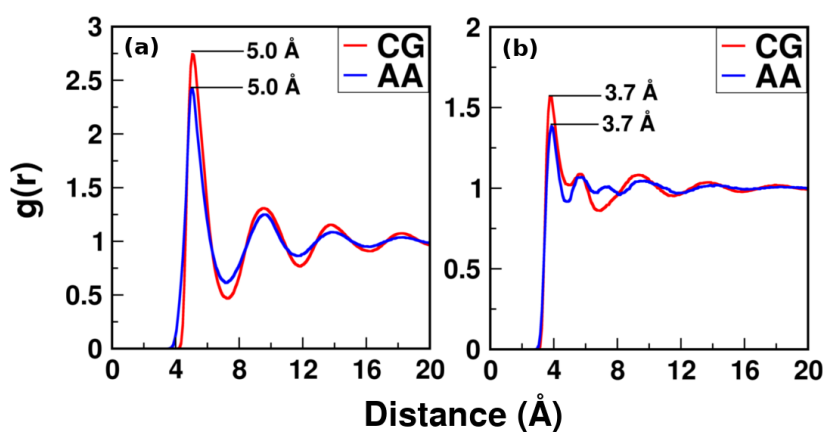
Figure C10.15: RDFs of (a) PL3 and (b) NCR1

Analogue 3: Toluene
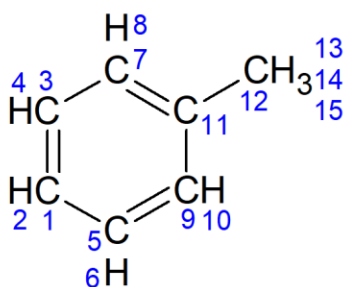
Figure C10.16: All-atom toluene with atom numbering
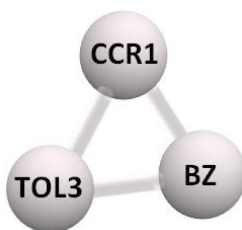
Figure C10.17: CG model of toluene showing bead types
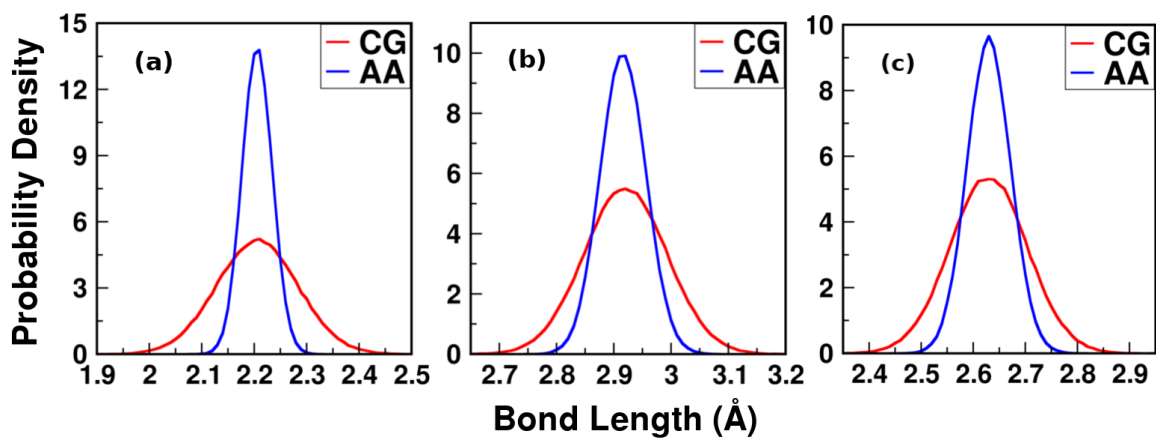
Figure C10.18: Distribution of bonds (a) TOL3-BZ (b) TOL3-CCR1, and (c) BZ-CCR1
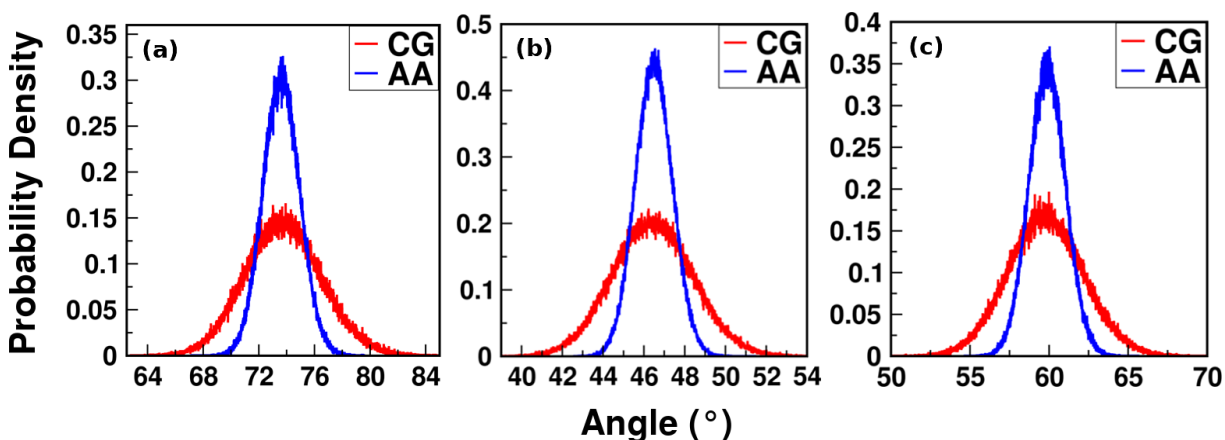
Figure C10.19: Distributions of angles (a) TOL3-BZ-CCR1, (b) TOL3-CCR1-BZ, and (c) BZ-TOL3-CCR1
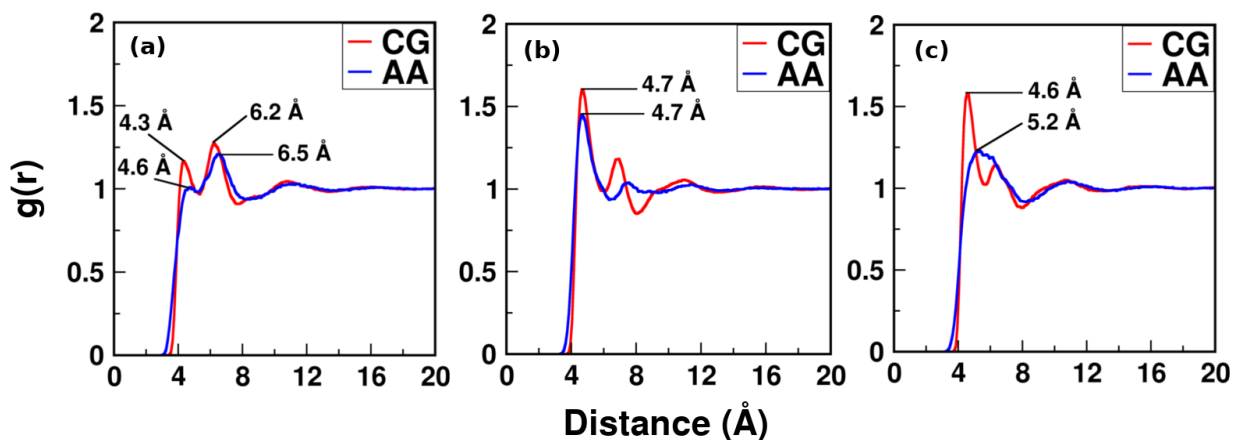
Figure C10.20: RDFs of (a) BZ, (b) CCR1, and (c) TOL3

C11: Isoleucine

Ile-Ile Dipeptide
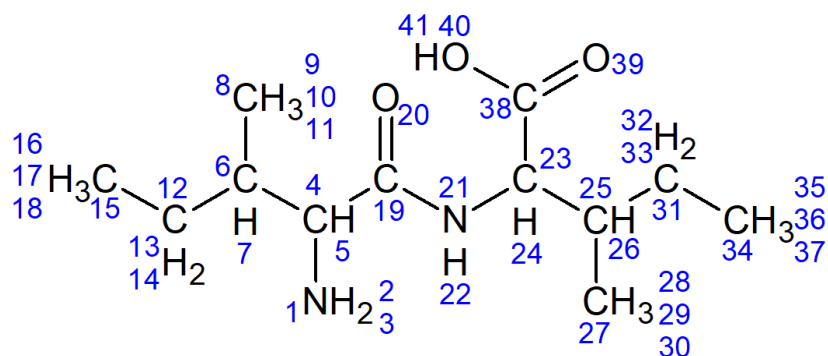
Figure C11.1: All-atom Isoleucine dipeptide with atom numbering
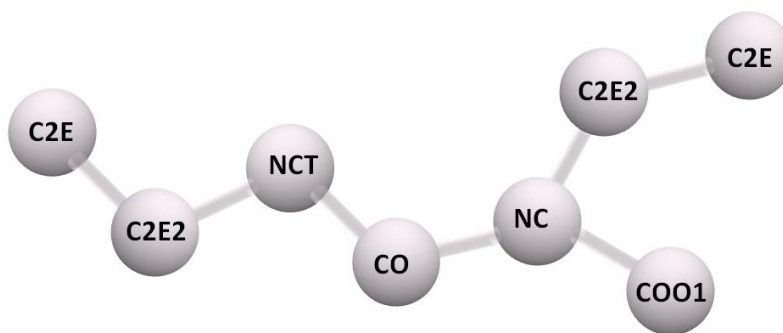
Figure C11.2: CG model of Isoleucine dipeptide showing bead types
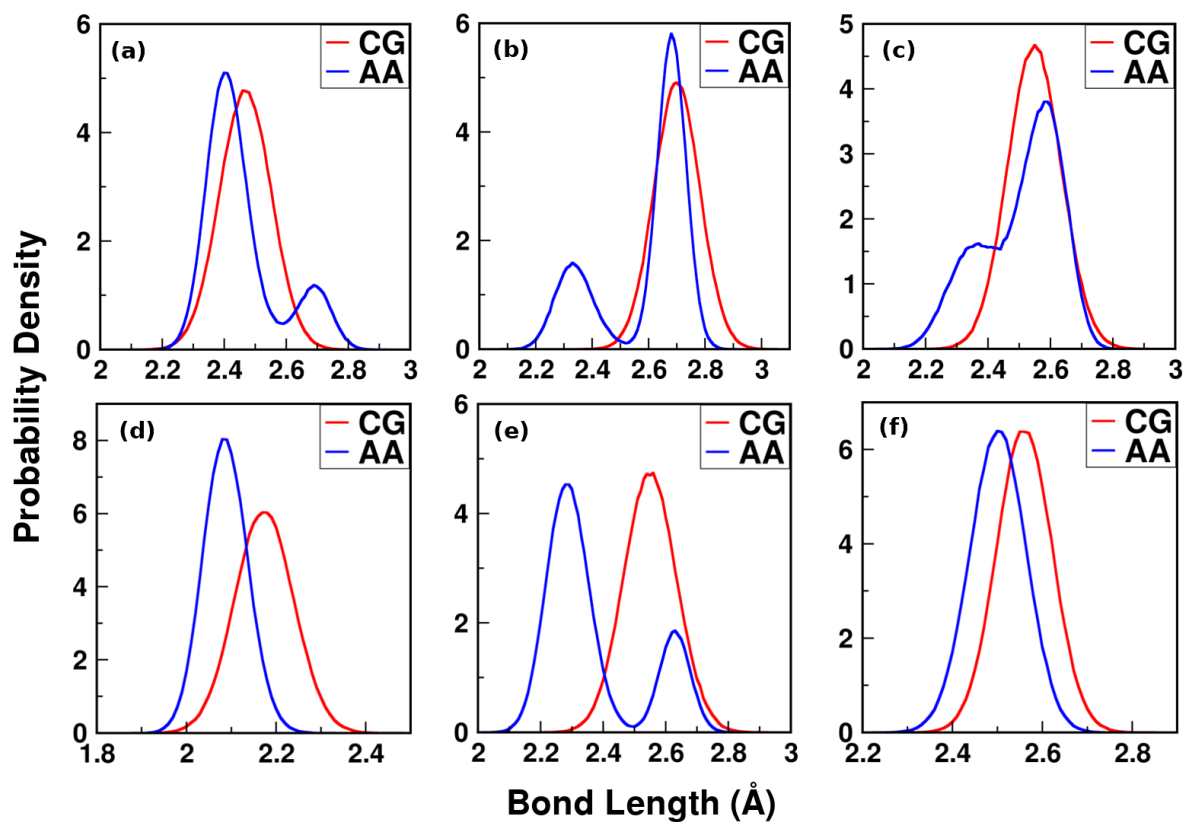
Figure C11.3: Distributions of bonds (a) C2E-C2E2, (b) C2E2-NCT, (c) NCT-CO, (d) CO-NC, (e) NC-C2E2, and (f) NC-COO1
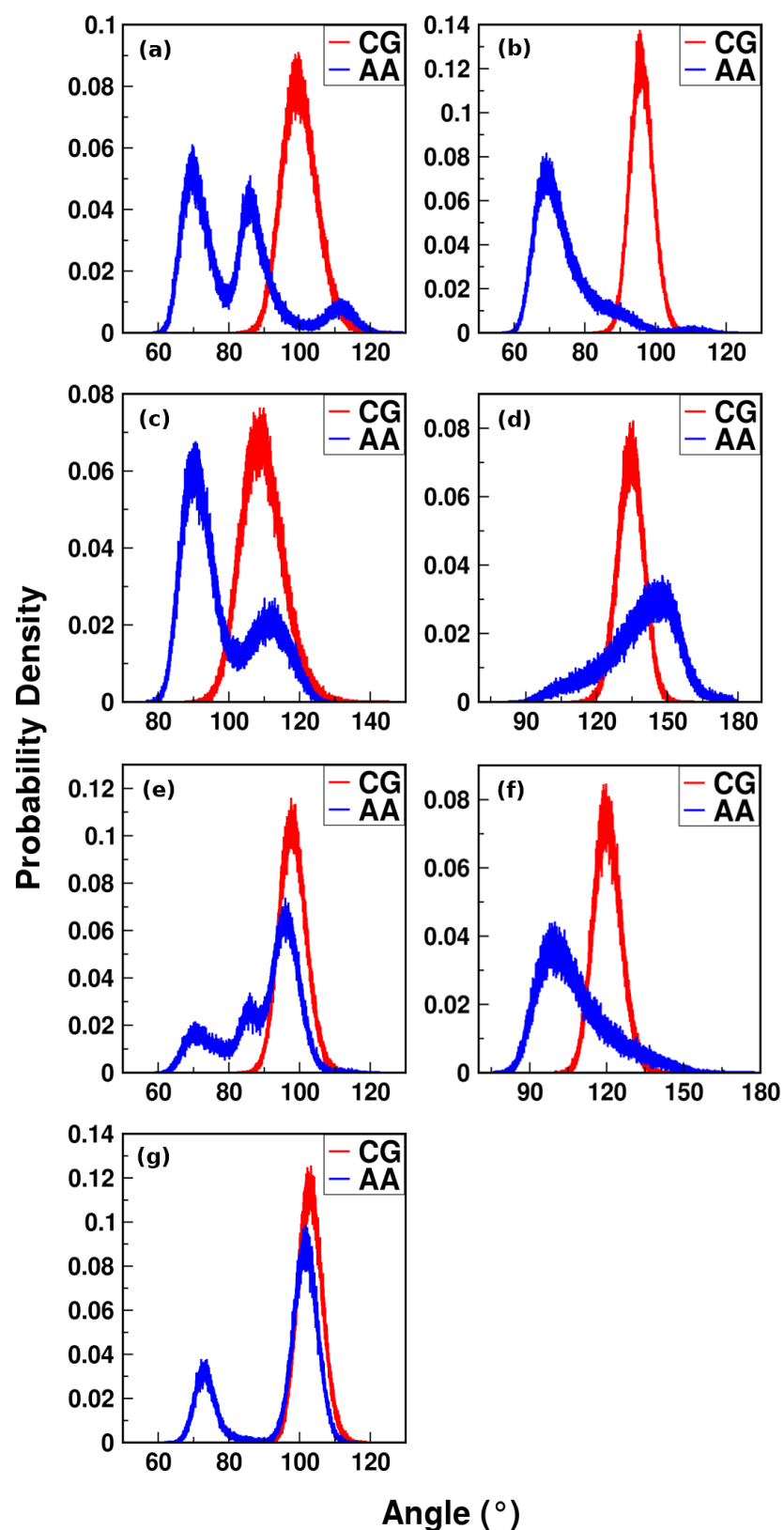
Figure C11.4: Distribution of angles (a) C2E-C2E2-NCT, (b) C2E2-NCT-CO, (c) NCT-CO-NC, (d) CO-NC-C2E2, (e) NC-C2E2-C2E, (f) CO-NC-COO1, and (g) C2E2-NC-COO1
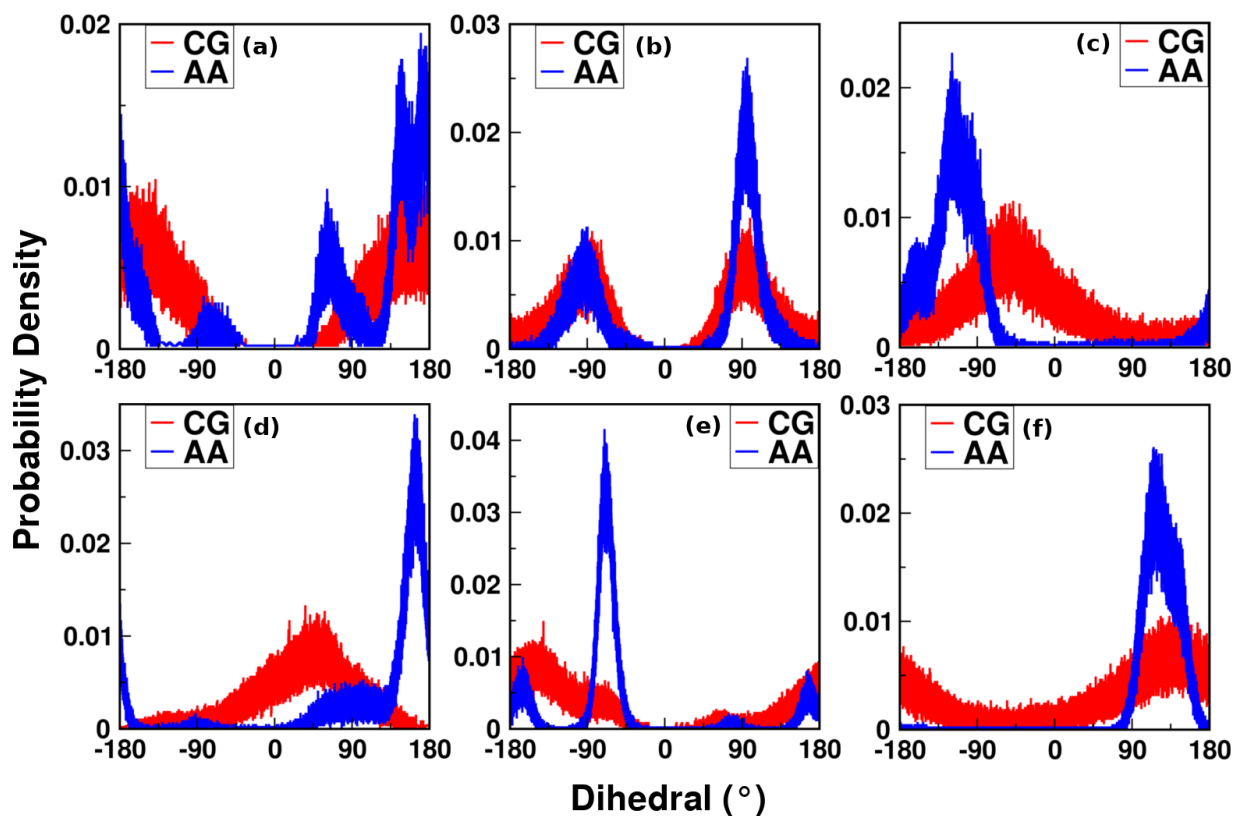
Figure C11.5: Distributions of dihedrals (a) C2E-C2E2-NCT-CO, (b) C2E2-NCT-CO-NC, (c) NCT-CO-NC-C2E2, (d) CO-NC-C2E2-C2E, (e) C2E-C2E2-NC-COO1, and (f) NCT-CO-NC-COO1
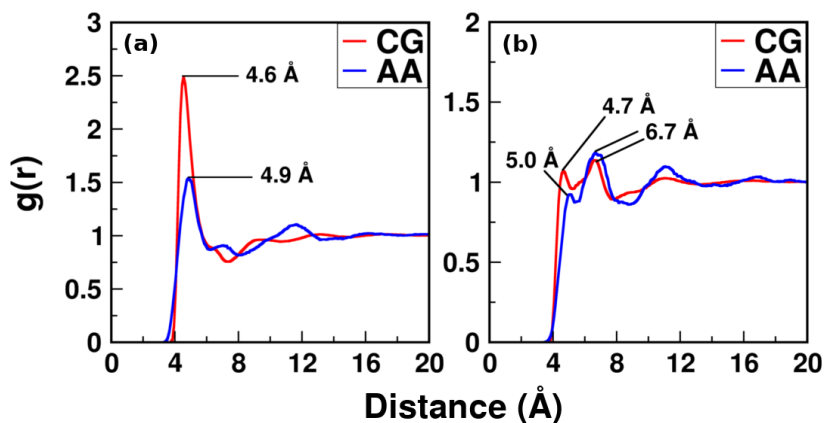

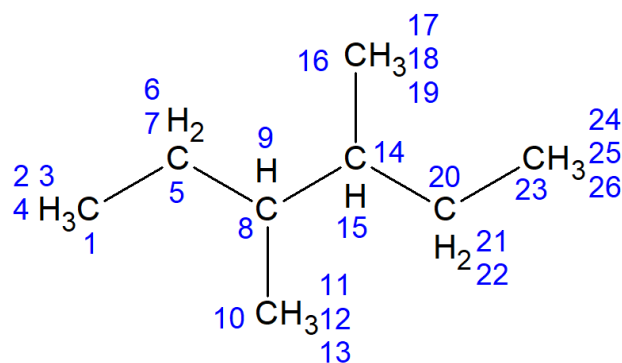**Figure C11.6:** RDFs of (a) C2E and (b) C2E2

Analogue: 3,4-dimethylhexane

Figure C11.7: All-atom 3,4-dimethylhexane with atom numbering
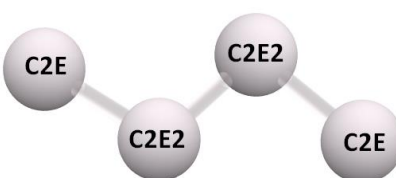

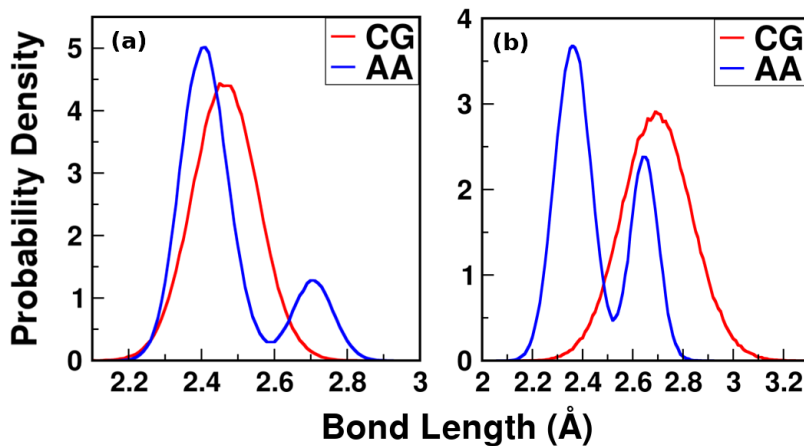**Figure C11.8:** CG model of 3,4-dimethylhexane showing bead types

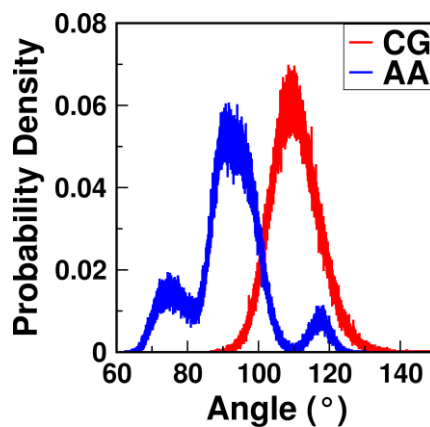**Figure C11.9:** Distributions of bonds (a) C2E-C2E2 and (b) C2E2-C2E2
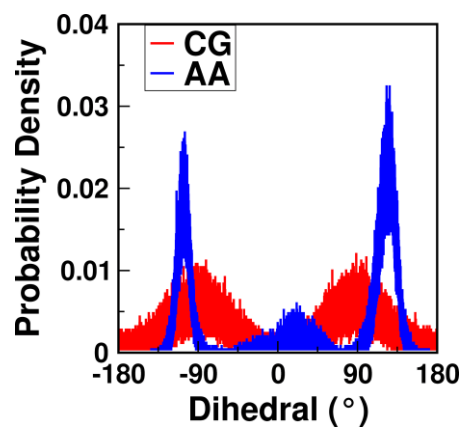

Figure C11.10: Distribution of angle C2E-C2E2-C2E2

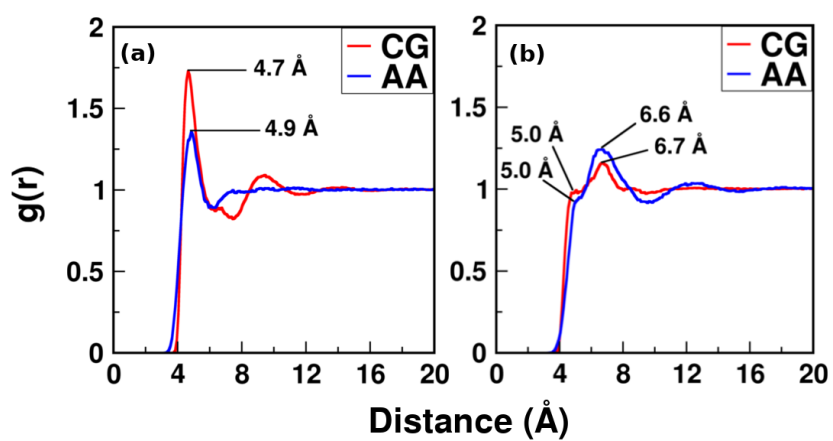**Figure C11.11:** Distribution of dihedral C2E-C2E2-C2E2-C2E

Figure C11.12: RDFs of (a) C2E and (b) C2E2
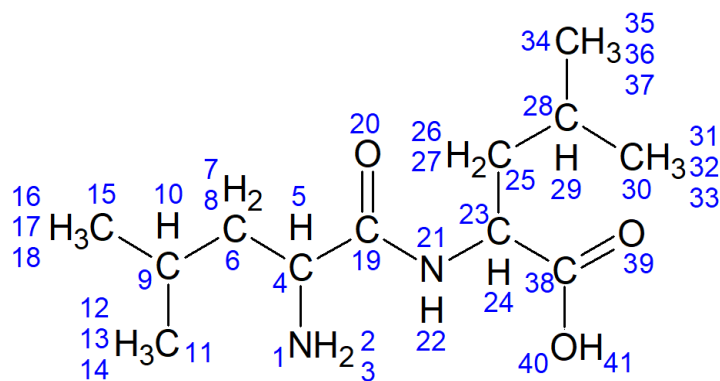
C12: Leucine

Leu-Leu Dipeptide

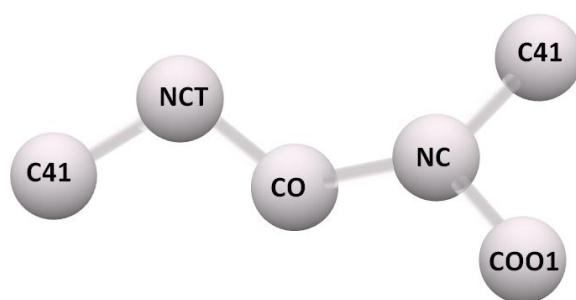**Figure C12.1:** All-atom Leucine dipeptide with atom numbering

Figure C12.2: CG model of Leucine dipeptide showing bead types
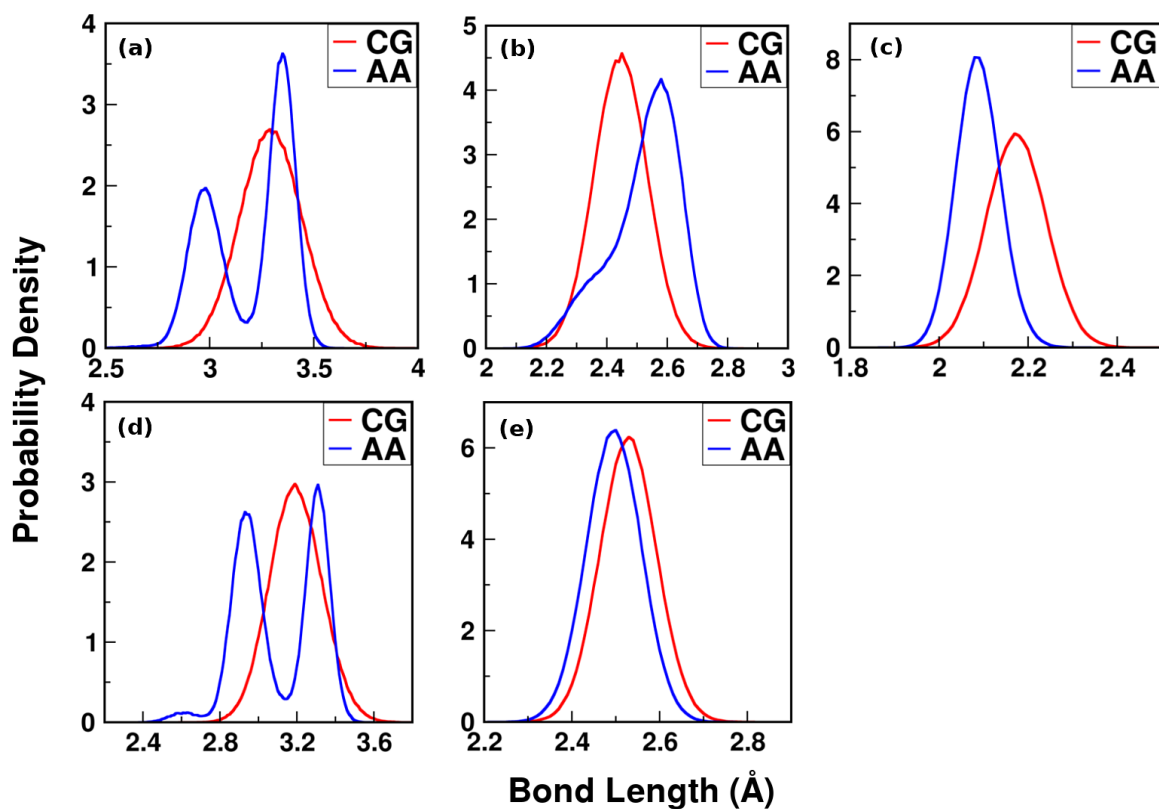
Figure C12.3: Distribution of bonds (a) C41-NCT, (b) NCT-CO, (c) CO-NC, (d) NC-C41, and (e) NC-COO1
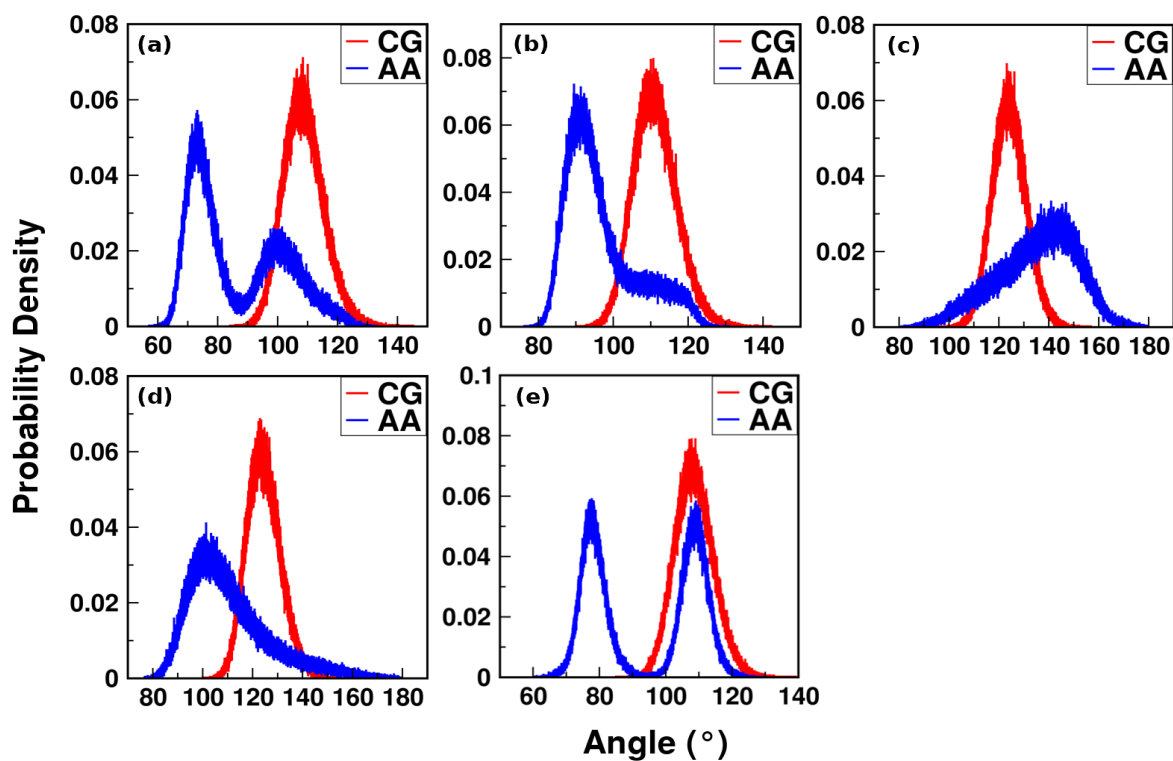
Figure C12.4: Distribution of angles (a) C41-NCT-CO, (b) NCT-CO-NC, (c) CO-NC-C41, (d) CO-NC-COO1, and (e) C41-NC-COO1
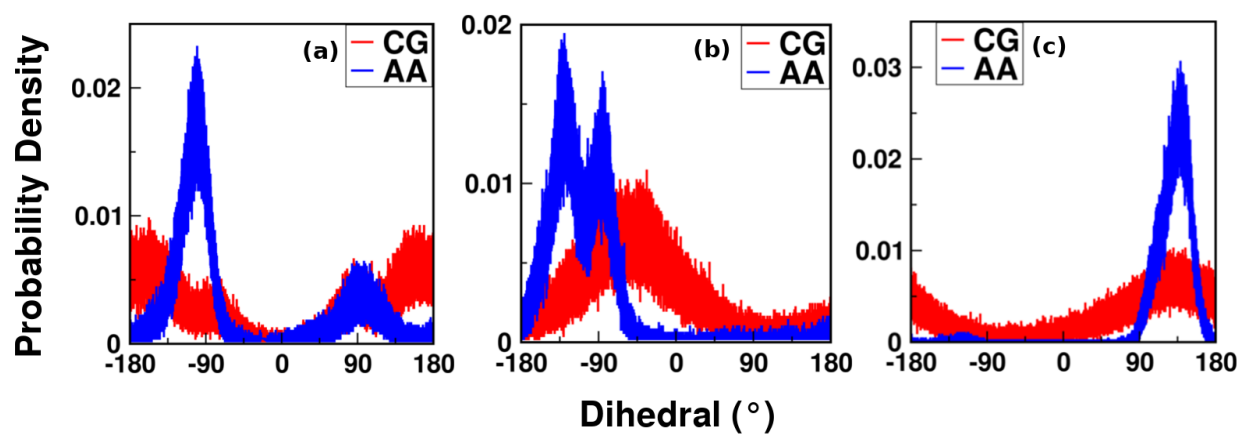
Figure C12.5: Distribution of dihedrals (a) C41-NCT-CO-NC, (b) NCT-CO-NC-C41, and (c) NCT-CO-NC-COO1
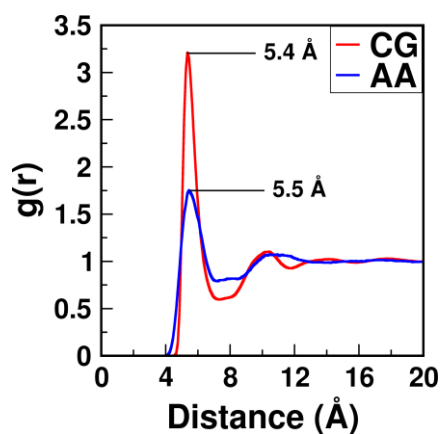
Figure C12.6: RDF of C41

Analogue: 2,5-dimethylhexane
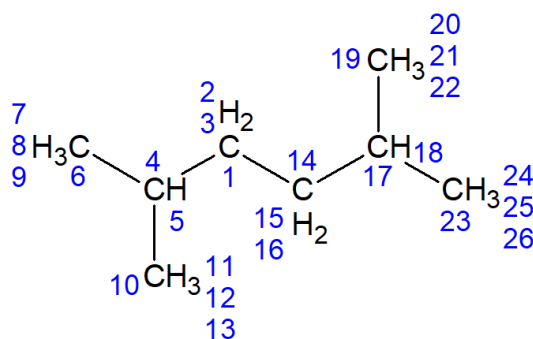
Figure C12.7: All-atom 2,5-dimethylhexane with atom numbering

Figure C12.8: CG model of 2,5-dimethylhexane showing bead types
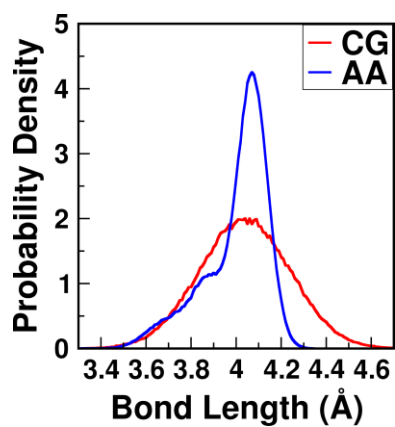
Figure C12.9: Distribution of bond C41-C41
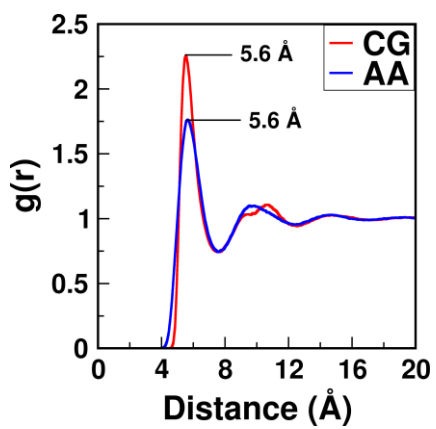
Figure C12.10: RDF of C41

C13: Lysine

Lys-Lys Dipeptide
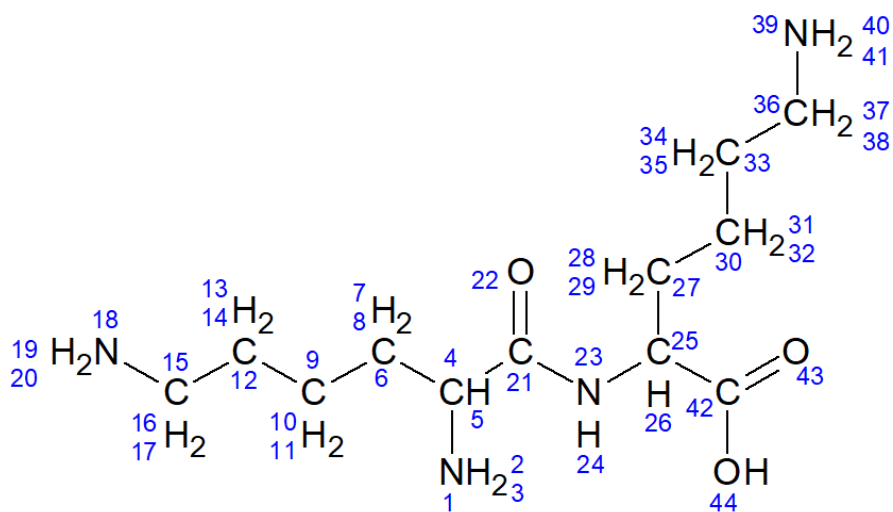
Figure C13.1: All-atom Lysine dipeptide with atom numbering
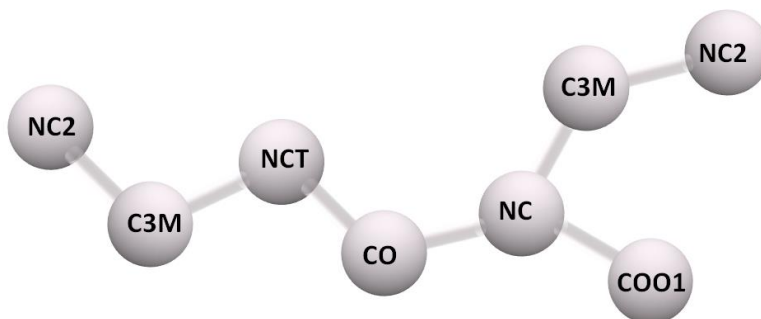
Figure C13.2: CG model of Lysine dipeptide
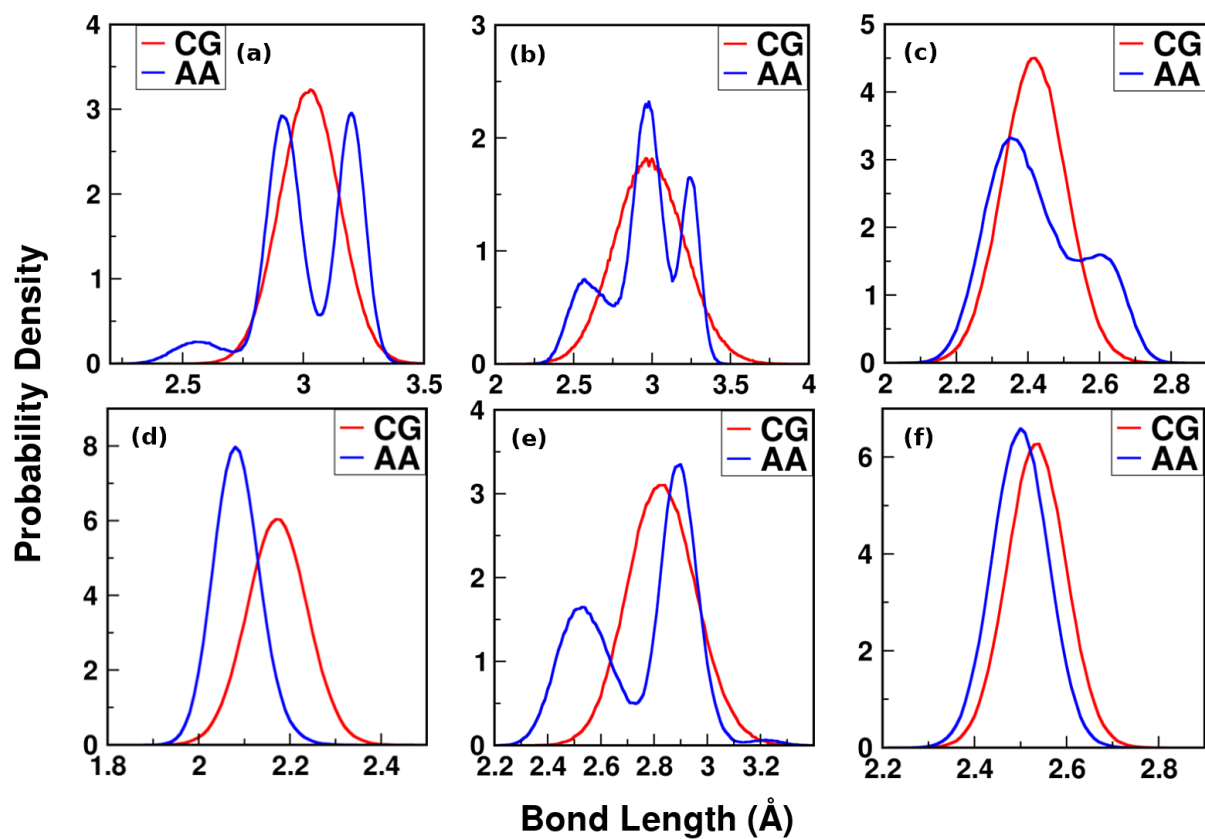
Figure C13.3: Distributions of bonds (a) NC2-C3M, (b) C3M-NCT, (c) NCT-CO, (d) CO-NC, (e) NC-C3M, and (f) NC-COO1
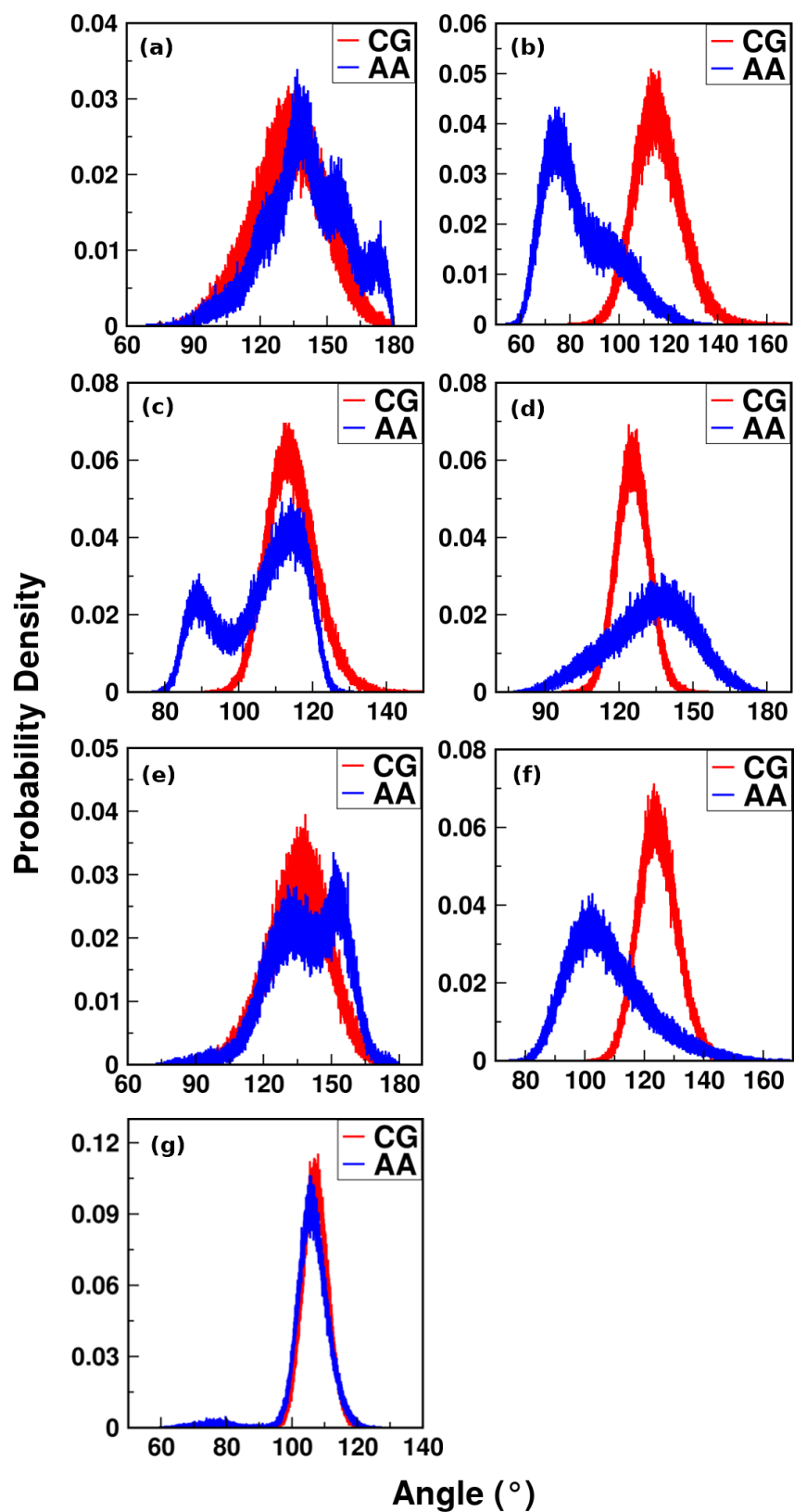
Figure C13.4: Distributions of angles (a) NC2-C3M-NCT, (b) C3M-NCT-CO, (c) NCT-CO-NC, (d) CO-NC-C3M, (e) NC-C3M-NC2, (f) CO-NC-COO1, and (g) C3M-NC-COO1
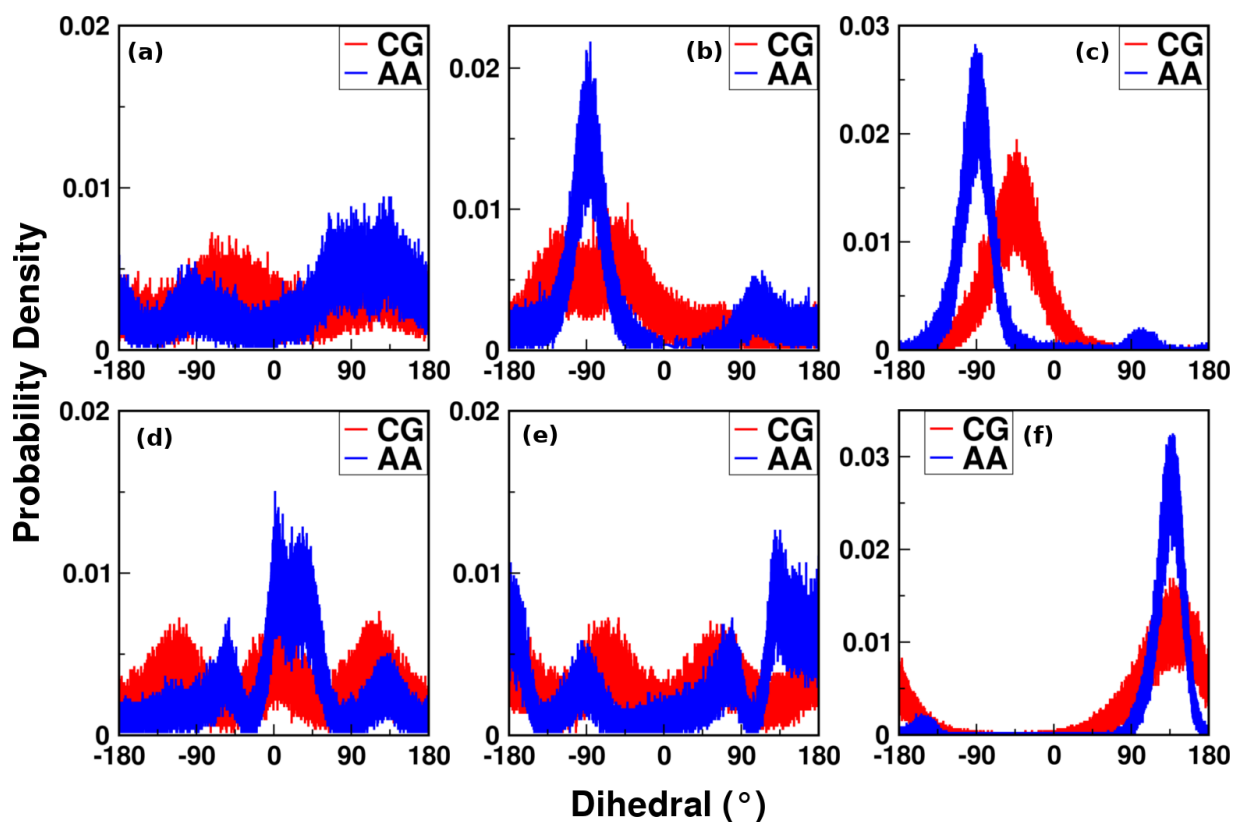
Figure C13.5: Distributions of dihedrals (a) NC2-C3M-NCT-CO, (b) C3M-NCT-CO-NC, (c) NCT-CO-NC-C3M, (d) CO-NC-C3M-NC2, (e) NC2-C3M-NC-COO1, and (f) NCT-CO-NC-COO1
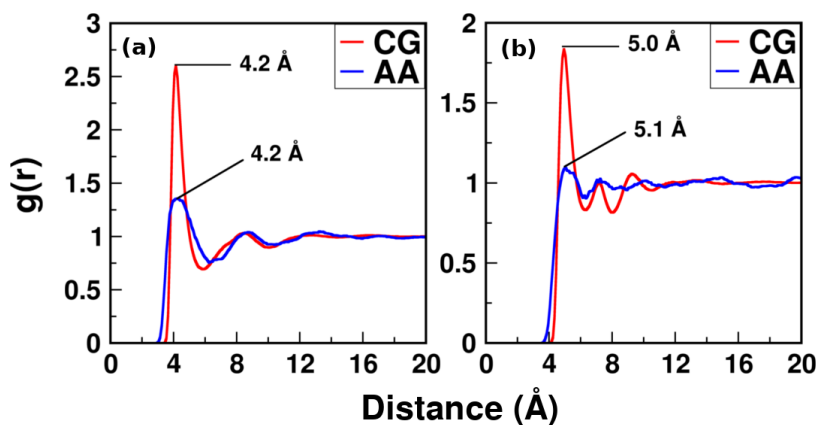
Figure C13.6: RDFs of (a) NC2 and (b) C3M

Analogue: N-butylamine
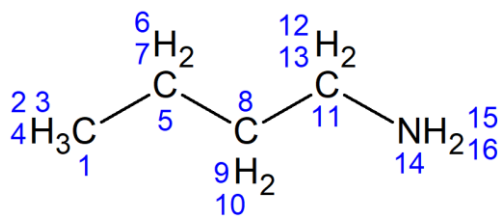
Figure C13.7: All-atom N-butylamine with atom numbering

139

Figure C13.8: CG model of N-butylamine showing bead types
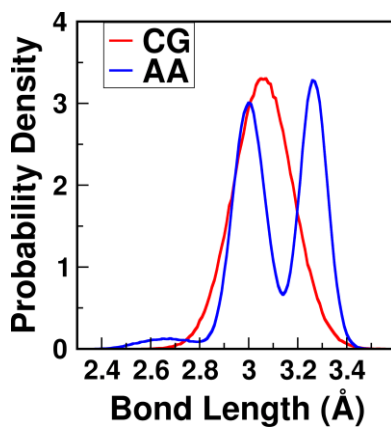
Figure C13.9: Distribution of bond NC2-C3M
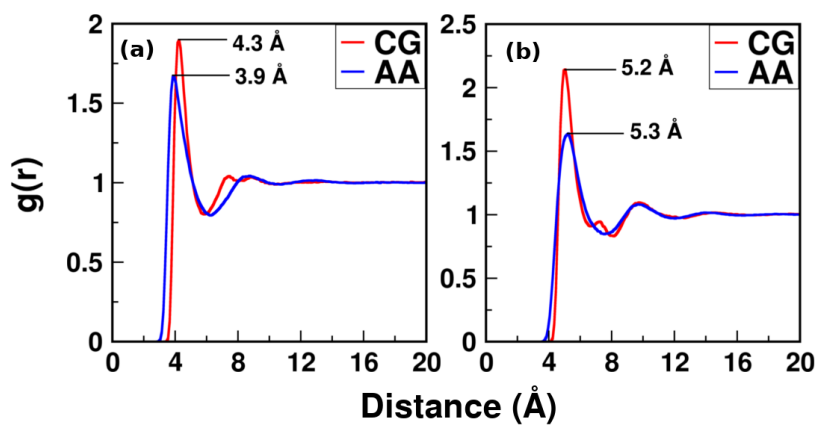
Figure C13.10: RDFs of (a) NC2 and (b) C3M

C14: Methionine

Met-Met Dipeptide
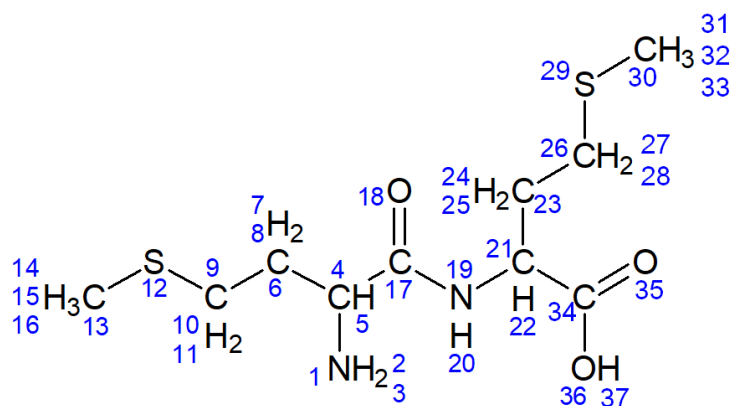
Figure C14.1: All-atom Methionine dipeptide with atom numbering
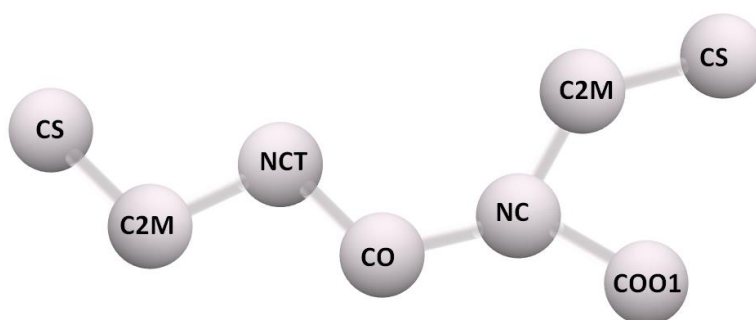
Figure C14.2: CG model of Methionine dipeptide showing bead types
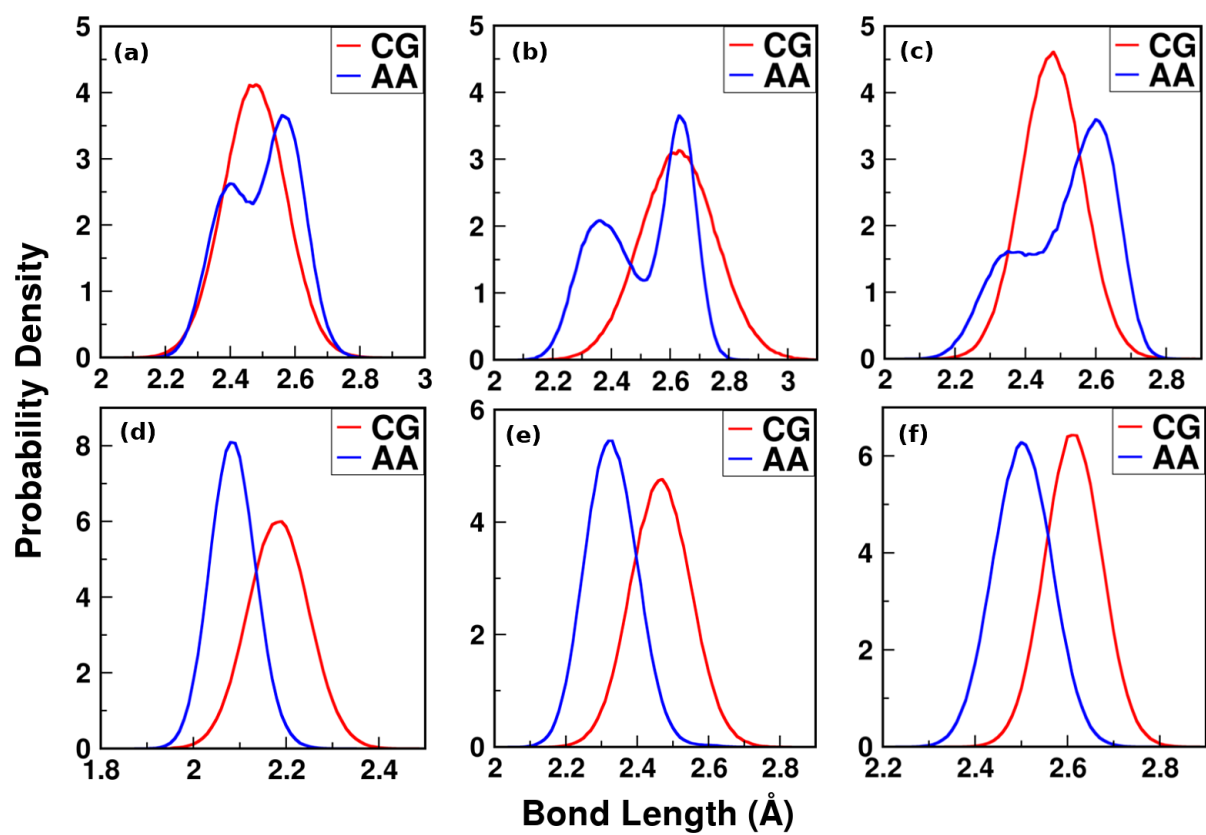
Figure C14.3: Distributions of bonds (a) CS-C2M, (b) C2M-NCT, (c) NCT-CO, (d) CO-NC, (e) NC-C2M, and (f) NC-COO1

Figure C14.4: Distributions of angles (a) CS-C2M-NCT, (b) C2M-NCT-CO, (c) NCT-CO-NC, (d) CO-NC-C2M, (e) NC-C2M-CS, (f) CO-NC-COO1, and (g) C2M-NC-COO1
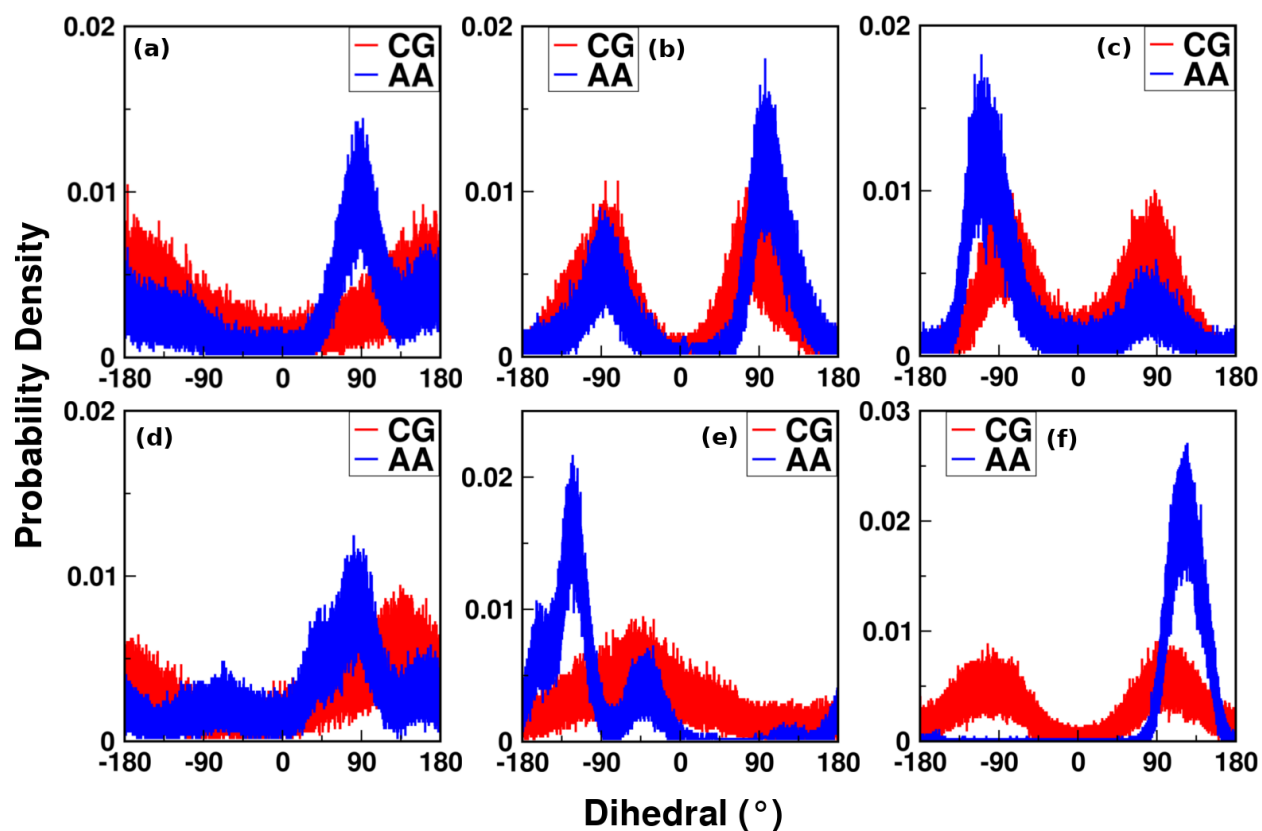
Figure C14.5: Distributions of dihedrals (a) CS-C2M-NCT-CO, (b) C2M-NCT-CO-NC, (c) NCT-CO-NC-C2M, (d) CO-NC-C2M-CS, (e) CS-C2M-NC-COO1, and (f) NCT-CO-NC-COO1
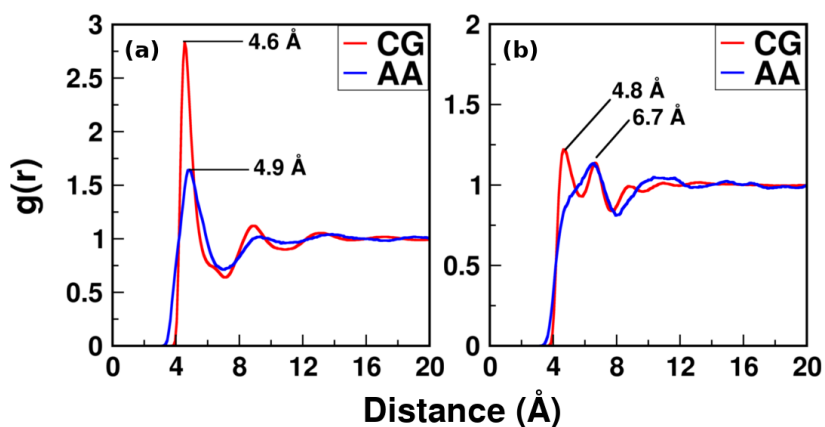
Figure C14.6: RDFs of (a) CS and (b) C2M

Analogue: Ethyl methyl sulfide
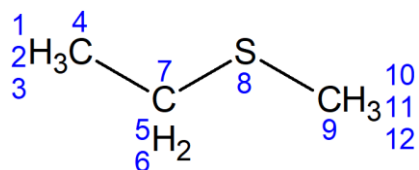
Figure C14.7: All-atom ethyl methyl sulfide with atom numbering

Figure C14.8: CG model of ethyl methyl sulfide showing bead types
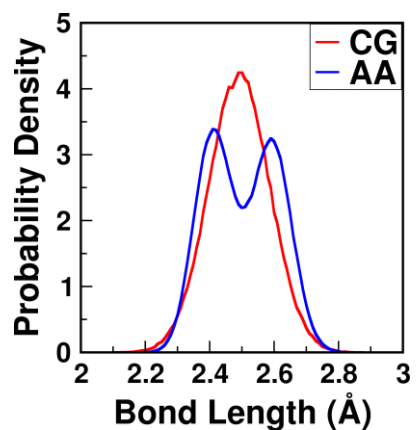
Figure C14.9: Distribution of bond CS-C2M
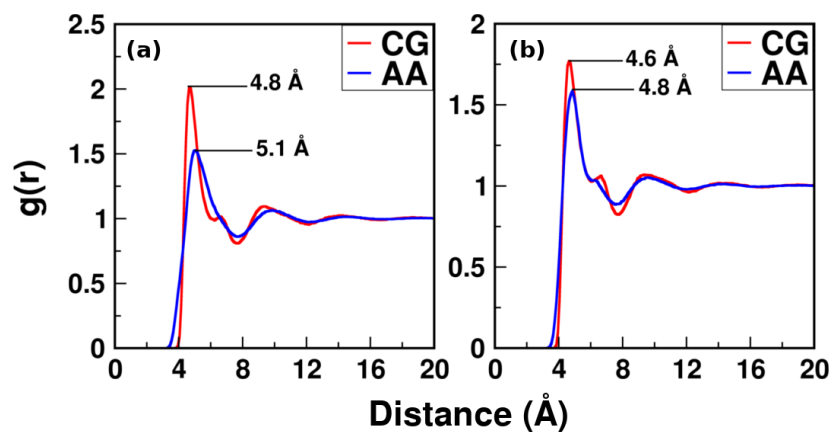
Figure C14.10: RDFs of (a) CS and (b) C2M

C15: Phenylalanine

Phe-Phe Dipeptide
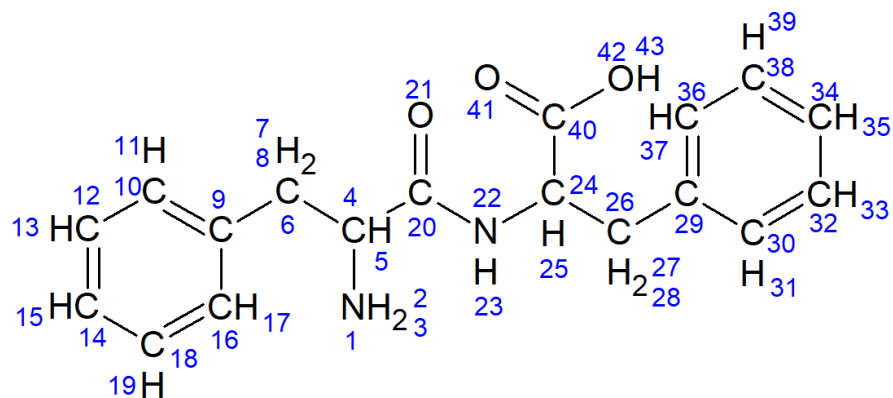
Figure C15.1: All-atom Phenylalanine dipeptide with atom numbering
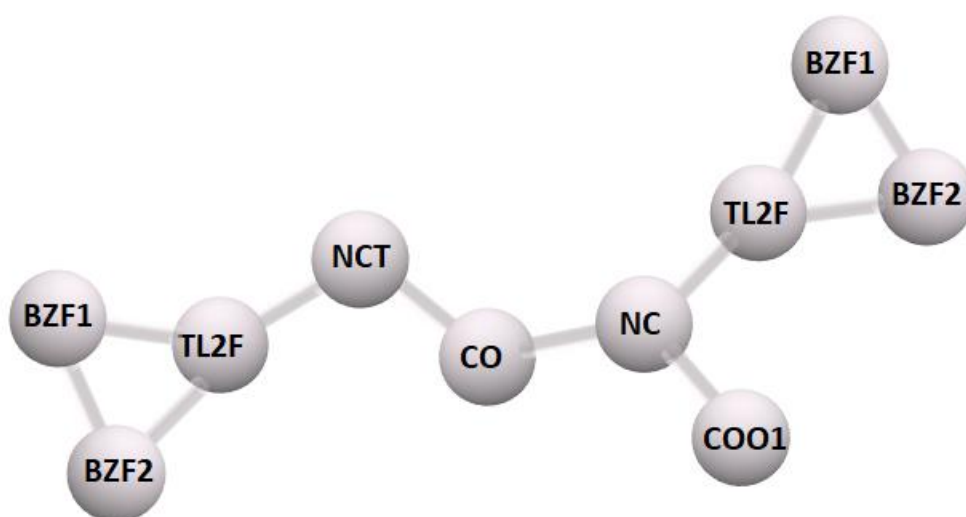
Figure C15.2: CG model of Phenylalanine dipeptide showing bead types
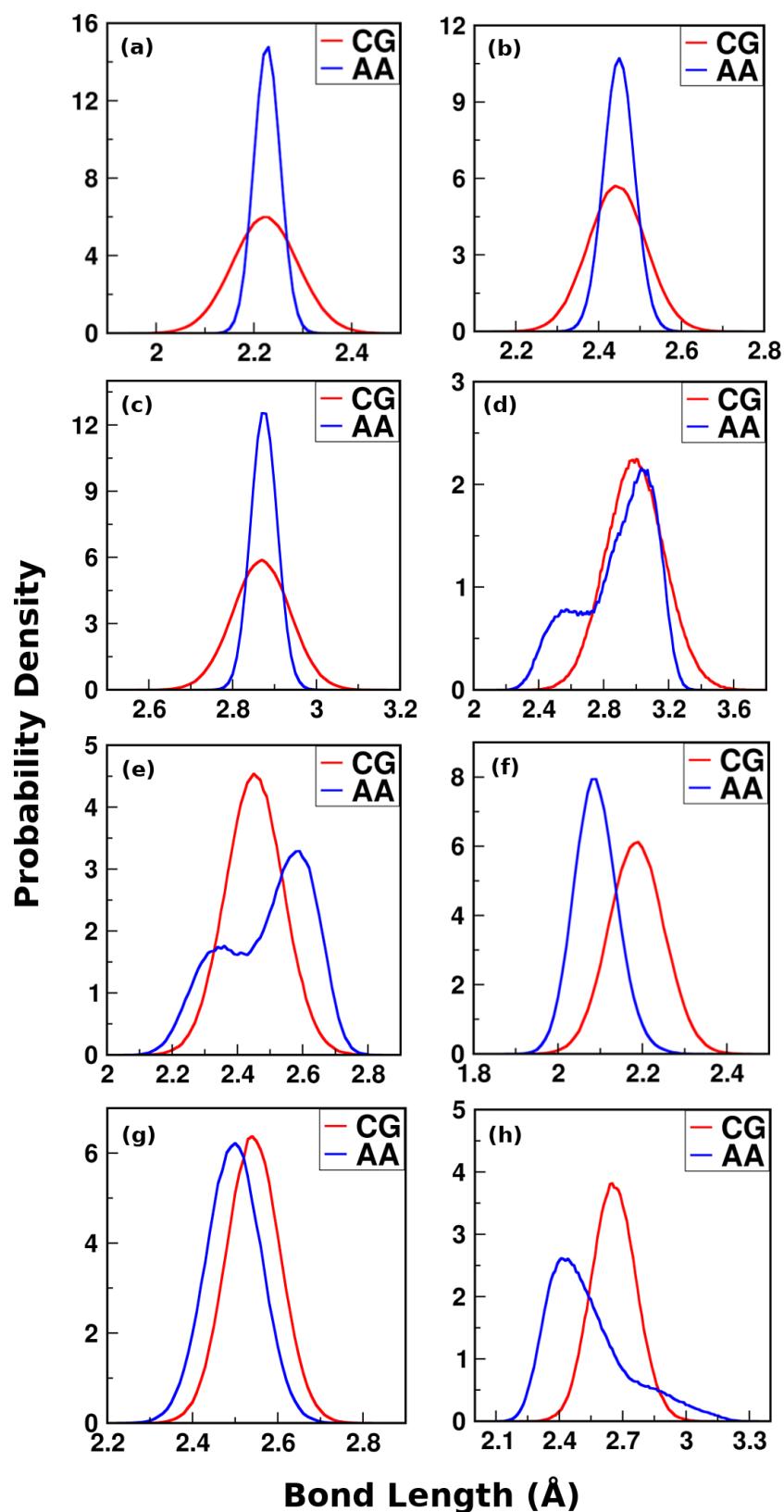
Figure C15.3: Distributions of bonds (a) BZF1-BZF2, (b) BZF1-TL2F, (c) BZF2-BZF1, (d) TL2F-NCT, (e) NCT-CO, (f) CO-NC, (g) NC-COO1, and (h) NC-TL2F
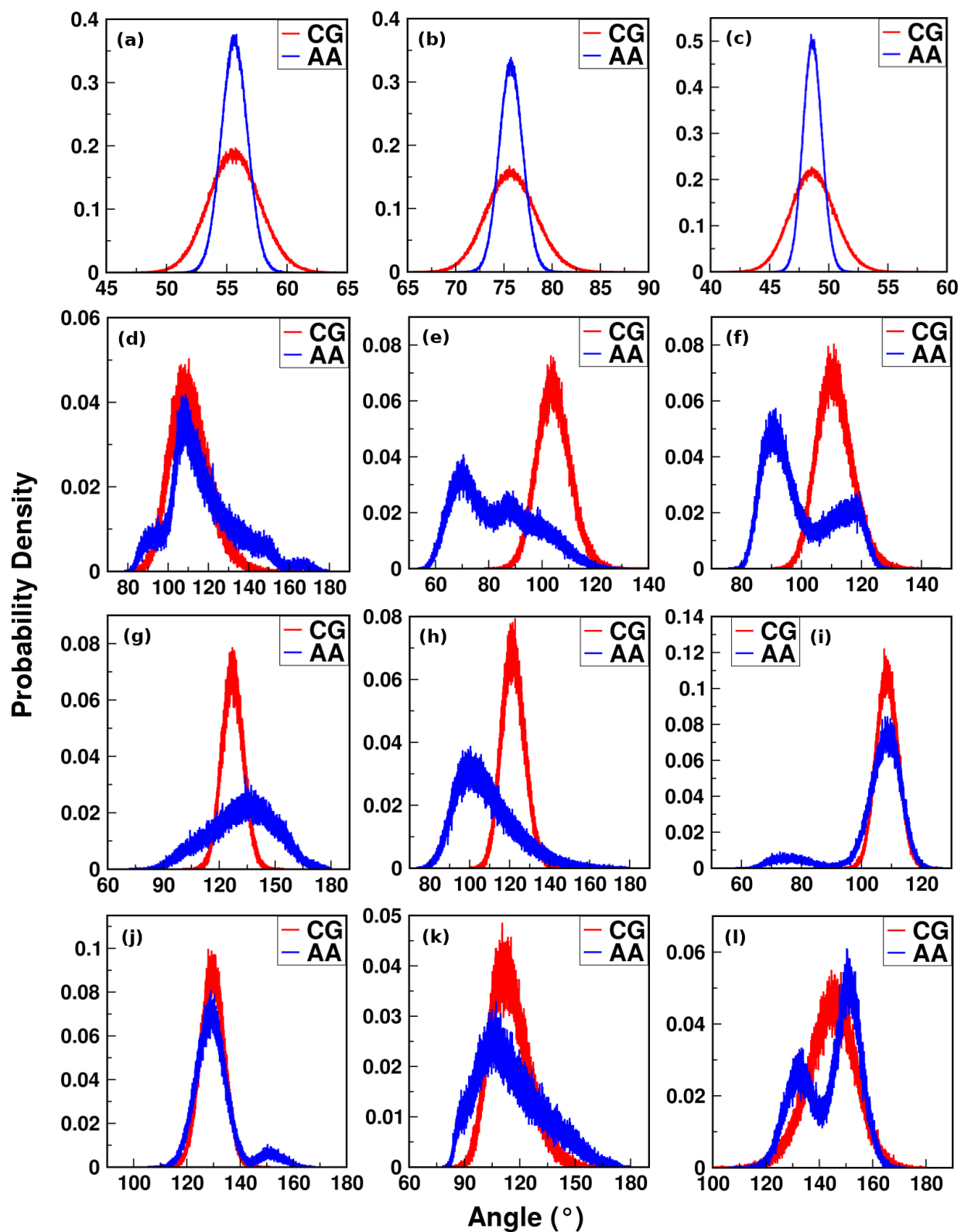
Figure C15.4: Distributions of angles (a) BZF2-BZF1-TL2F, (b) BZF1-BZF2-TL2F (c) BZF1-TL2F-BZF2, (d) BZF2-TL2F-NCT, (e) TL2F-NCT-CO, (f) NCT-CO-NC, (g) CO-NC-TL2F, (h) CO-NC-COO1, (i) COO1-NC-TL2F, (j) NC-TL2F-BZF1, (k) NC-TL2F-BZF2, and (l) BZF1-TL2F-NCT
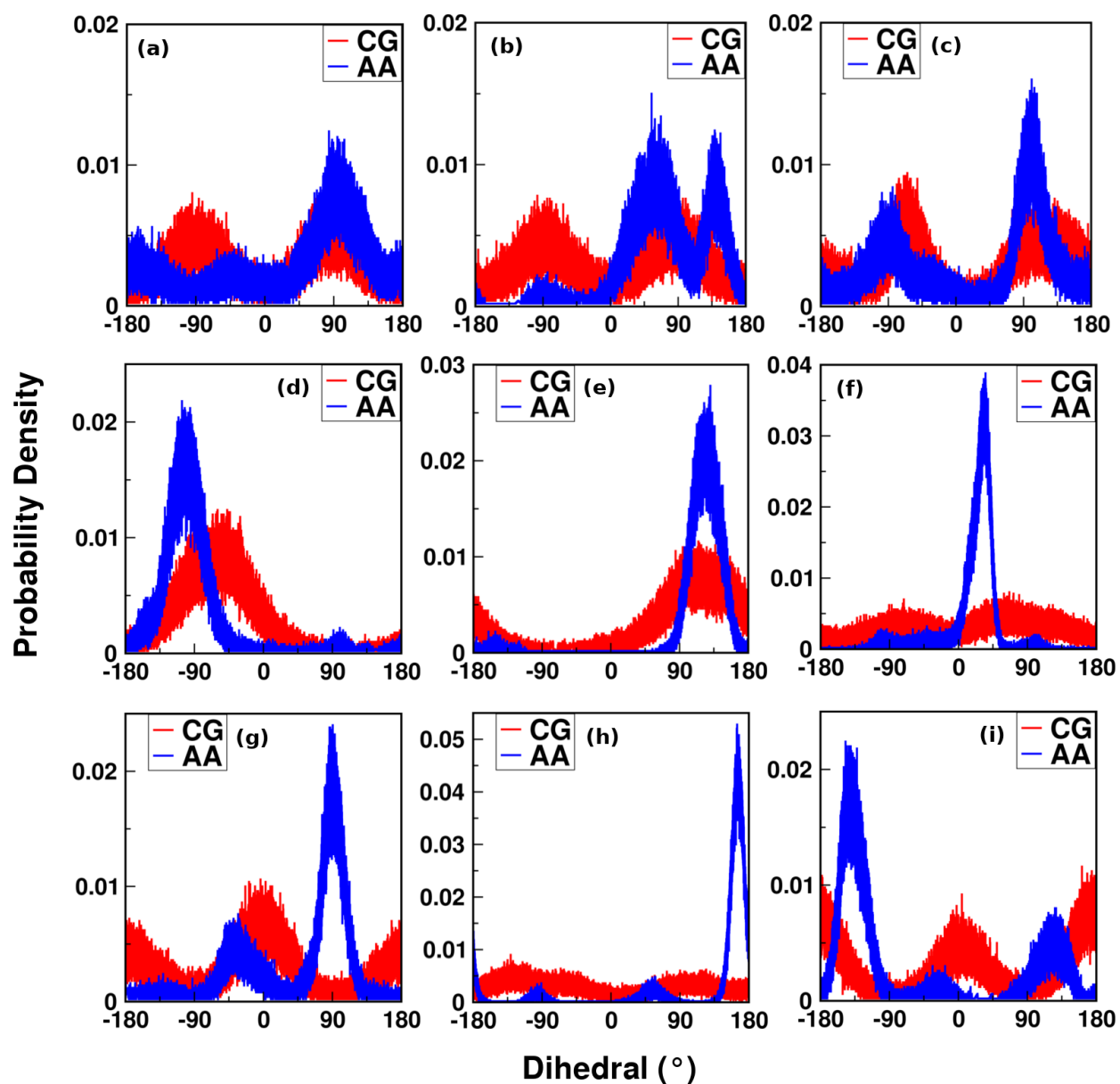
Figure C15.5: Distributions of dihedrals (a) BZF1-TL2F-NCT-CO, (b) BZF2-TL2F-NCT-CO, (c) TL2F-NCT-CO-NC, (d) NCT-CO-NC-TL2F, (e) NCT-CO-NC-COO1, (f) CO-NC-TL2F-BZF1, (g) CO-NC-TL2F-BZF2, (h) BZF1-TL2F-NC-COO1, and (i) BZF2-TL2F-NC-COO1
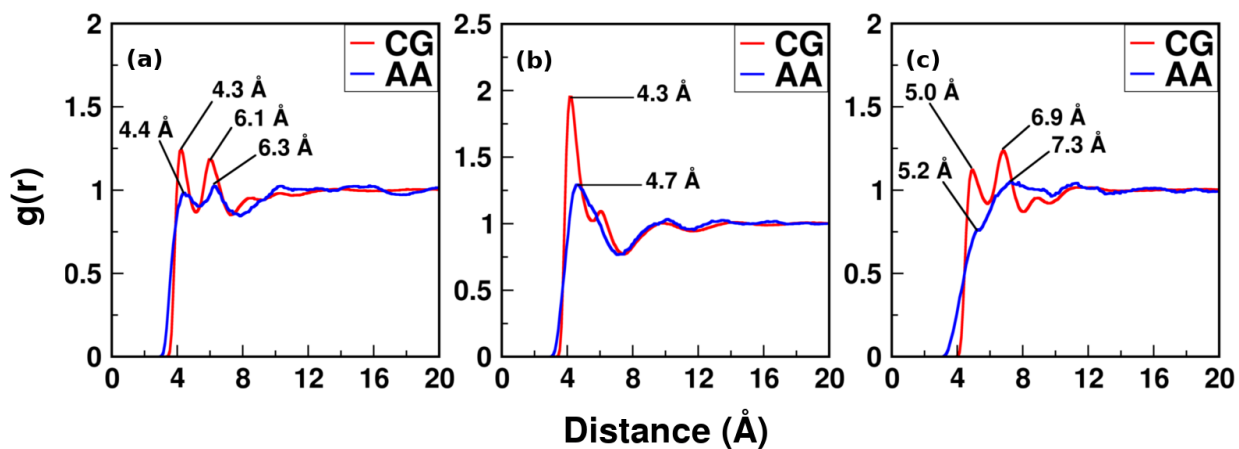
Figure C15.6: RDFs of (a) BZF2, (b) BZF1, and (c) TL2F

Analogue: Toluene

Figure C15.7: All-atom toluene with atom numbering
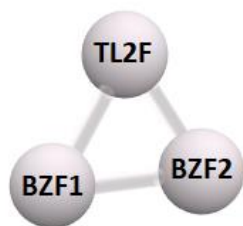
Figure C15.8: CG model of toluene showing bead types
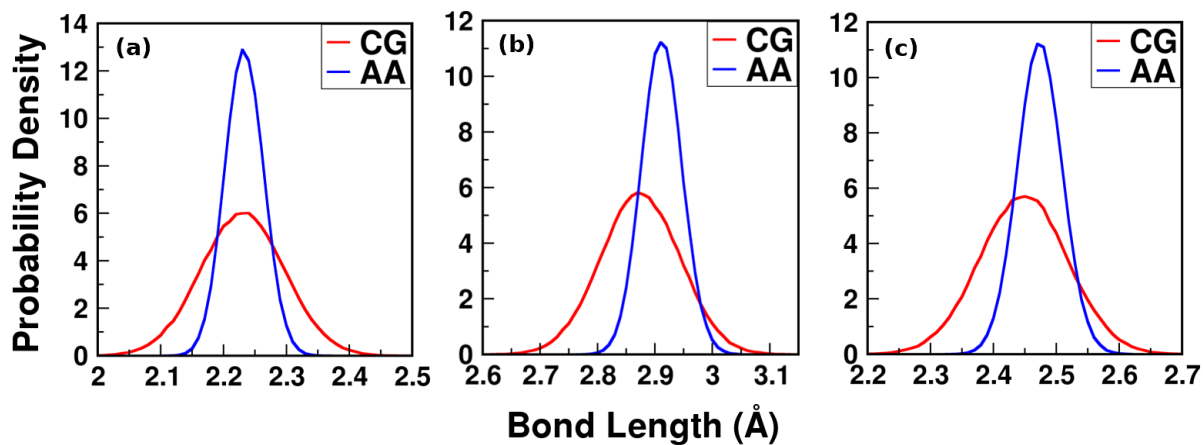
Figure C15.9: Distribution of bonds (a) BZF1-BZF2, (b) BZF1-TL2F, and (c) BZF2-TL2F

150
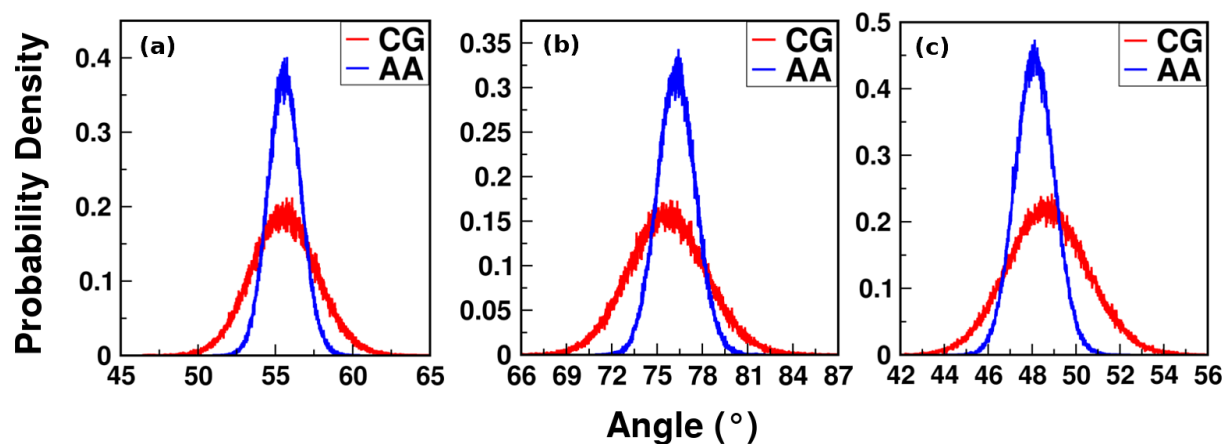
Figure C15.10: Distributions of angles (a) BZF2-BZF1-TL2F, (b) BZF1-BZF2-TL2F, and (c) BZF1-TL2F-BZF2
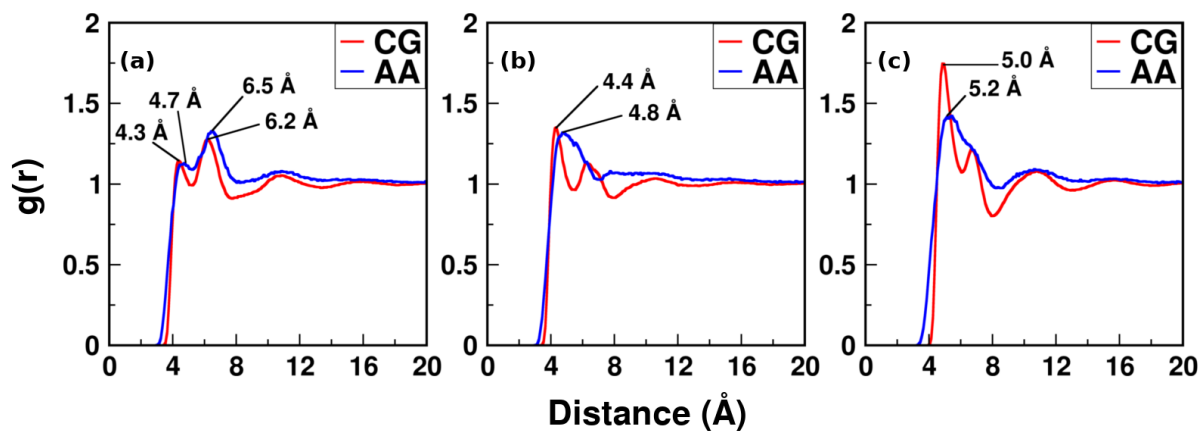
Figure C15.11: RDFs of (a) BZF2, (b) BZF1, and (c) TL2F

C16: Proline

Val-Pro-Leu Tripeptide

Figure C16.1: All-atom Val-Pro-Leu tripeptide with atom numbering

Figure C16.2: CG model of Val-Pro-Leu tripeptide showing bead types
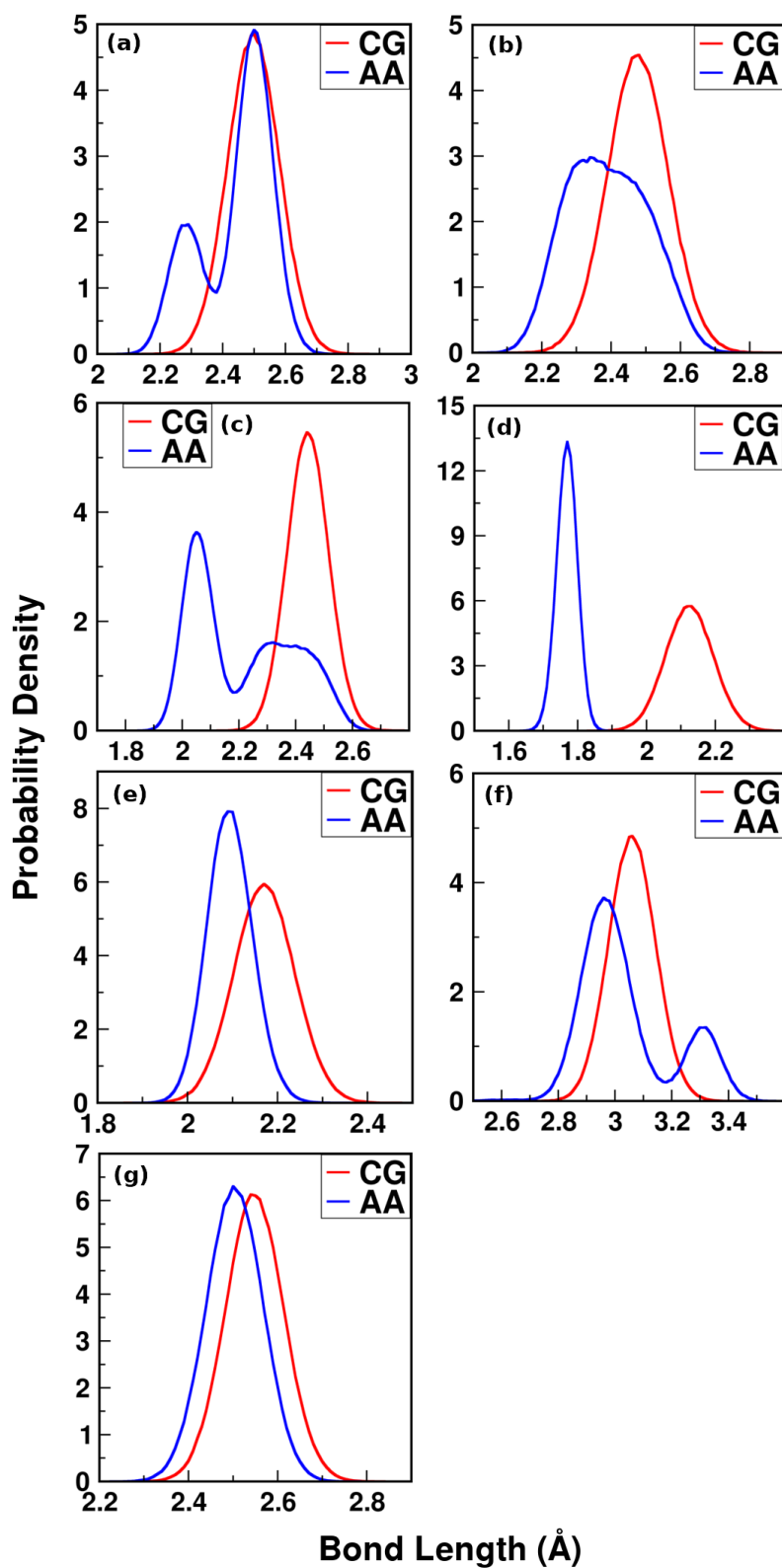
Figure C16.3: Distributions of bonds (a) C3E2-NCT, (b) NCT-CO, (c) CO-PNC, (d) PNC-PS, (e) CO-NC, (f) NC-C41, and (g) NC-COO1
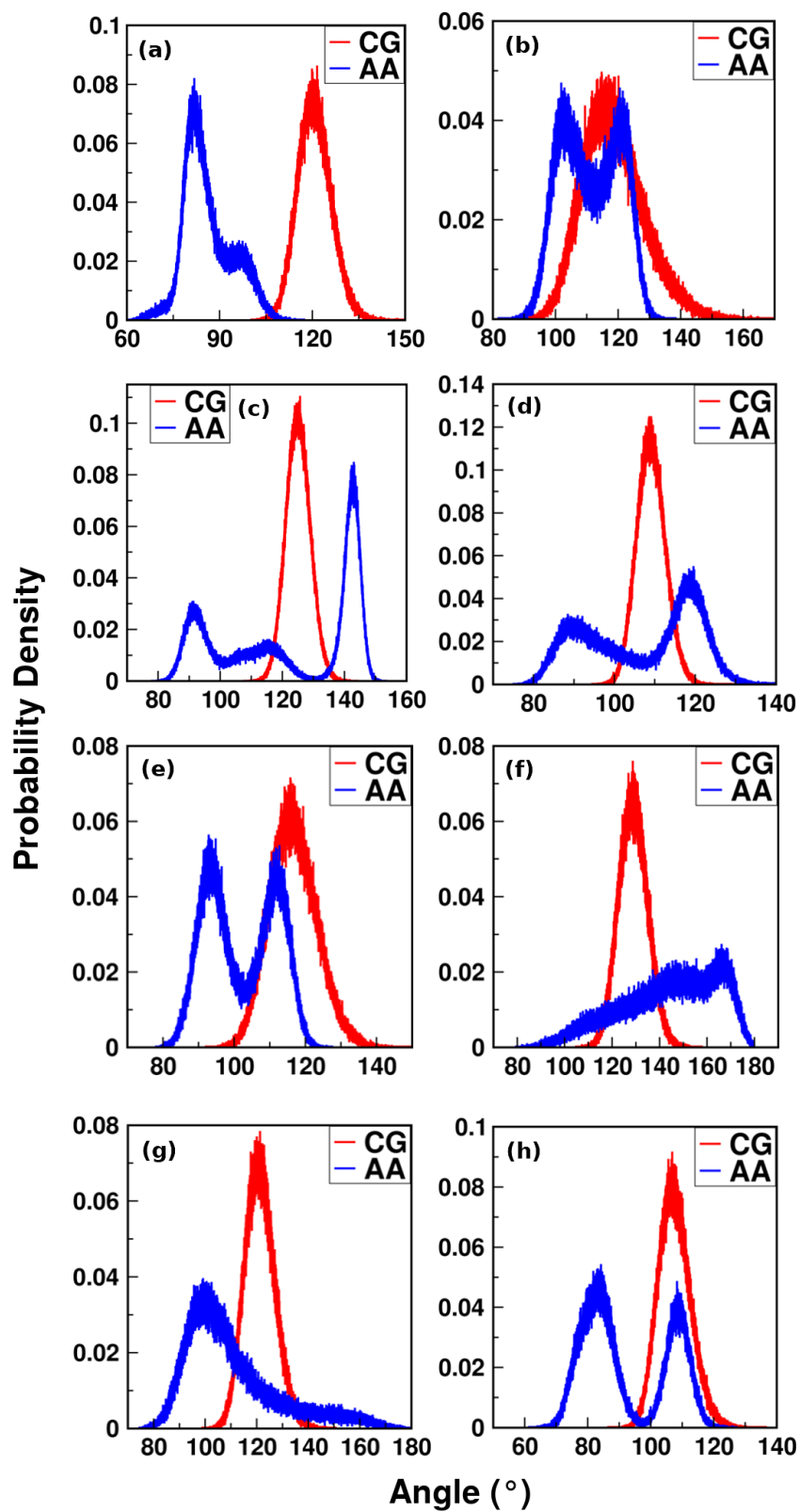
Figure C16.4: Distributions of angles (a) C3E2-NCT-CO, (b) NCT-CO-PNC, (c) CO-PNC-PS, (d) CO-PNC-CO, (e) PNC-CO-NC, (f) CO-NC-C41, (g) CO-NC-COO1, and (h) C41-NC-COO1
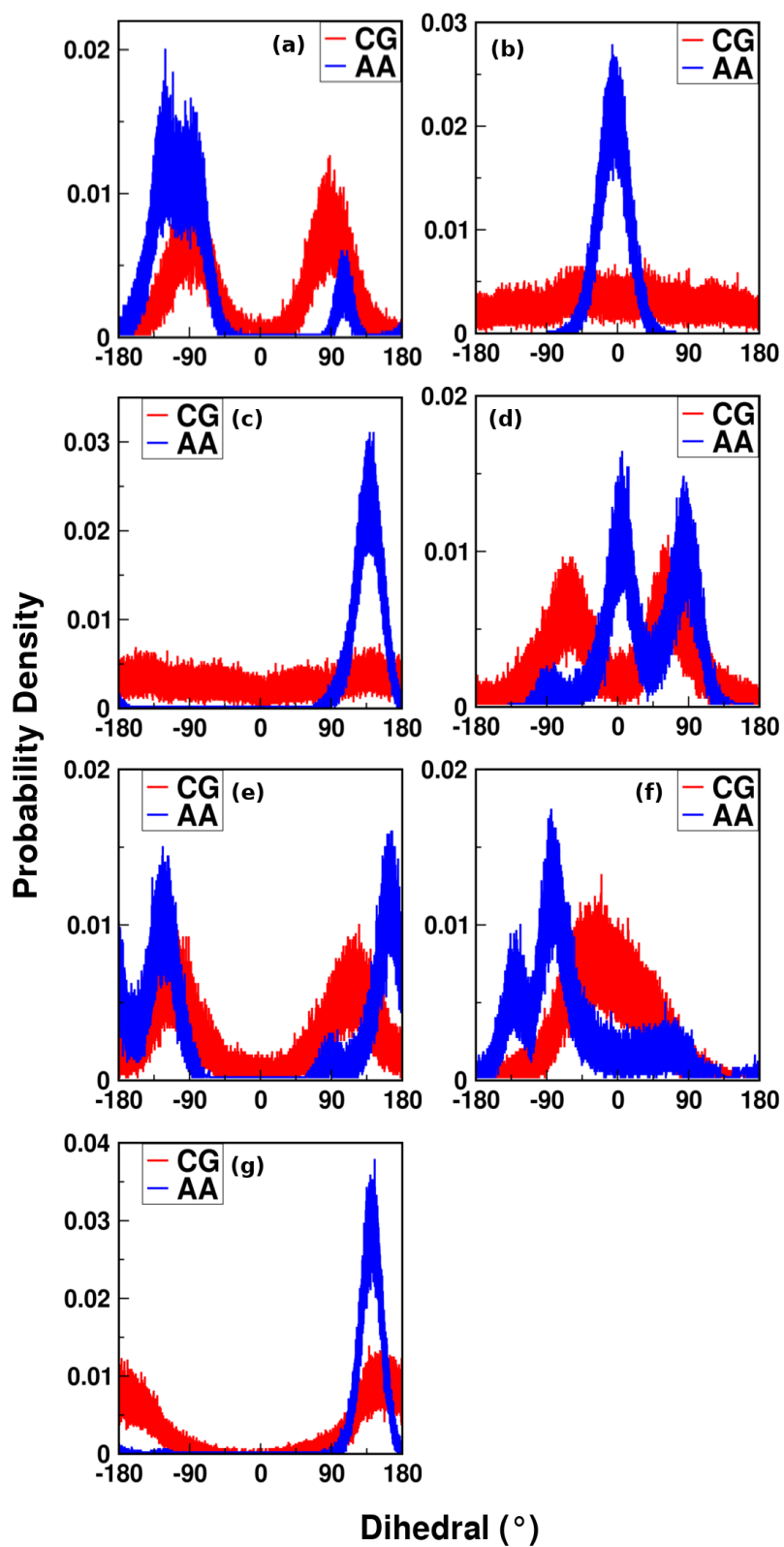
Figure C16.5: Distributions of dihedrals (a) C3E2-NCT-CO-PNC, (b) NCT-CO-PNC-PS (c) NCT-CO-PNC-CO, (d) CO-PNC-CO-NC, (e) PS-PNC-CO-NC, (f) PNC-CO-NC-C41, and (g) PNC-CO-NC-COO1
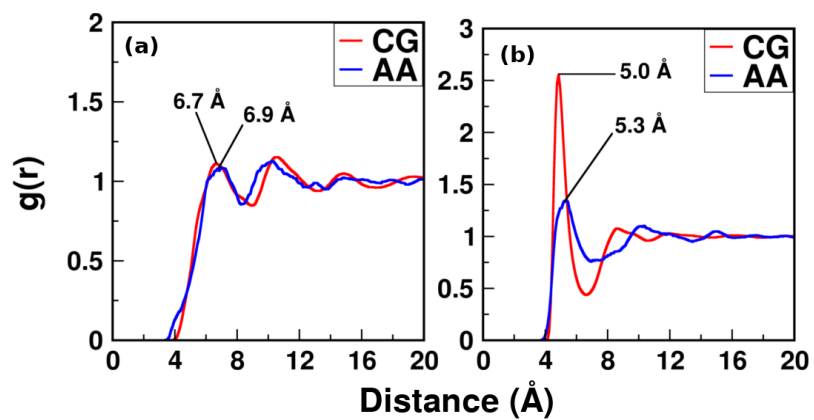
Figure C16.6: RDFs of (a) PNC and (b) PS

C17: Serine

Ser-Ser Dipeptide
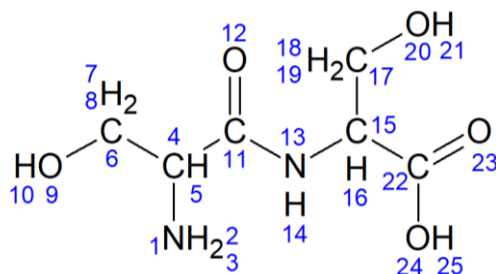
Figure C17.1: All-atom Serine dipeptide with atom numbering
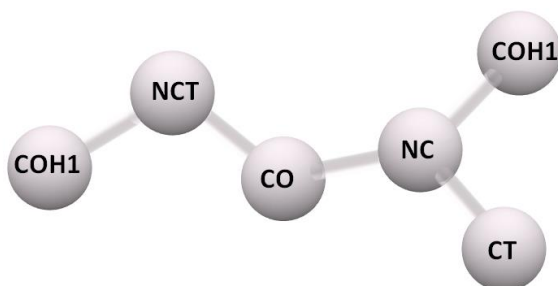
**Figure C17.2:** CG model of Serine dipeptide showing bead types
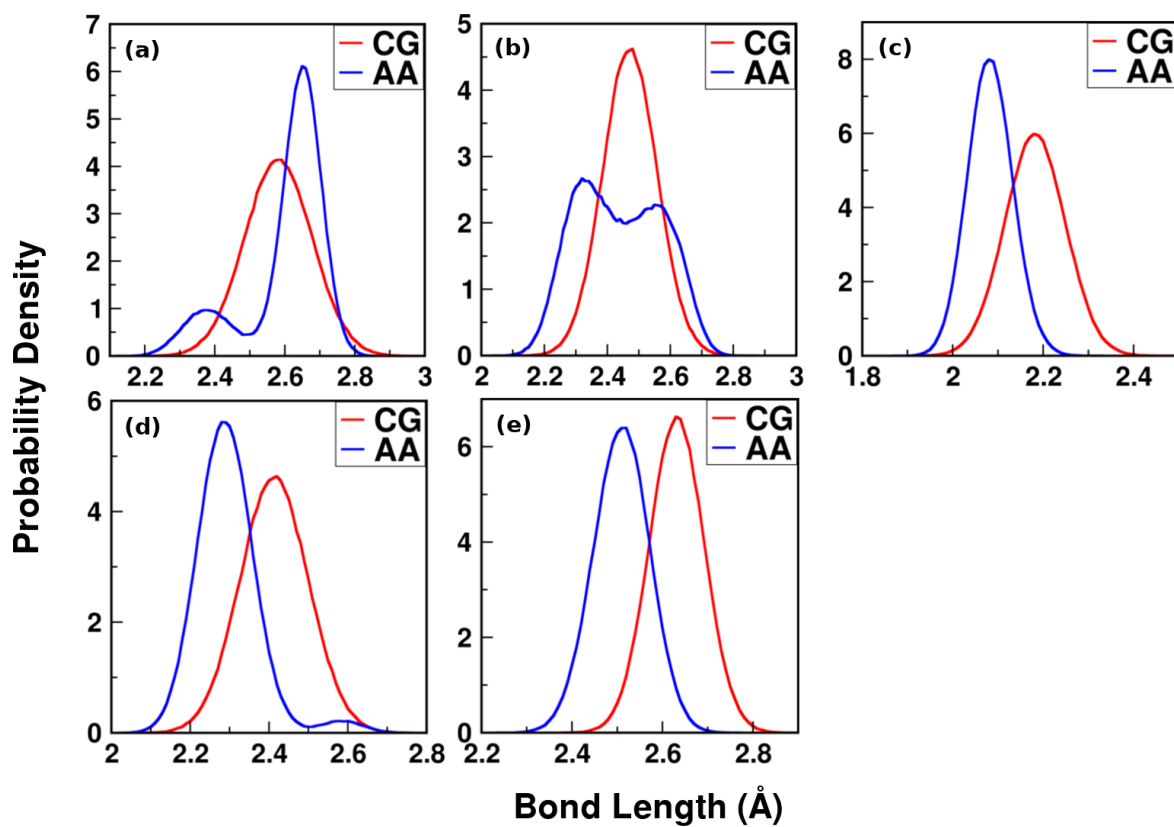
Figure C17.3: Distributions of bonds (a) COH1-NCT, (b) NCT-CO, (c) CO-NC, (d) NC-COH1, and (e) NC-COO1
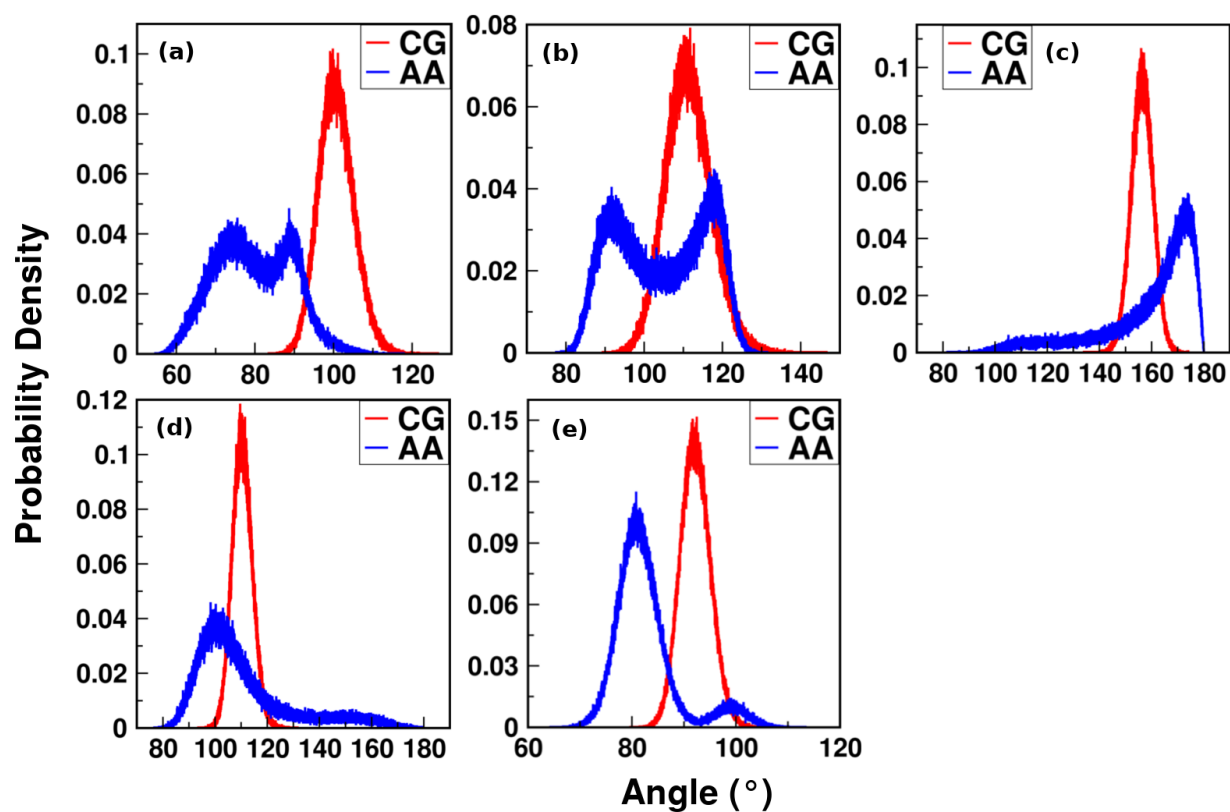
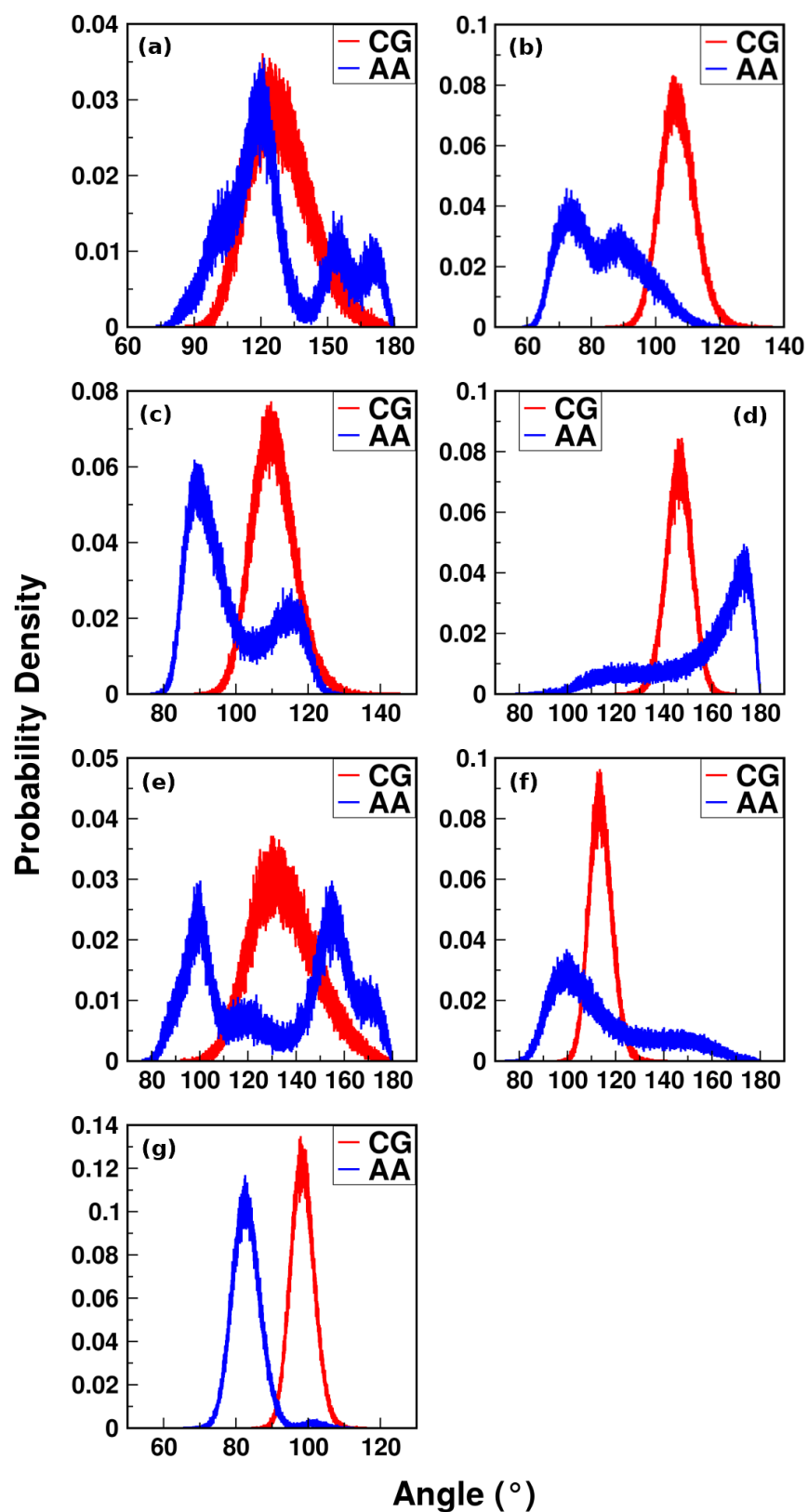**Figure C17.4:** Distributions of angles (a) COH1-NCT-CO, (b) NCT-CO-NC, (c) CO-NC-COH1, (d) CO-NC-COO1, and (e) COH1-NC-COO1
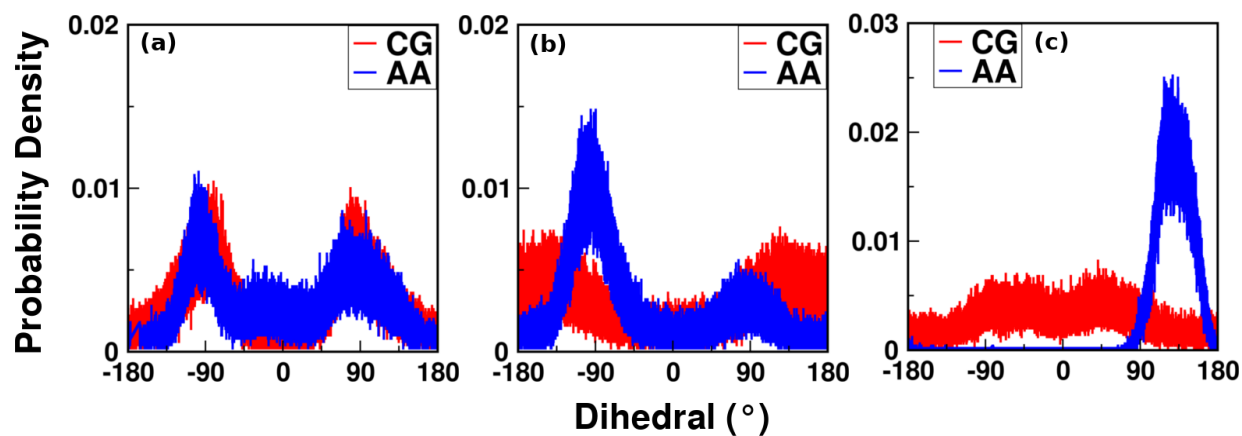
Figure C17.5: Distributions of dihedrals (a) COH1-NCT-CO-NC, (b) NCT-CO-NC-COH1, and (c) NCT-CO-NC-COO1
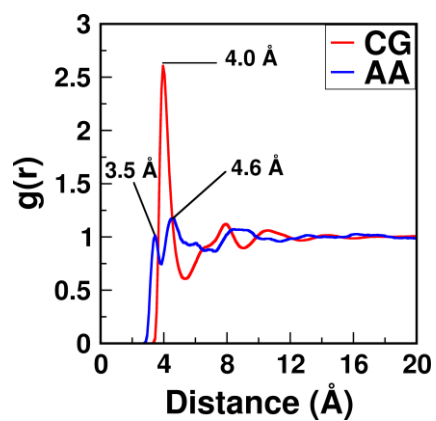
Figure C17.6: RDF of COH1

Analogue: Methanol
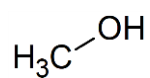
Figure C17.7: All-atom methanol

Figure C17.8: CG model of methanol showing bead type
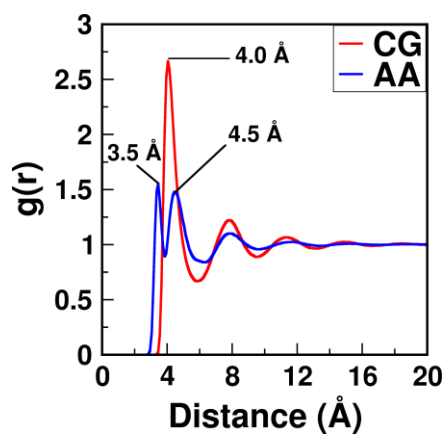
Figure C17.9: RDF of COH1

C18: Threonine

Thr-Thr Dipeptide
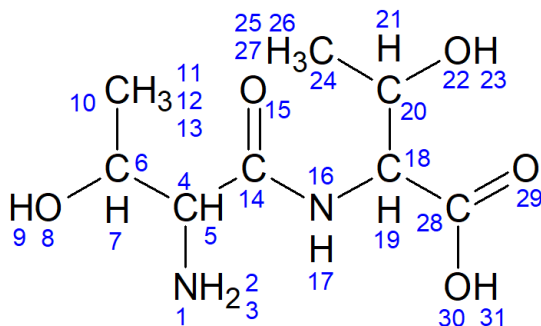
Figure C18.1: All-atom model of Threonine dipeptide with atom numbering
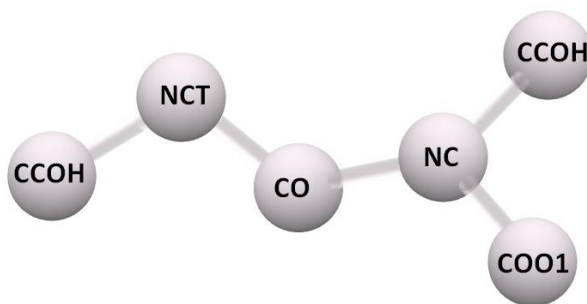
Figure C18.2: CG model of Threonine dipeptide
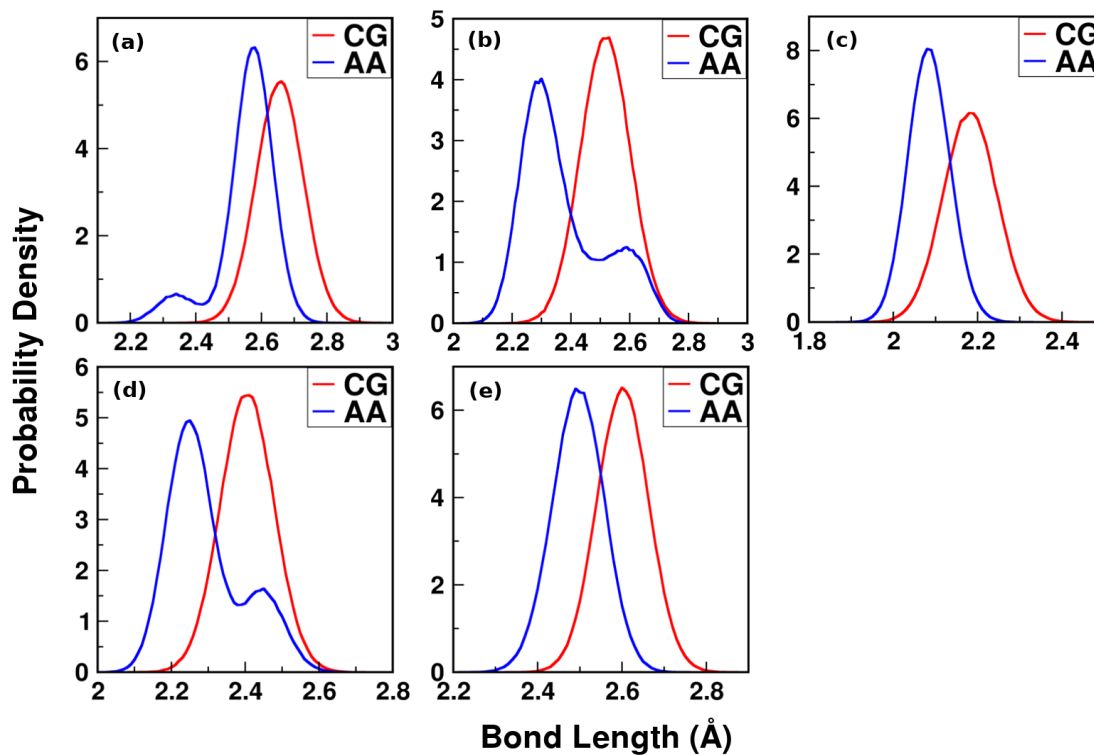
Figure C18.3: Distributions of bonds (a) CCOH-NCT, (b) NCT-CO, (c) CO-NC, (d) NC-CCOH, and (e) NC-COO1
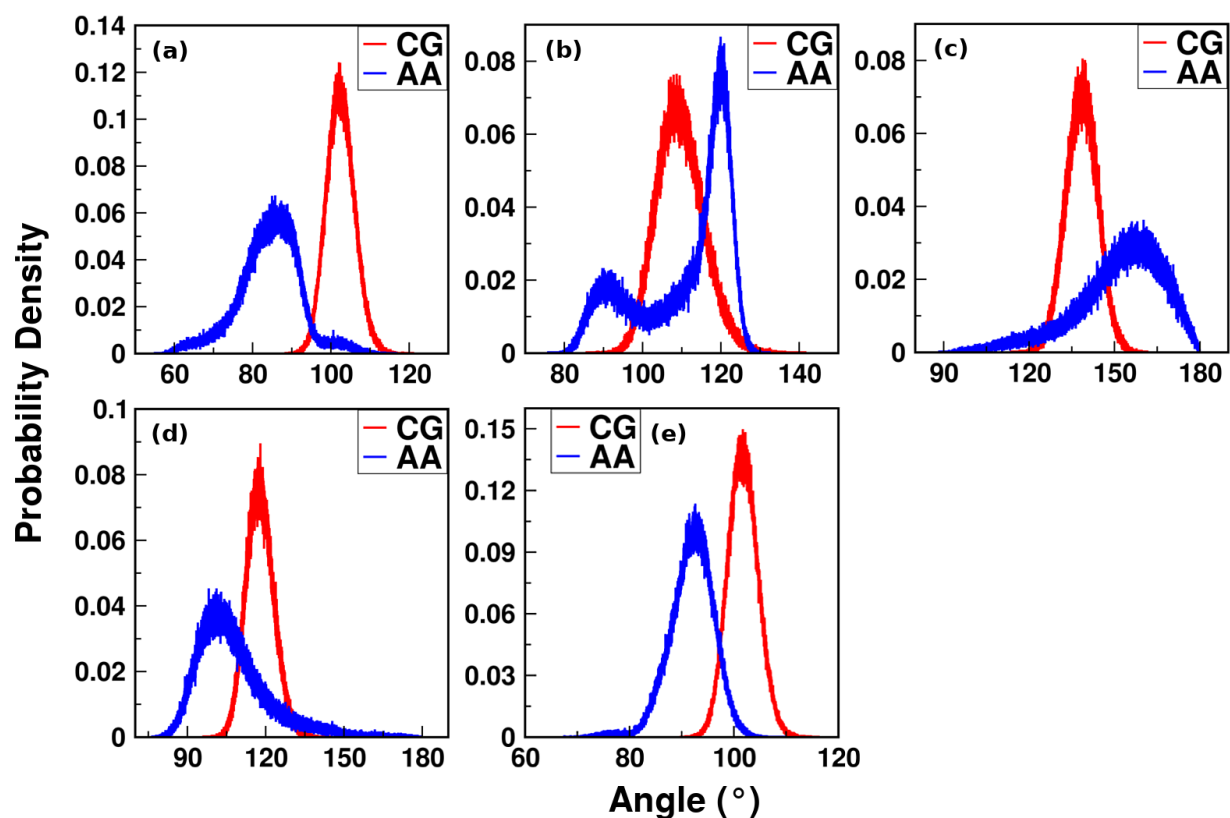
Figure C18.4: Distributions of angles (a) CCOH-NCT-CO, (b) NCT-CO-NC, (c) CO-NC-CCOH, (d) CO-NC-COO1, and (e) CCOH-NC-COO1
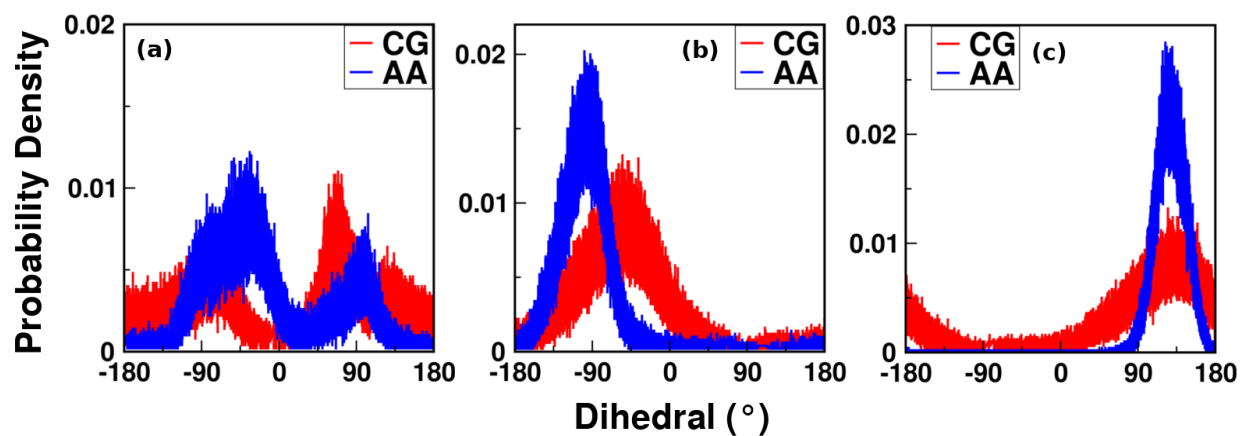
Figure C18.5: Distribution of dihedrals (a) CCOH-NCT-CO-NC, (b) NCT-CO-NC-CCOH, and (c) NCT-CO-NC-COO1
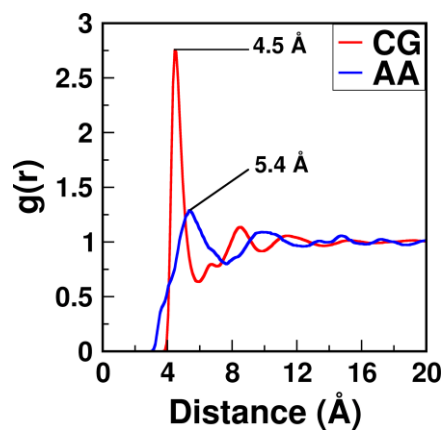
Figure C18.6: RDF of CCOH

Analogue: Ethanol
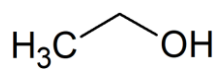
Figure C18.7: All-atom ethanol

Figure C18.8: CG model of ethanol showing bead type
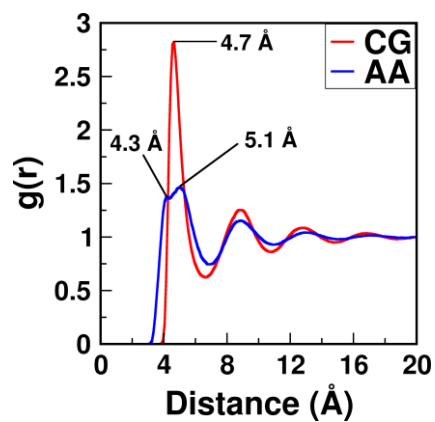
Figure C18.9: RDF of CCOH

C19: Tryptophan

Trp-Trp Dipeptide
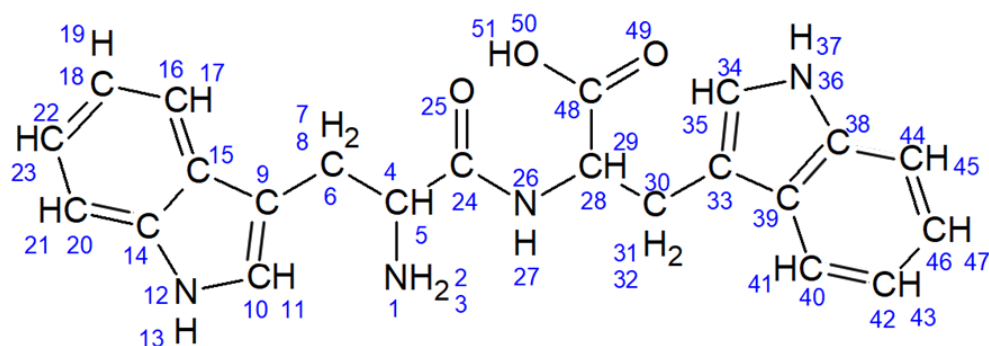
Figure C19.1: All-atom Tryptophan dipeptide with atom numbering
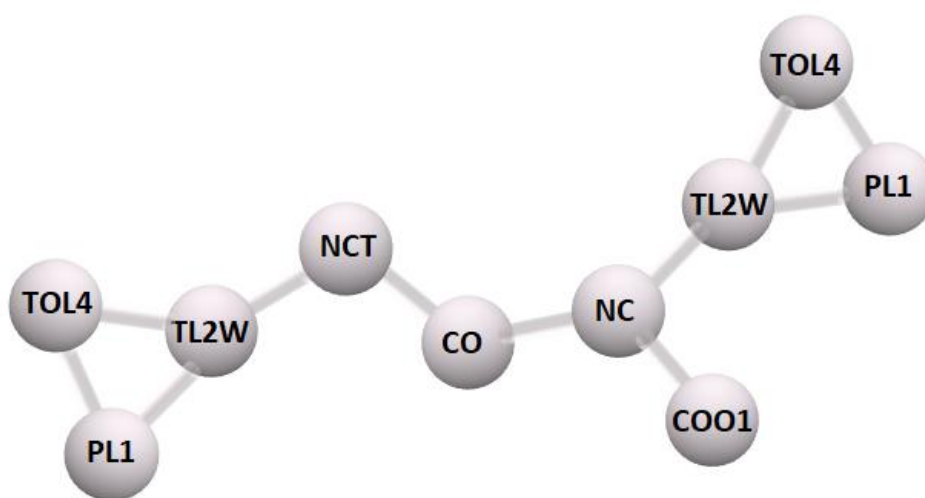
Figure C19.2: CG model of Tryptophan dipeptide showing bead types
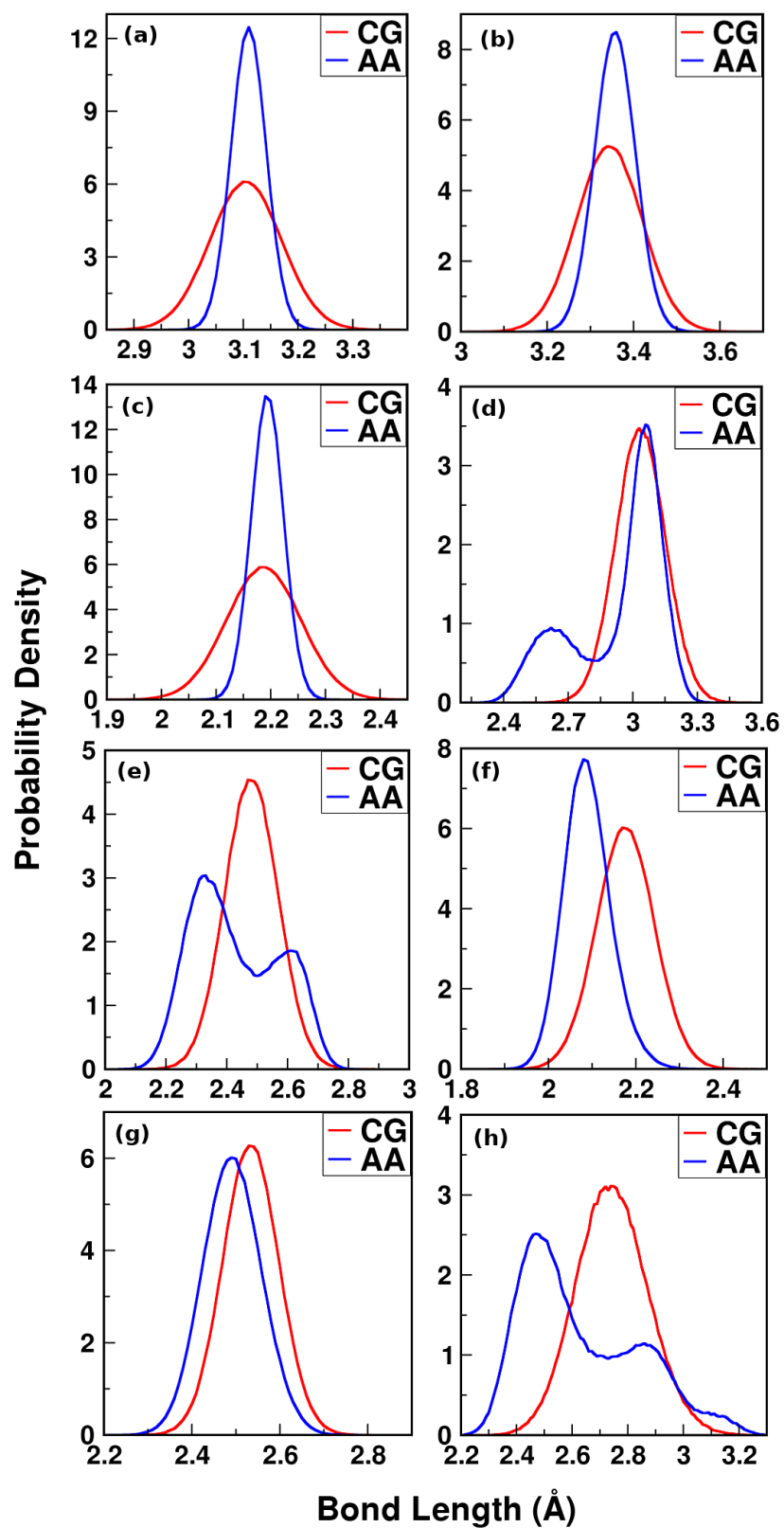
Figure C19.3: Distributions of bonds (a) TOL4-PL1, (b) TOL4-TL2W, (c) TL2W-PL1, (d) TL2W-NCT, (e) NCT-CO, (f) CO-NC, (g) NC-COO1, and (h) NC-TL2W

164
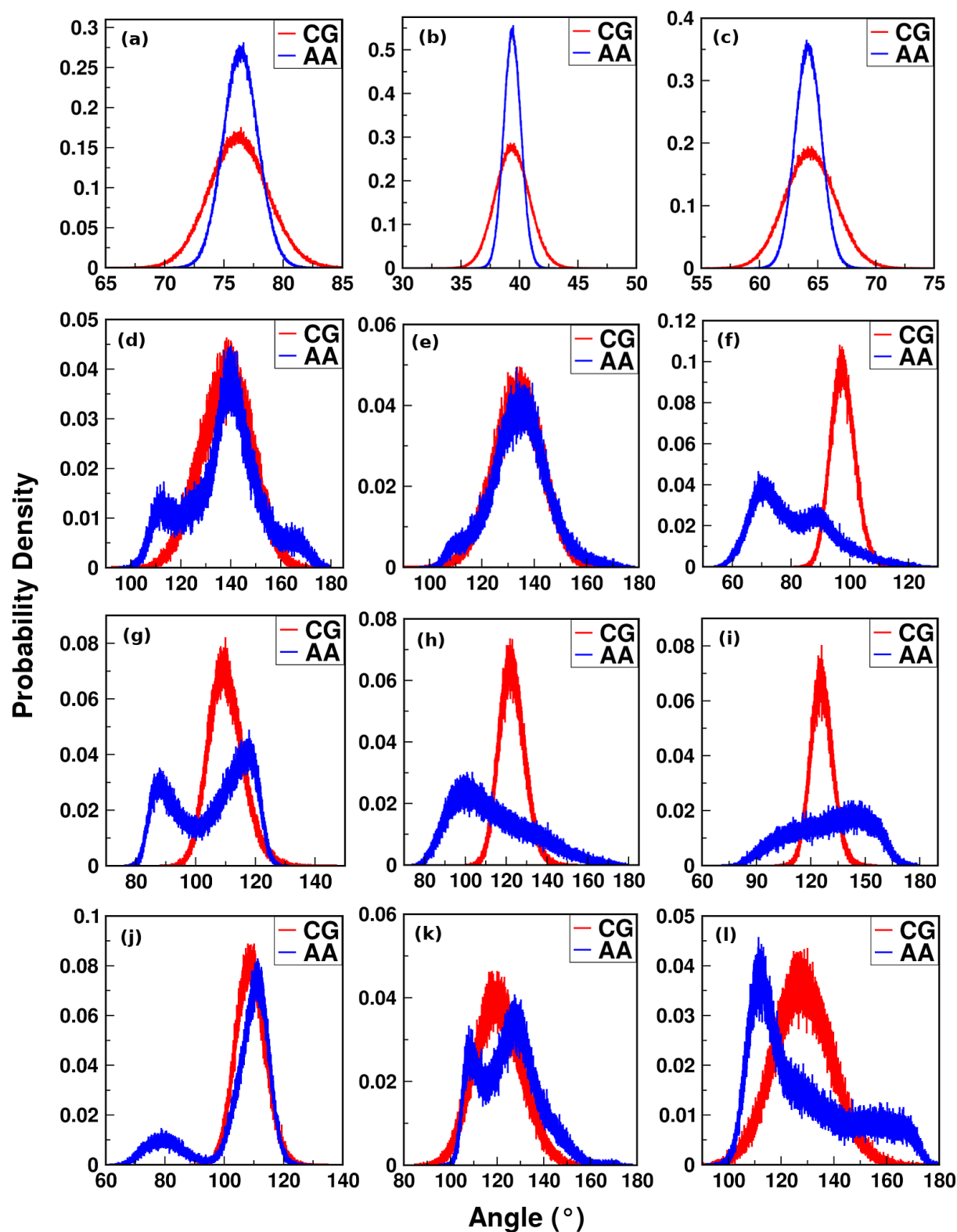
Figure C19.4: Distributions of angles (a) TL2W-PL1-TOL4, (b) PL1-TOL4-TL2W, (c) TOL4-TL2W-PL1, (d) TOL4-TL2W-NCT, (e) PL1-TL2W-NCT, (f) TL2W-NCT-CO, (g) NCT-CO-NC, (h) CO-NC-COO1, (i) CO-NC-TL2W, (j) COO1-NC-TL2W, (k) NC-TL2W-PL1, and (l) NC-TL2W-TOL4
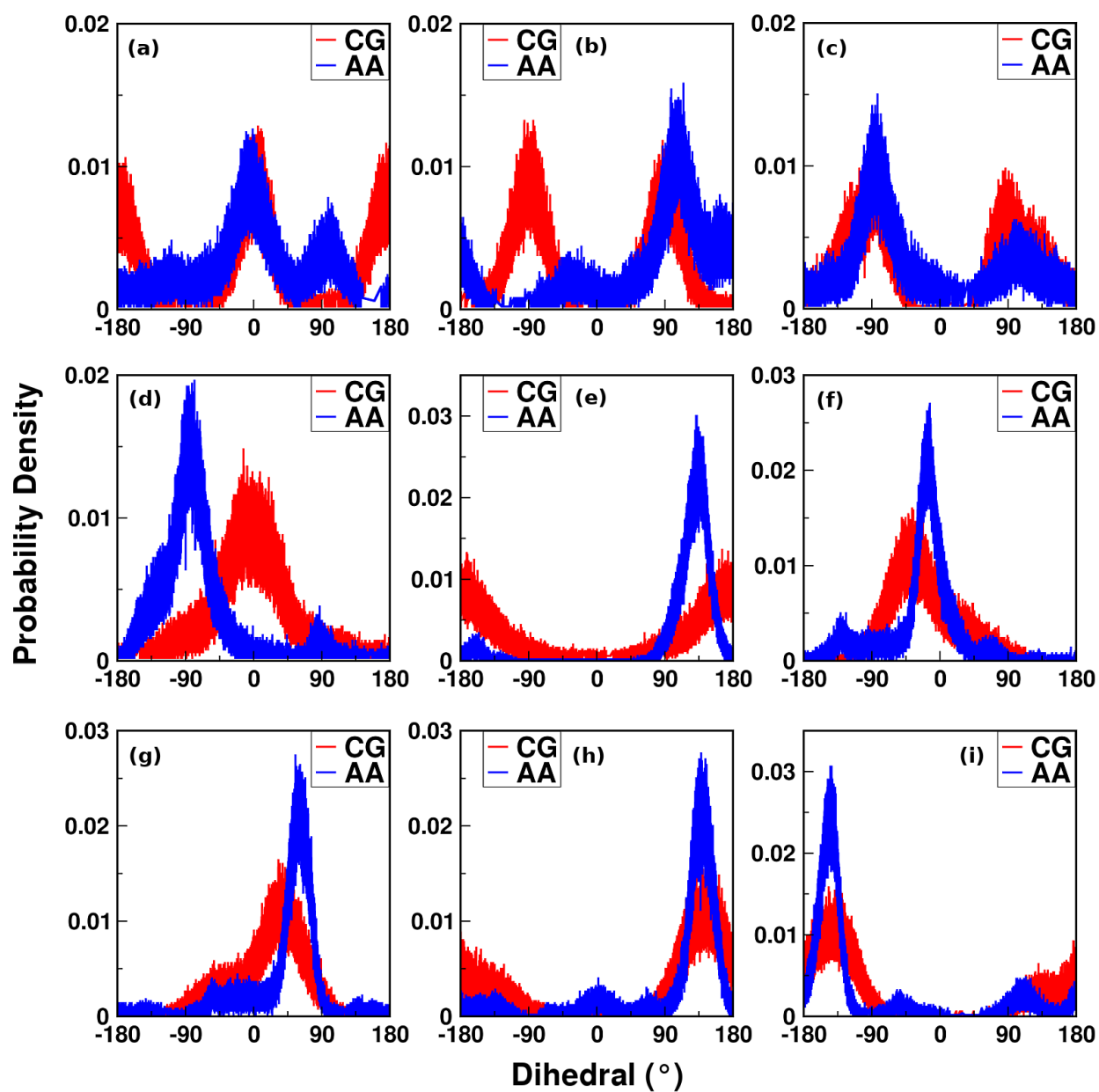
Figure C19.5: Distributions of dihedrals (a) TOL4-TL2W-NCT-CO, (b) PL1-TL2W-NCT-CO, (c) TL2W-NCT-CO-NC, (d) NCT-CO-NC-TL2W, (e) NCT-CO-NC-COO1, (f) CO-NC-TL2W-PL1, (g) CO-NC-TL2W-TOL4, (h) COO1-NC-TL2W-PL1, and (i) COO1-NC-TL2W-TOL4
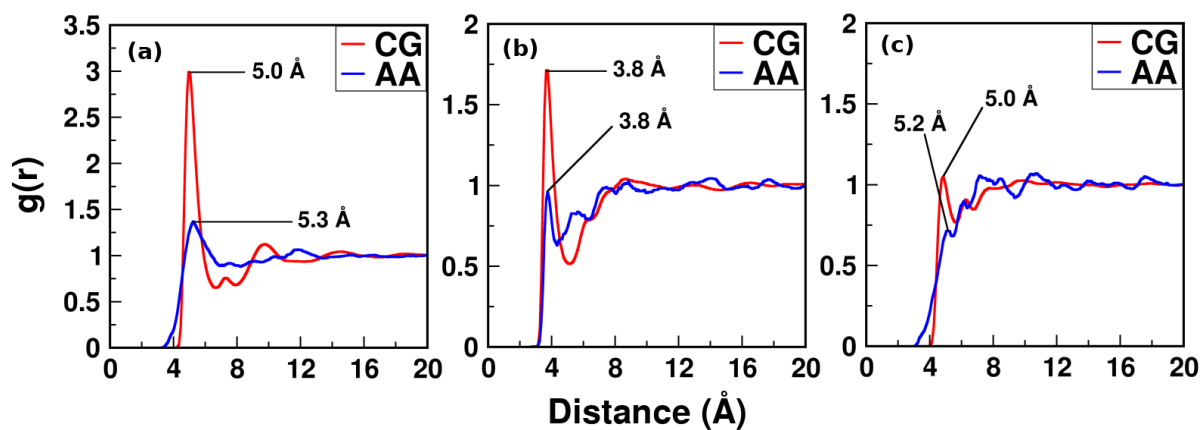
Figure C19.6: RDFs of (a) TOL4, (b) PL1, and (c) TL2W

Analogue 1: Pyrrole
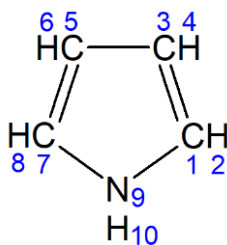
Figure C19.7: All-atom pyrrole with atom numbering
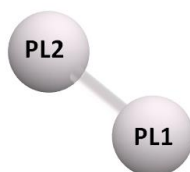
Figure C19.8: CG model of pyrrole showing bead types
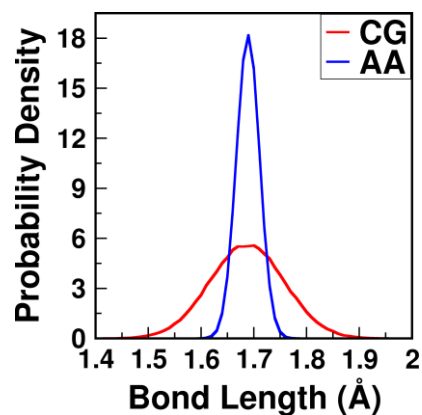
Figure C19.9: Distribution of bond PL1-PL2
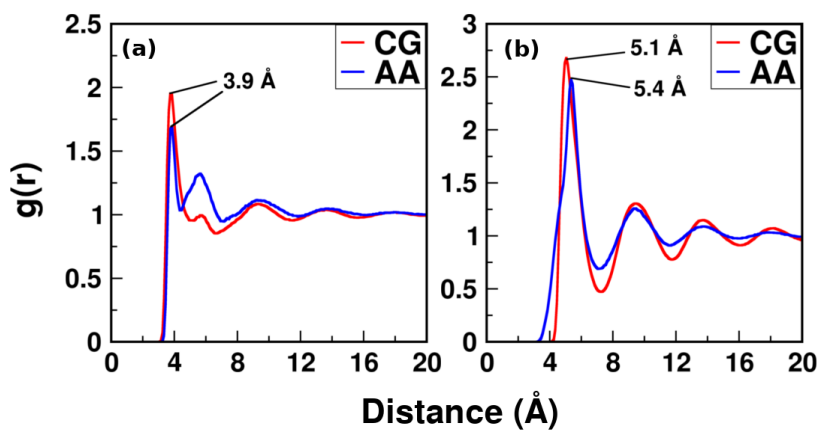
Figure C19.10: RDFs of (a) PL1 and (b) PL2

Analogue 2: Toluene
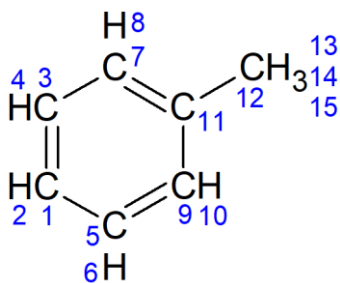
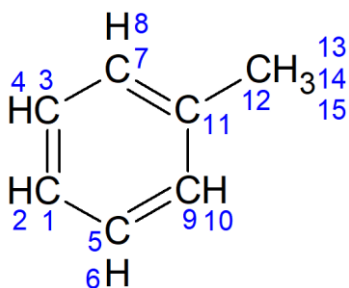**Figure C19.11:** All-atom toluene with atom numbering
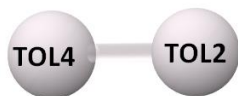
Figure C19.12: CG model of toluene showing bead types
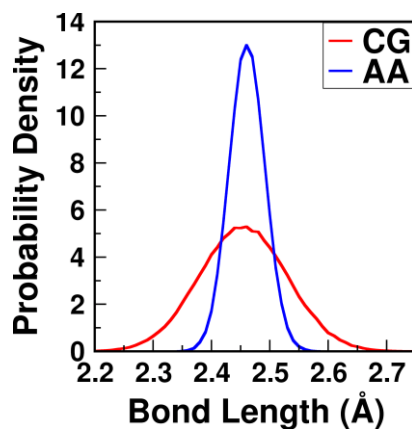
Figure C19.13: Distribution of bond TOL4-TOL2
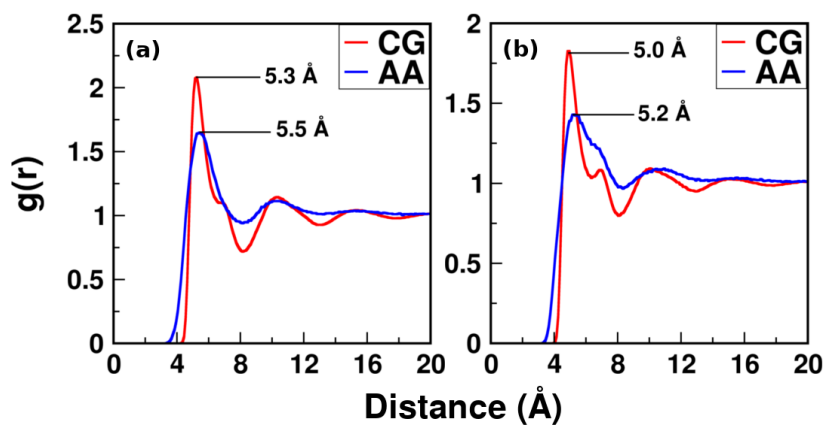
Figure C19.14: RDFs of (a) TOL4 and (b) TOL2

C20: Tyrosine

Tyr-Tyr Dipeptide
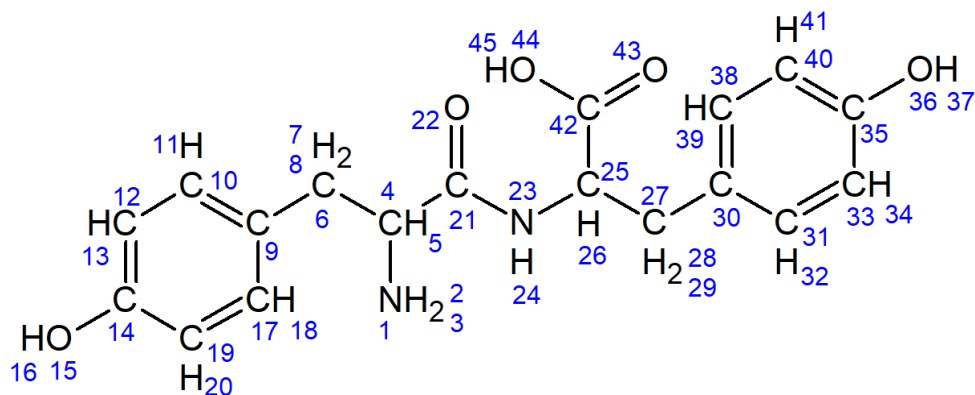
Figure C20.1: All-atom Tyrosine dipeptide with atom numbering
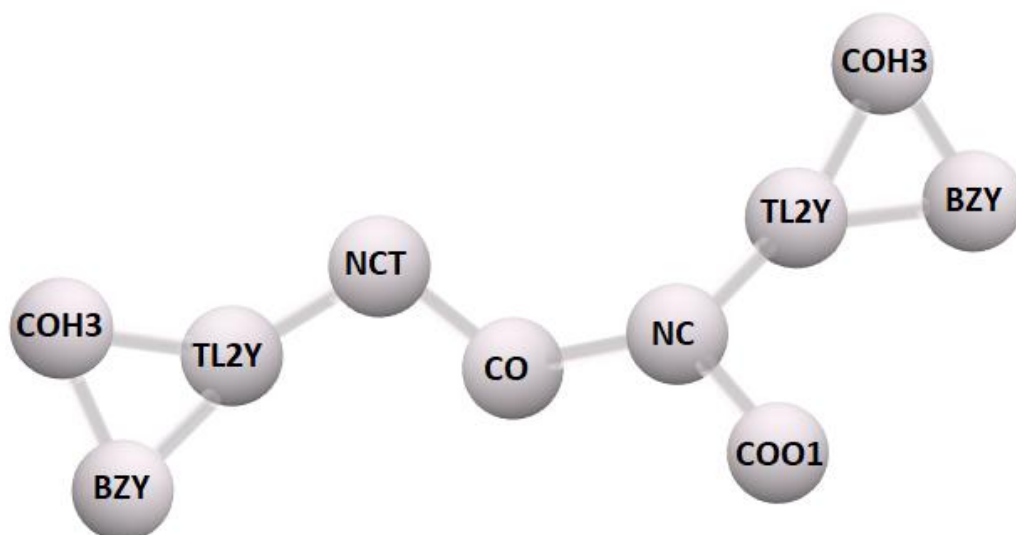
Figure C20.2: CG model of Tyrosine dipeptide showing bead types
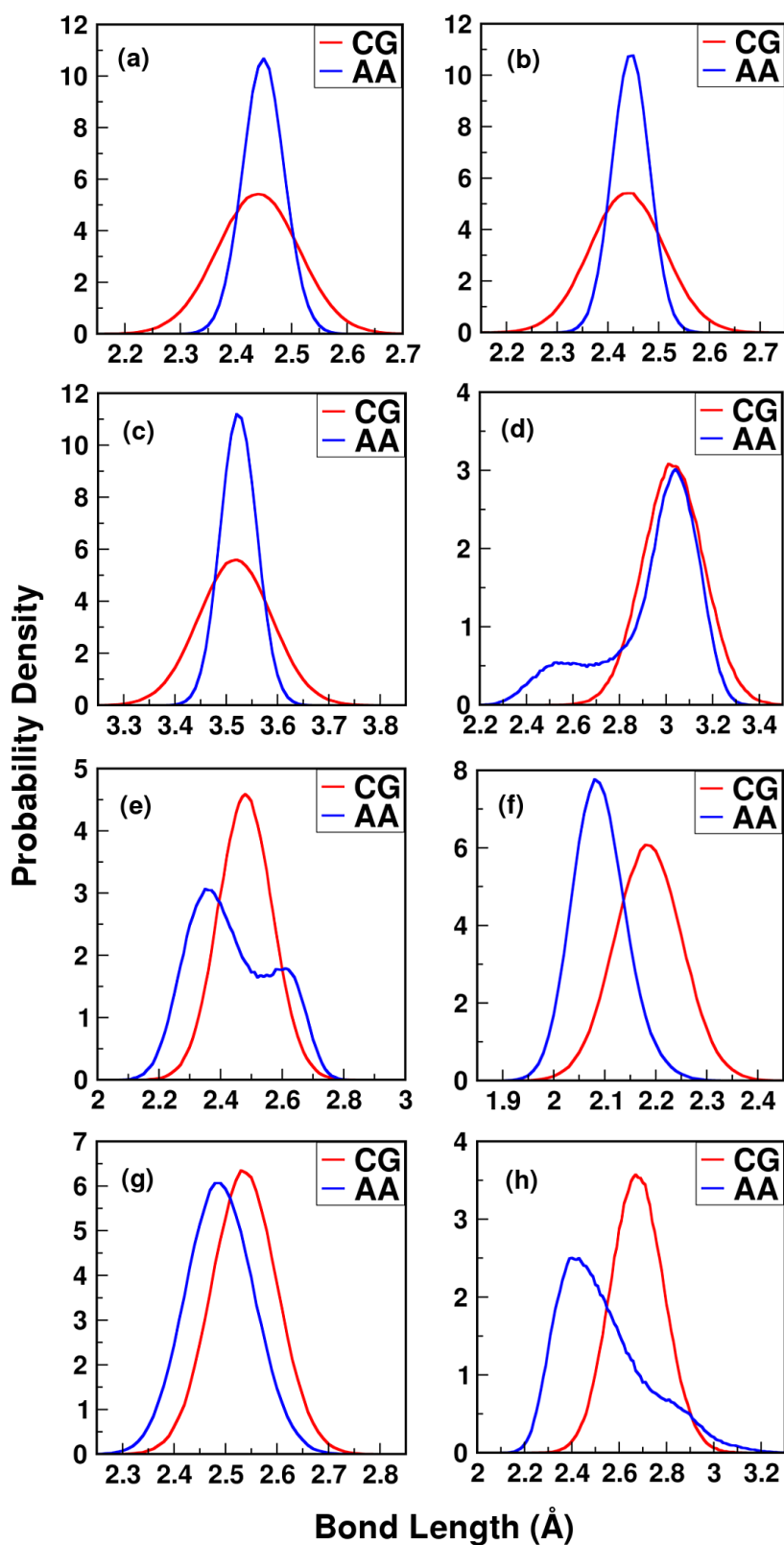
Figure C20.3: Distributions of bonds (a) TL2Y-BZY, (b) BZY-COH3, (c) COH3-TL2Y, (d) TL2Y-NCT, (e) NCT-CO, (f) CO-NC, (g) NC-COO1, and (h) NC-TL2Y
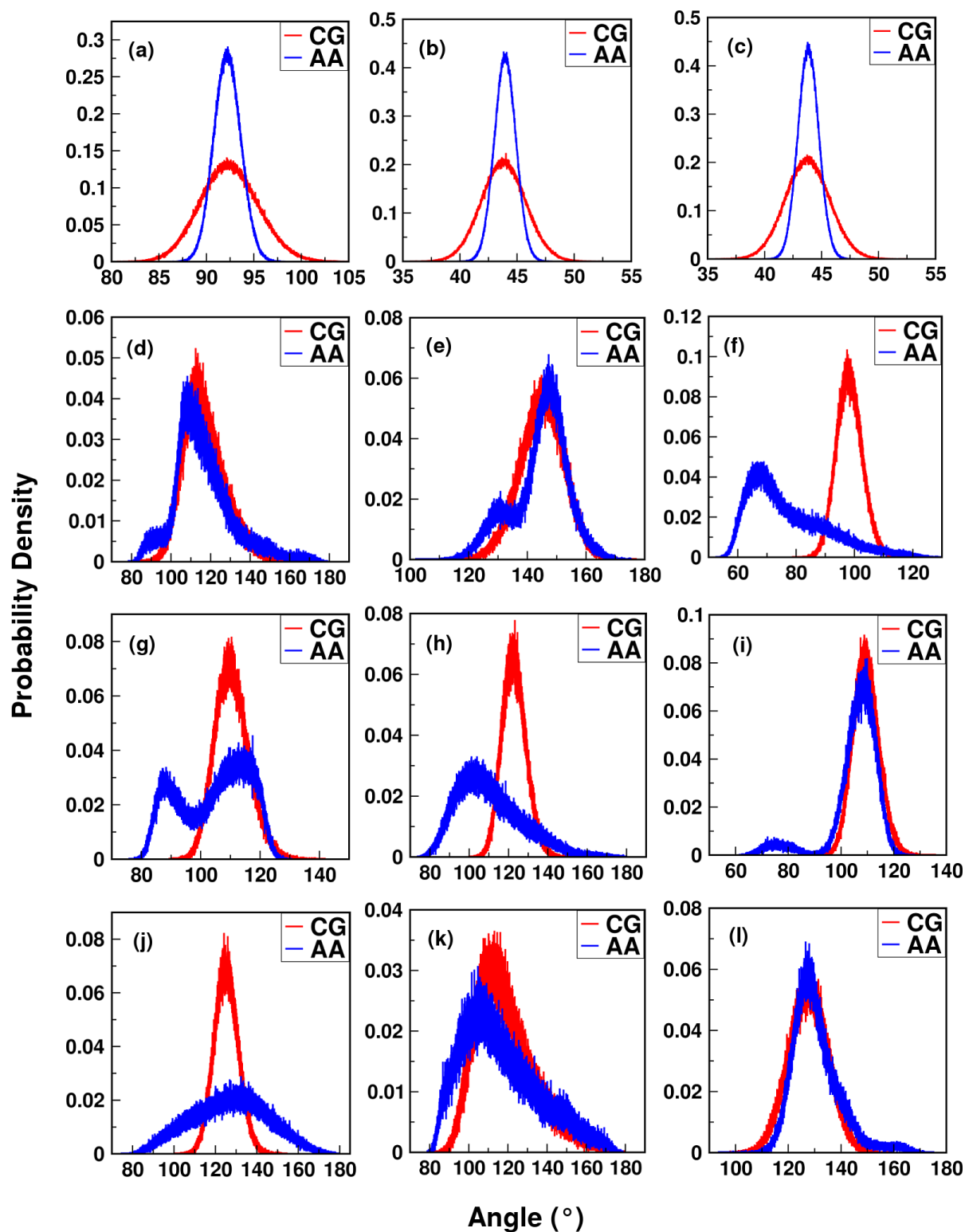
Figure C20.4: Distributions of angles (a) TL2Y-BZY-COH3, (b) BZY-YS3-TL2Y, (c) COH3-TL2Y-BZY, (d) BZY-TL2Y-NCT, (e) COH3-TL2Y-NCT, (f) TL2Y-NCT-CO, (g) NCT-CO-NC, (h) CO-NC-COO1, (i) TL2Y-NC-COO1, (j) CO-NC-TL2Y, (k) BZY-TL2Y-NC, and (l) COH3-TL2Y-NC

172
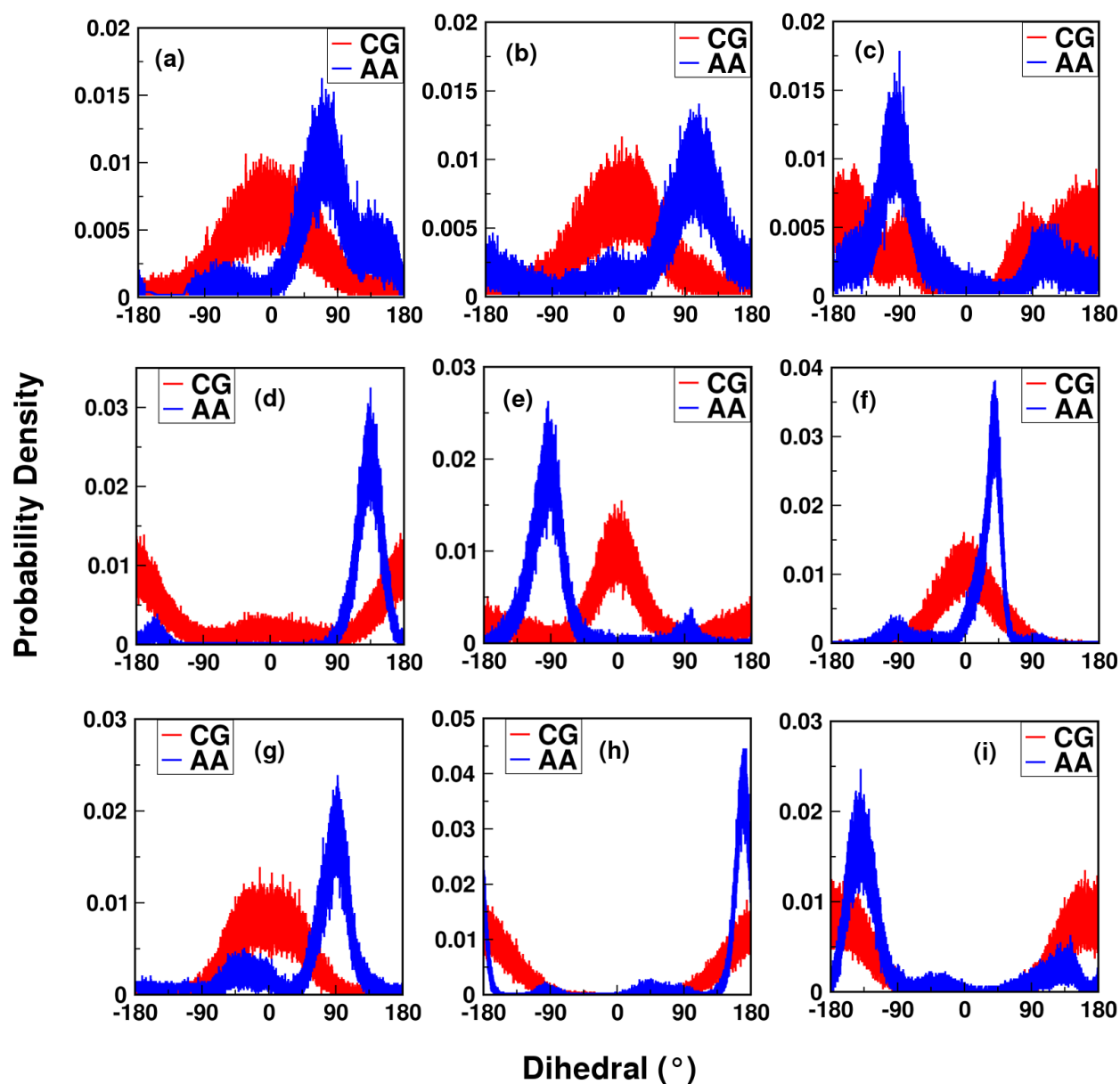
Figure C20.5: Distributions of dihedrals (a) COH3-TL2Y-NCT-CO, (b) BZ-TL2Y-NCT-CO, (c) TL2Y-NCT-CO-NC, (d) NCT-CO-NC-COO1, (e) NCT-CO-NC-TL2Y, (f) CO-NC-TL2Y-COH3, (g) CO-NC-TL2Y-BZ, (h) COH3-TL2Y-NC-COO1, and (i) BZ-TL2Y-NC-COO1
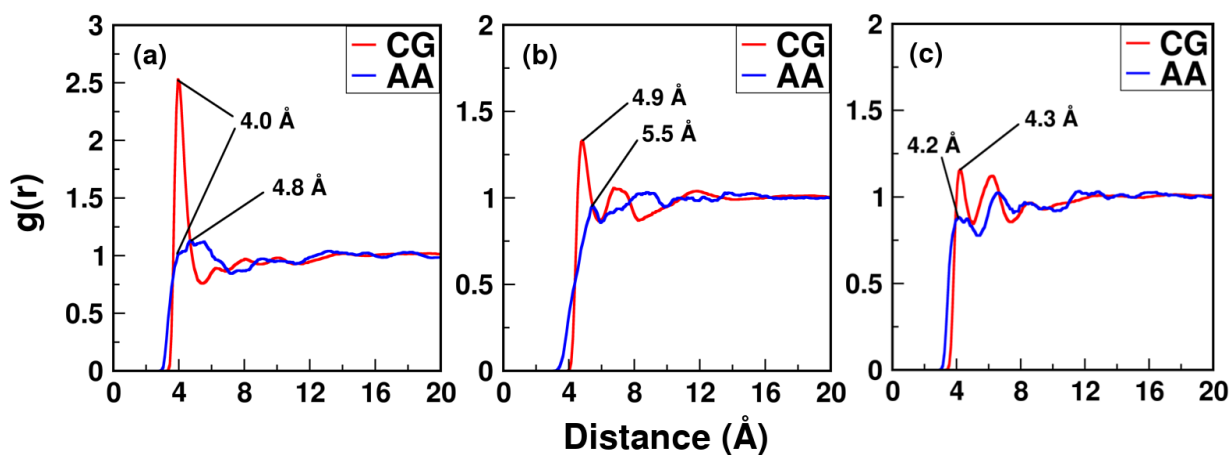
Figure C20.6: RDFs of (a) COH3, (b) TL2Y, and (c) BZY

Analogue: m-Cresol
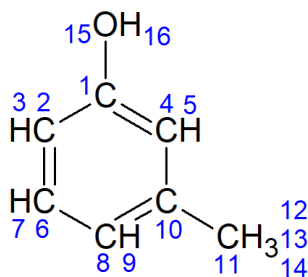
Figure C20.7: All-atom m-Cresol with atom numbering
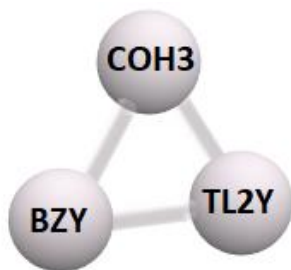
Figure C20.8: CG model of m-Cresol showing bead types
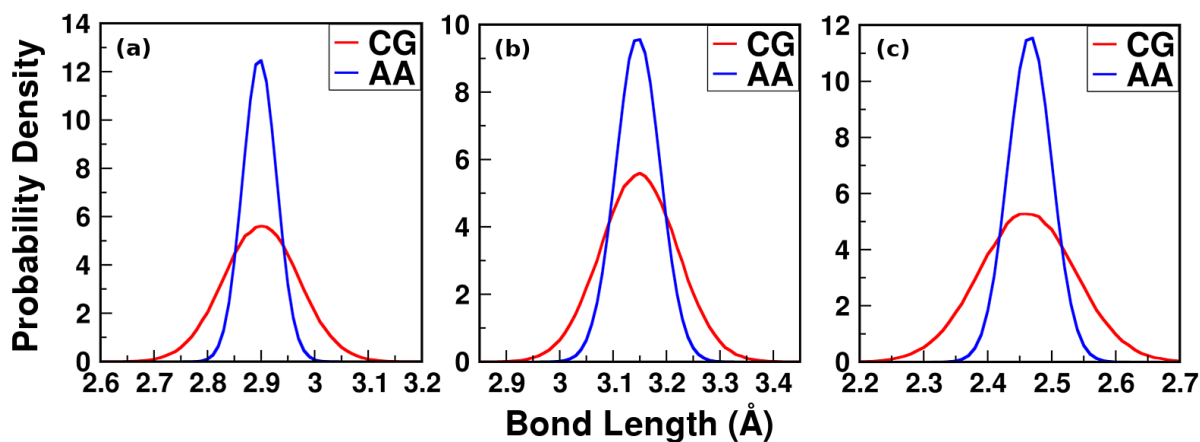
Figure C20.9: Distributions of bonds (a) COH3-BZY, (b) COH3-TL2Y, and (c) BZY-TL2Y
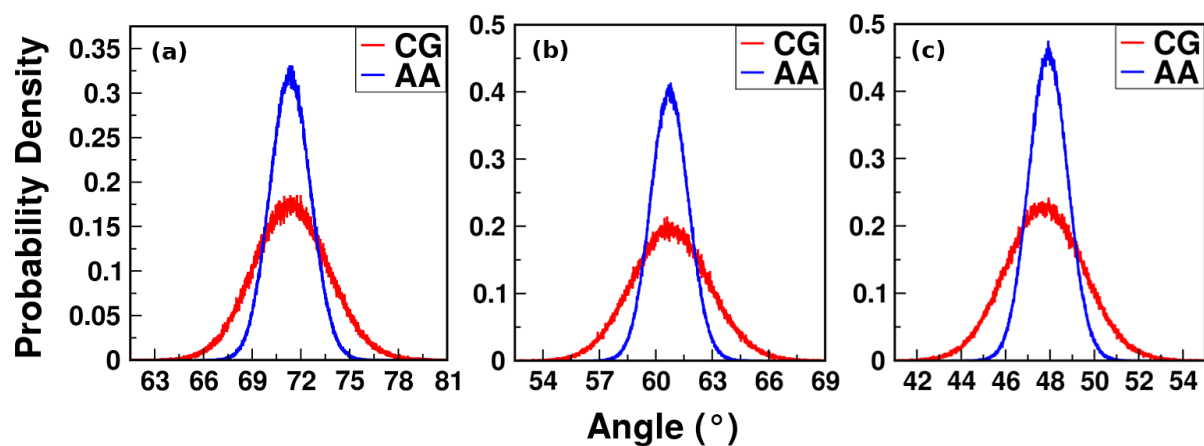
Figure C20.10: Distributions of angles (a) COH3-BZY-TL2Y, (b) COH3-TL2Y-BZY, and (c) BZY-COH3-TL2Y
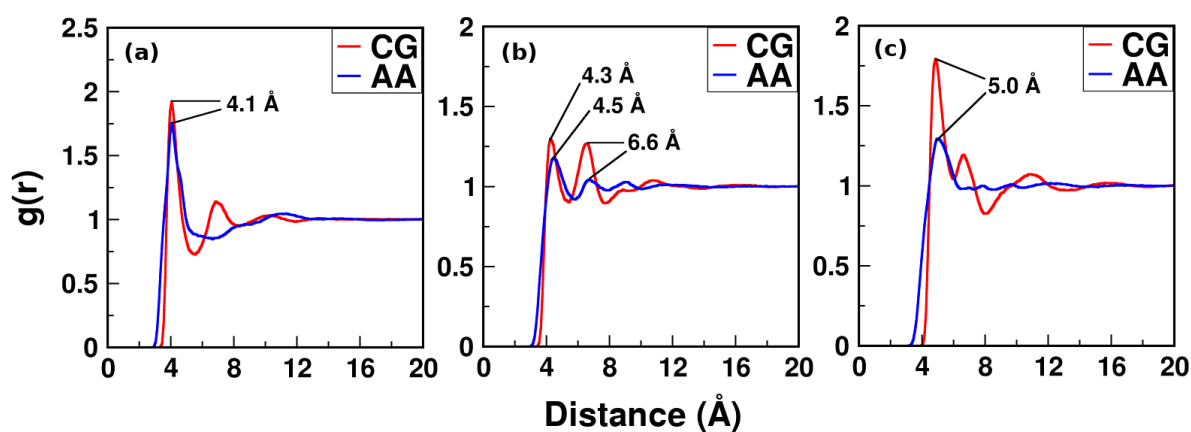
Figure C20.11: RDFs of (a) COH3, (b) BZY, and (c) TL2Y
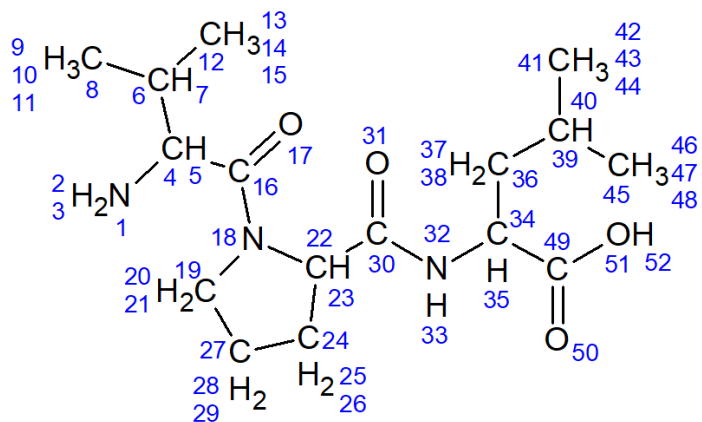
C21: Valine

Val-Val Dipeptide
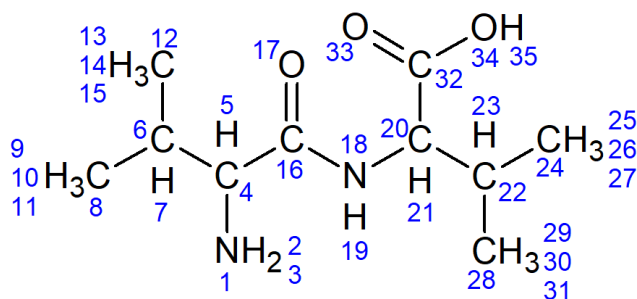
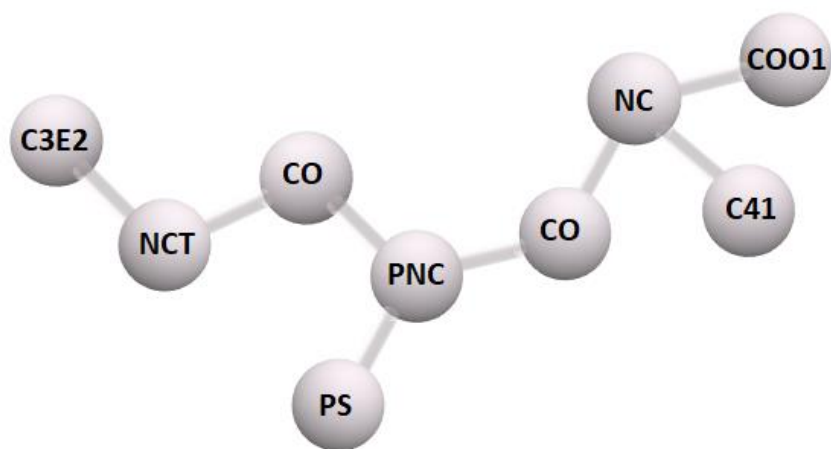**Figure C21.1:** All-atom Valine dipeptide with atom numbering
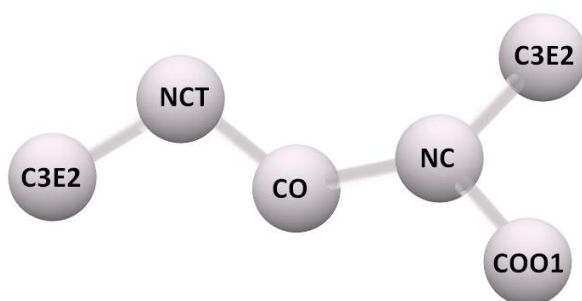
Figure C21.2: CG model of valine dipeptide showing bead types
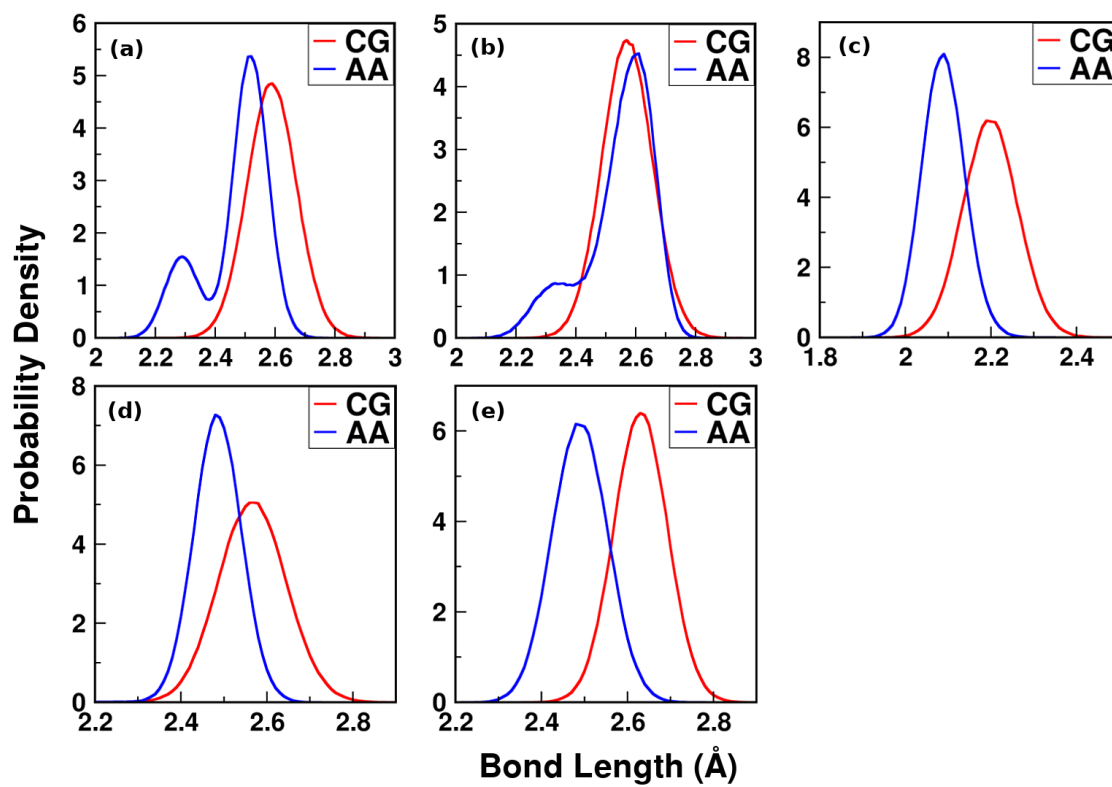
Figure C21.3: Distributions of bonds (a) C3E2-NCT, (b) NCT-CO, (c) CO-NC, (d) NC-C3E2, and (e) NC-COO1
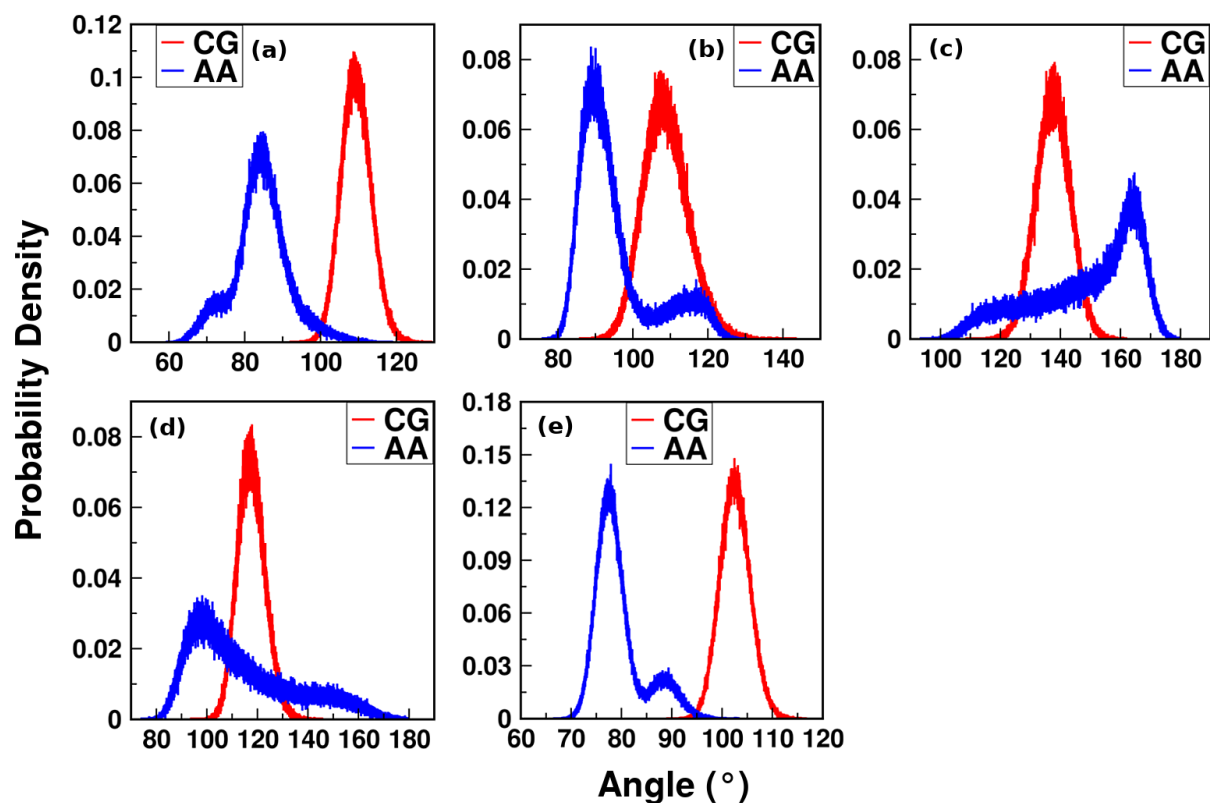
Figure C21.4: Distribution of angles (a) C3E2-NCT-CO (b) NCT-CO-NC, (c) CO-NC-C3E2, (d) CO-NC-COO1, and (e) C3E2-NC-COO1
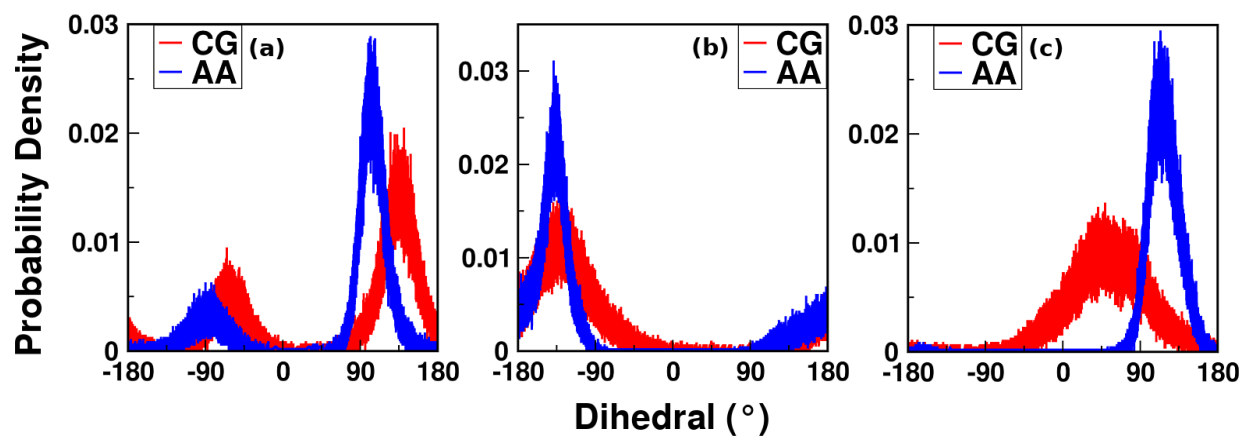
Figure C21.5: Distribution of dihedrals (a) C3E2-NCT-CO-NC, (b) NCT-CO-NC-C3E2, and (c) NCT-CO-NC-COO1
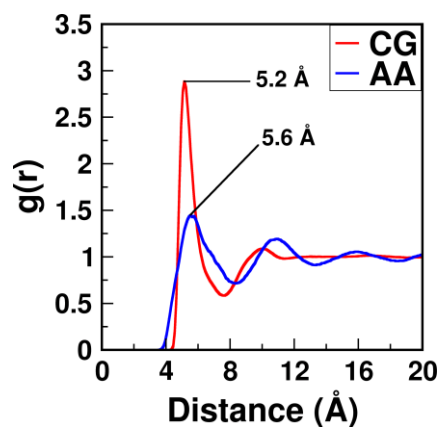
Figure C21.6: RDF of C3E2

Analogue: 2,3-dimethylbutane
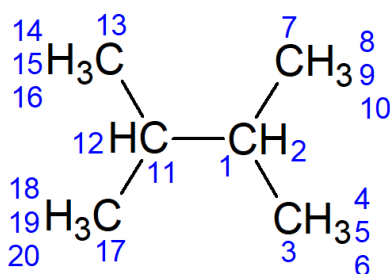
Figure C21.7: All-atom 2,3-dimethylbutane with atom numbering
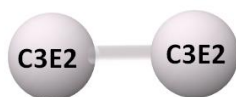
Figure C21.8: CG model of 2,3-dimethylbutane showing bead types
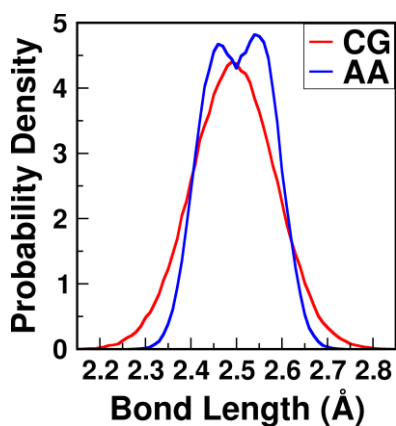
Figure C21.9: Distribution of bond C3E2-C3E2

179
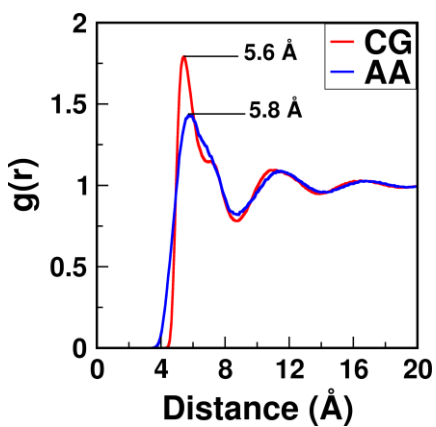
Figure C21.10: RDF of C3E2

Val-Val-Val Tripeptide
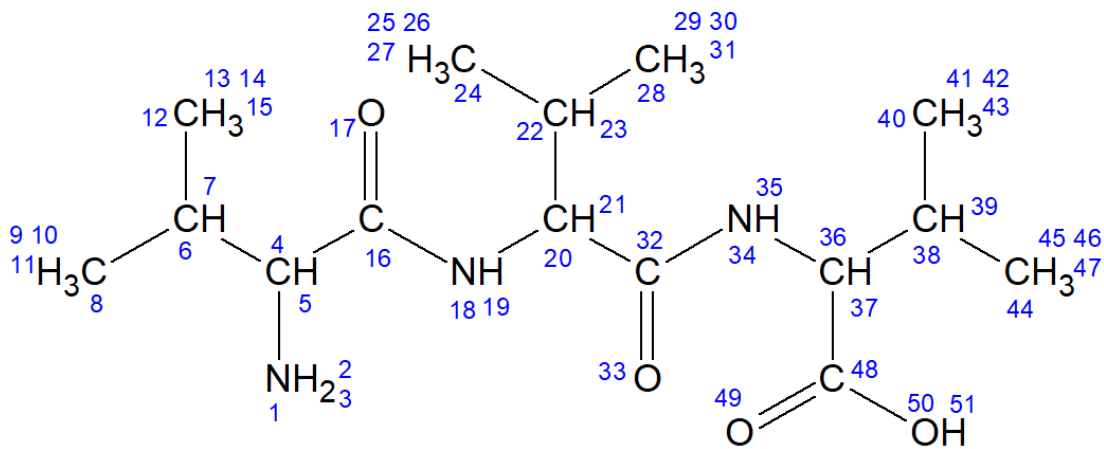
Figure C21.11: All-atom valine tripeptide with atom numbering
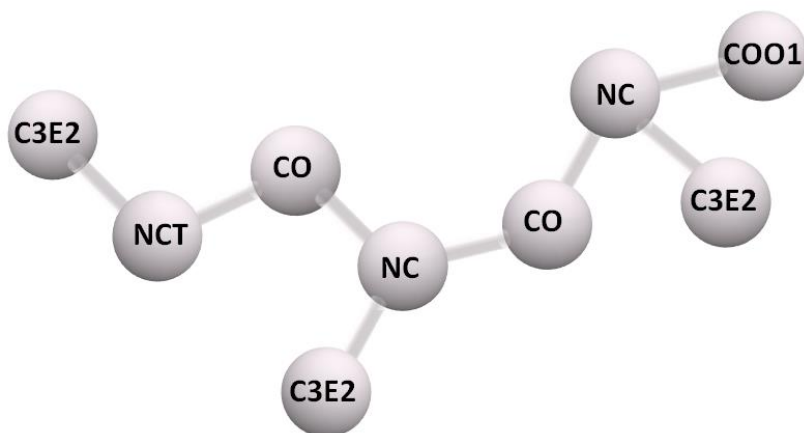
**Figure C21.12:** CG model of valine tripeptide showing bead types
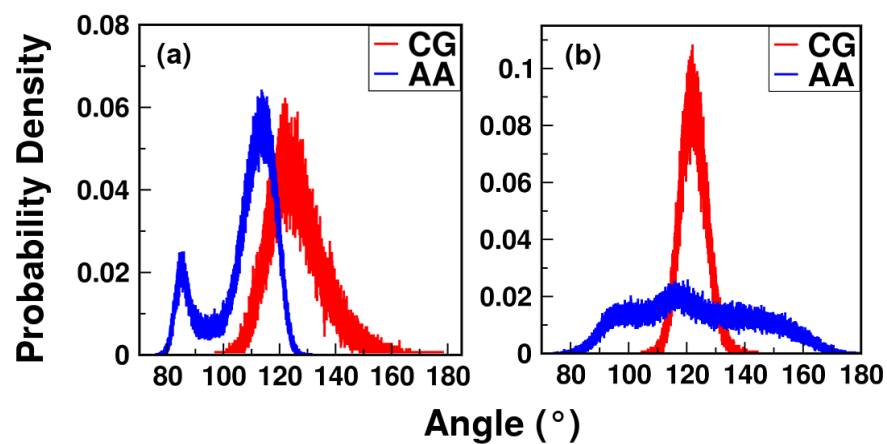
Figure C21.11: Angle distributions of (a) NC-CO-NC and (b) CO-NC-CO.
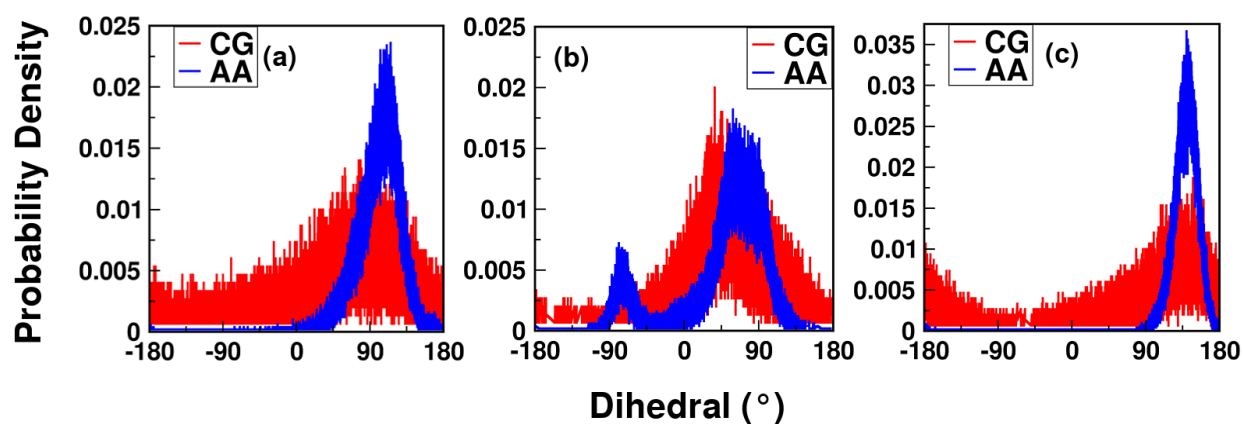
Figure C21.12: Dihedral distributions of (a) NT-CO-NC-CO, (b) CO-NC-CO-NC, and (c) NC-CO-NC-CT.

Appendix D: Topology, Parameter, and Configuration Files

D1: Topology File

!! Coarse-Grained Topology File for amino acids/peptides !!

```
MASS -1 C2E        29.07
MASS -1 C2M        28.06
MASS -1 C3E        43.10
MASS -1 C3M        42.09
MASS -1 CCR1       26.04 ! His
MASS -1 NCR1       28.04 ! His
MASS -1 NCR2       27.03 ! His
MASS -1 NCC1       44.09 ! Ala N-terminus
MASS -1 NCC2       43.08 ! Ala backbone
MASS -1 C41        57.13 ! Leu
MASS -1 CON2       58.07 ! Asn
MASS -1 COO2       59.05 ! Asp
MASS -1 RS1        28.06 ! Arginine side chain bead
MASS -1 RS2        29.05 ! Arginine side chain bead
MASS -1 RS3        43.06 ! Arginine side chain bead
MASS -1 SC         47.10 ! Cys
MASS -1 CON1       44.04 ! Glutamine side chain outer
MASS -1 GNT        30.06 ! Glycine N-terminus
MASS -1 GNC        29.05 ! Glycine backbone
MASS -1 C2E2       28.06 ! Ile side chain inner
MASS -1 CS         47.10 ! Methionine side chain outer
MASS -1 BZF2       26.04
MASS -1 BZF1       26.04
MASS -1 TL2F       39.06 ! Phenylalanine toluene bead
MASS -1 TL2W       39.06 ! Tryptophan toluene bead
MASS -1 TL2Y       39.06 ! Tyrosine toluene bead
MASS -1 PNC        27.03 ! Proline N-C bead
MASS -1 PS         42.09 ! Proline 3-C side chain bead
MASS -1 COH1       31.04 ! Serine side chain bead
MASS -1 CCOH       45.07 ! Threonine side chain
MASS -1 TOL4       52.08 ! Trp
MASS -1 PL1        40.05 ! Trp
MASS -1 COH3       42.04
MASS -1 BZY        26.04
MASS -1 C3E2       43.10
MASS -1 CO         28.01 ! backbone carbonyl
MASS -1 NC         28.04 ! backbone N-C
```

```
MASS -1 NC2          30.06
MASS -1 COO1         45.02 ! neutral C-terminus
MASS -1 NCT          29.05 ! neutral N-terminus
MASS -1 H2O1         36.04


RESI HCT             0.00
ATOM C1 C3E          0.00
ATOM C2 C3M          0.00
ATOM C3 C3M          0.00
ATOM C4 C3M          0.00
ATOM C5 C3M          0.00
ATOM C6 CO           0.00


BOND C1 C2  C2 C3  C3 C4  C4 C5  C5 C6


RESI ALA             0.00
ATOM AB     NCC2 0.00
ATOM BCO    CO   0.00


BOND AB BCO


RESI ARG             0.00
ATOM BNC    NC   0.00
ATOM RS1    RS1  0.00
ATOM RS2    RS2  0.00
ATOM RS3    RS3  0.00
ATOM BCO    CO   0.00


BOND BNC RS1  RS1 RS2  RS2 RS3  BNC BCO


RESI ASN             0.00
ATOM BNC    NC   0.00
ATOM NS     CON2 0.00
ATOM BCO    CO   0.00


BOND BNC NS  BNC CO


RESI ASP             0.00
ATOM BNC    NC   0.00
ATOM DS     COO2 0.00
ATOM BCO    CO   0.00


BOND BNC DS  BNC CO
```

```
 RESI CYS              0.00
ATOM BNC     NC     0.00
ATOM CS      SC     0.00
ATOM BCO     CO     0.00


BOND  BNC CS  BNC CO

RESI GLU              0.00
ATOM BNC     NC     0.00
ATOM ES1     C2M    0.00
ATOM ES2     COO1   0.00
ATOM BCO     CO     0.00


BOND  BNC ES1  ES1 ES2  BNC BCO

RESI GLN              0.00
ATOM BNC     NC     0.00
ATOM QS1     C2M    0.00
ATOM QS2     CON1   0.00
ATOM BCO     CO     0.00


BOND  BNC QS1  QS1 QS2  BNC BCO

RESI GLY              0.00
ATOM GNC     GNC    0.00
ATOM BCO     CO     0.00


BOND GNC BCO

RESI HIS              0.00
ATOM BNC     NC     0.00
ATOM HS1     CCR1   0.00
ATOM HS2     NCR1   0.00
ATOM HS3     NCR2   0.00
ATOM BCO     CO     0.00


BOND  BNC HS1  HS1 HS2  HS1 HS3  HS2 HS3  BNC BCO

RESI ILE              0.00
ATOM BNC     NC     0.00
ATOM IS1     C2E2   0.00
ATOM IS2     C2E    0.00
ATOM BCO     CO     0.00
```

BOND BNC IS1 IS1 IS2 BNC BCO

RESI LEU   0.00
ATOM BNC   NC   0.00
ATOM LS   C41   0.00
ATOM BCO   CO   0.00

BOND BNC LS BNC BCO

RESI LYS   0.00
ATOM BNC   NC   0.00
ATOM KS1   C3M   0.00
ATOM KS2   NC2   0.00
ATOM BCO   CO   0.00

BOND BNC KS1 KS1 KS2 BNC BCO

RESI MET   0.00
ATOM BNC   NC   0.00
ATOM MS1   C2M   0.00
ATOM MS2   CS   0.00

BOND BNC MS1 MS1 MS2 BNC BCO

RESI PHE   0.00
ATOM BNC   NC   0.00
ATOM FS1   TL2F   0.00
ATOM FS21   BZF1   0.00
ATOM FS22   BZF2   0.00
ATOM BCO   CO   0.00

BOND BNC FS1 FS1 FS21 FS1 FS22 FS21 FS22 BNC BCO

RESI PRO   0.00
ATOM PNC   PNC   0.00
ATOM PS   PS   0.00
ATOM BCO   CO   0.00

BOND PNC PS PNC BCO

RESI SER   0.00
ATOM BNC   NC   0.00

ATOM SS   COH1   0.00
ATOM BCO   CO   0.00

BOND BNC SS BNC BCO

RESI THR   0.00
ATOM BNC   NC   0.00
ATOM TS   CCOH   0.00
ATOM BCO   CO   0.00

BOND BNC TS BNC BCO

RESI TRP   0.00
ATOM BNC   NC   0.00
ATOM WS1   TL2W   0.00
ATOM WS2   PL1   0.00
ATOM WS3   TOL4   0.00

BOND BNC WS1 WS1 WS2 WS1 WS3 WS2 WS3 BNC BCO

RESI TYR   0.00
ATOM BNC   NC   0.00
ATOM YS1   TL2Y   0.00
ATOM YS2   BZY   0.00
ATOM YS3   COH3   0.00
ATOM BCO   CO   0.00

BOND BNC YS1 YS1 YS2 YS1 YS3 YS2 YS3 BNC BCO

RESI VAL   0.00
ATOM BNC   NC   0.00
ATOM VS   C3E2   0.00
ATOM BCO   CO   0.00

BOND BNC VS BNC BCO

PRES PEPB   ! Normal amino acid linkage
ATOM 1BCO   CO   0.00
ATOM 2BNC   NC   0.00

BOND 1BCO 2BNC

```
PRES PEPG                    ! Amino acid to glycine linkage
ATOM 1BCO    CO      0.00
ATOM 2GNC    GNC     0.00

BOND 1BCO 2GNC

PRES PEPP                    ! Amino acid to proline linkage
ATOM 1BCO    CO      0.00
ATOM 2PNC    PNC     0.00

BOND 1BCO 2PNC

PRES AT                      0.00 ! Neutral alanine N-terminus
ATOM ANT     NCC1    0.00
ATOM BCO     CO      0.00
DELETE ATOM AB
BOND ANT BCO

PRES NTER                    0.00 ! Neutral standard N-terminus
ATOM NT      NCT     0.00
ATOM BCO     CO      0.00
DELETE ATOM BNC
BOND NT BCO

PRES GT                      0.00 ! Neutral glycine N-terminus
ATOM GNT     GNT     0.00
ATOM BCO     CO      0.00
DELETE ATOM GNC
BOND GNT BCO

PRES CTER                    0.00 ! Neutral C-terminus
ATOM CT      COO1    0.00
ATOM BNC     NC      0.00
DELETE ATOM BCO
BOND  BNC CT

PRES HCTA                    0.00
ATOM 1C6     CO      0.00
ATOM 2AB     NCC2    0.00
BOND 1C6 2AB

RESI WAT                     0.00
ATOM H2O     H2O1    0.00
```

D2: Parameter File

!! Coarse-Grained Parameter File for the amino acids and peptides !!
!! This is for use with "exclude 1-2" ONLY.   !!
!! Only neutral amino acids and neutral termini are represented.   !!

```
ATOMS
MASS -1 C2E      29.07
MASS -1 C2M      28.06
MASS -1 C3E      43.10
MASS -1 C3M      42.09
MASS -1 CCR1     26.04 ! His
MASS -1 NCR1     28.04 ! His
MASS -1 NCR2     27.03 ! His
MASS -1 NCC1     44.09 ! Ala N-terminus
MASS -1 NCC2     43.08 ! Ala backbone
MASS -1 C41      57.13 ! Leu
MASS -1 CON2     58.07 ! Asn
MASS -1 COO2     59.05 ! Asp
MASS -1 RS1      28.06 ! Arginine side chain bead
MASS -1 RS2      29.05 ! Arginine side chain bead
MASS -1 RS3      43.06 ! Arginine side chain bead
MASS -1 SC       47.10 ! Cys
MASS -1 CON1     44.04 ! Glutamine side chain outer
MASS -1 GNT      30.06 ! Glycine N-terminus
MASS -1 GNC      29.05 ! Glycine backbone
MASS -1 C2E2     28.06 ! Ile side chain inner
MASS -1 CS       47.10 ! Methionine side chain outer
MASS -1 BZF2     26.04
MASS -1 BZF1     26.04
MASS -1 TL2F     39.06 ! Phenylalanine toluene bead
MASS -1 TL2W     39.06 ! Tryptophan toluene bead
MASS -1 TL2Y     39.06 ! Tyrosine toluene bead
MASS -1 PNC      27.03 ! Proline N-C bead
MASS -1 PS       42.09 ! Proline 3-C side chain bead
MASS -1 COH1     31.04 ! Serine side chain bead
MASS -1 CCOH     45.07 ! Threonine side chain
MASS -1 TOL4     52.08 ! Trp
MASS -1 PL1      40.05 ! Trp
MASS -1 COH3     42.04
MASS -1 BZY      26.04
MASS -1 C3E2     43.10
MASS -1 CO       28.01 ! backbone carbonyl
```

```
MASS  -1  NC            28.04 ! backbone N-C
MASS  -1  NC2           30.06
MASS  -1  COO1          45.02 ! neutral C-terminus
MASS  -1  NCT           29.05 ! neutral N-terminus
MASS  -1  H2O1          36.04


BONDS
!
!V(bond) = Kb(r - r0)**2
!
!Kb: kcal/mole/A**2
!r0: A
!
!bead types            Kb      r0

! Alanine
NCC1  CO               15      2.57
CO      NCC2           10      2.63
NCC2  COO1             30      2.54

! Asparagine
CON2   NCT             8       3.00
NC      CON2           8       2.75

! Aspartic Acid
COO2   NCT             8       3.00
NC      COO2           55      2.71

! Arginine
NC      RS1            10      2.30 ! RS1=C2M
RS1     RS2            15      2.53
RS2     RS3            50      2.4

! Cysteine
SC      NCT            20      2.65
NC      SC             55      2.58

! Glutamic Acid
COO1   C2M            90.98   2.5
C2M     NCT            15      2.45 ! Same as Gln, Met
NC      C2M            34      2.30 ! Same as Gln, Met
```

! Glutamine
```
CON1  C2M        42    2.44
C2M   NCT        15    2.45 ! Same as Glu, Met
NC    C2M        34    2.30 ! Same as Glu, Met
```

! Glycine
```
GNT   CO         50    2.50
CO    GNC        70    2.15
GNC   COO1       60    2.50
```

! Histidine
```
NCR2  NCR1       50    1.86
NCR1  CCR1       30    2.45
NCR2  CCR1       50    2.53
CCR1  NCT        30    2.35
NC    CCR1       80    2.26
```

! Isoleucine
```
C2E   C2E2       44    2.41
C2E2  NCT        40    2.50
NC    C2E2       30    2.28
```

! Leucine
```
C41   NCT        12    3.18
NC    C41        10    2.95
```

! Lysine
```
NC2   C3M        20    3.07
C3M   NCT        5     2.90
NC    C3M        10    2.52
```

! Methionine
```
CS    C2M        32    2.5
C2M   NCT        15    2.45 ! Same as Glu, Gln
NC    C2M        34    2.30 ! Same as Glu, Gln
```

! Phenylalanine
```
BZF1  BZF2       62    2.23
BZF1  TL2F       55    2.875
BZF2  TL2F       53    2.45
TL2F  NCT        7     2.75
NC    TL2F       17    2.43
```

190

! Proline
CO   PNC   40   2.04
PNC   PS   45   1.77

! Serine
COH1   NCT   30   2.5
NC   COH1   35   2.28

! Threonine
CCOH   NCT   55   2.58
NC   CCOH   45   2.25

! Tryptophan
TOL4   PL1   60   3.11
TOL4   TL2W   40   3.358
TL2W   PL1   60   2.195
TL2W   NCT   20   2.90
NC   TL2W   10   2.50

! Tyrosine
TL2Y   BZY   50   2.45
BZY   COH3   50   2.45
COH3   TL2Y   50   3.52
TL2Y   NCT   15   2.83
NC   TL2Y   15   2.42

! Valine
C3E2   NCT   40   2.44
NC   C3E2   35   2.26

! Backbone
NCT   CO   35   2.35
CO   NC   60   2.08
NC   COO1   70   2.48

! Carbon-Carbon & Carbon-Peptide
C3E   C3M   37.695   3.561
C3M   C3M   37.695   3.561 ! from C3E-C3M
C3M   CO   38.537   2.914 ! from C2M-C3M
C2E   C2M   35.293   2.561
C2M   C2M   36.609   2.520

```
ANGLES
!
!V(angle) = Ktheta(Theta - Theta0)**2
!
!Ktheta: kcal/mole/rad**2
!Theta0: degrees
!
!bead types                   Ktheta Theta0

! Alanine
NCC1   CO     NCC2        20     97
CO     NCC2   CO          5      91
NCC2   CO     NCC2        17     96
CO     NCC2   COO1        6      83

! Asparagine
CON2   NCT    CO          6      68
CO     NC     CON2        3      138
CON2   NC     COO1        50     114

! Aspartic Acid
COO2   NCT    CO          5      75
CO     NC     COO2        2      138
COO2   NC     COO1        27     114

! Arginine
CO     NC     RS1         10     140
RS1    NC     COO1        20     83
NC     RS1    RS2         10     153
RS1    RS2    RS3         5      157

! Cysteine
SC     NCT    CO          10     71
CO     NC     SC          30     166
SC     NC     COO1        35     84

! Glutamic Acid
COO1   C2M    NCT         25     114
C2M    NCT    CO          10     73     ! Same in Gln, Met
CO     NC     C2M         12     173    ! Same in Gln, Met
NC     C2M    COO1        25     162
C2M    NC     COO1        20     83     ! Same in Gln, Met
```

192

! Glutamine

CON1	C2M	NCT	30	100	
C2M	NCT	CO	10	73	! Same in Gln, Met
CO	NC	C2M	12	173	! Same in Gln, Met
NC	C2M	CON1	30	160	
C2M	NC	COO1	20	83	! Same in Gln, Met

! Glycine

GNT	CO	GNC	30	90
CO	GNC	COO1	20	97

! Histidine

NCR2	CCR1	NCR1	40	44	! This and next 3 must add to 180deg
CCR1	NCR1	NCR2	25	70	
NCR1	NCR2	CCR1	25	66	
NCR1	CCR1	NCT	5	140	
CCR1	NCT	CO	40	68	
CO	NC	CCR1	1	140	
COO1	NC	CCR1	100	90	
NC	CCR1	NCR2	7	124	
NCR2	CCR1	NCT	10	135	
NC	CCR1	NCR1	7	120	

! Isoleucine

C2E	C2E2	NCT	10	70
C2E2	NCT	CO	30	70
CO	NC	C2E2	10	146
NC	C2E2	C2E	20	70
C2E2	NC	COO1	40	100

! Leucine

C41	NCT	CO	6	73
CO	NC	C41	2	144
C41	NC	COO1	6	110

! Lysine

NC2	C3M	NCT	3	140
C3M	NCT	CO	3	75
CO	NC	C3M	1	140
NC	C3M	NC2	5	140
C3M	NC	COO1	40	106

! Methionine
```
CS     C2M    NCT        2      120
C2M    NCT    CO         10     73      ! Same in Gln, Glu
CO     NC     C2M        12     173     ! Same in Gln, Glu
NC     C2M    CS         2      130
C2M    NC     COO1       20     83      ! Same in Gln, Glu
```

! Phenylalanine
```
BZF2   BZF1   TL2F       31     55.6
BZF1   BZF2   TL2F       32     75.75
BZF1   TL2F   BZF2       33     48.65
BZF2   TL2F   NCT        7      110
TL2F   NCT    CO         8      70
CO     NC     TL2F       5      135
COO1   NC     TL2F       40     108
NC     TL2F   BZF1       50     129.5
NC     TL2F   BZF2       5      110
BZF1   TL2F   NCT        10     145
```

! Proline
```
NCT    CO     PNC        5      111
CO     PNC    PS         8      144
CO     PNC    CO         3      106
PNC    CO     NC         10     103
```

! Serine
```
COH1   NCT    CO         15     74
CO     NC     COH1       30     173
COH1   NC     COO1       25     81
```

! Threonine
```
CCOH   NCT    CO         30     83
CO     NC     CCOH       7      157
CCOH   NC     COO1       50     93
```

! Tryptophan
```
TL2W   PL1    TOL4       40     76.4  ! this and next three = 180
PL1    TOL4   TL2W       40     39.4
TOL4   TL2W   PL1        40     64.2
TOL4   TL2W   NCT        10     140
PL1    TL2W   NCT        10     135
TL2W   NCT    CO         20     75
```

! Tyrosine

TL2Y	BZY	COH3	25	92
BZY	COH3	TL2Y	25	44
COH3	TL2Y	BZY	25	44
BZY	TL2Y	NCT	5	117
COH3	TL2Y	NCT	15	145
TL2Y	NCT	CO	15	67
TL2Y	NC	COO1	12	108
CO	NC	TL2Y	2	123
BZY	TL2Y	NC	2	110
COH3	TL2Y	NC	20	128

! Valine

C3E2	NCT	CO	20	75
CO	NC	C3E2	2	160
C3E2	NC	COO1	28	77

! Carbon-Carbon & Carbon-Peptide

C3E	C3M	C3M	3.4	145	! based on C3E-C3M-C3E
C3M	C3M	C3M	3.4	145	! based on C3E-C3M-C3E
C3M	C3M	CO	3.4	145	! based on C3E-C3M-C3E
C3M	CO	NCC2	1	140	! from Lys CO-NC-C3M

C2E	C2M	C2M	3.886	147.766
C2M	C2M	C2M	3.460	144.226

! Backbone

NCT	CO	NC	7	90	
CO	NC	COO1	1	100	
NC	CO	NC	8	103	! from valylvalylvaline
CO	NC	CO	1	120	! from valylvalylvaline
NC	CO	NCC2	20	97	! from NCC1 CO NCC2 (alanine)

DIHEDRALS
!
!V(dihedral) = Kchi(1 + cos(n(chi) - delta))
!
!Kchi: kcal/mole
!n: multiplicity
!delta: degrees
!

! bead types				Kchi	n	delta
! Alanine						
NCC1	CO	NCC2	CO	1.0	1	90
CO	NCC2	CO	NCC2	0.6	1	0
NCC2	CO	NCC2	COO1	1.2	1	0
! Asparagine						
NS	NCT	CO	NC	0.3	2	0
NCT	CO	NC	NS	1.45	1	100
! Aspartic Acid						
DS	NCT	CO	NC	0.3	2	0
NCT	CO	NC	DS	1.7	1	102
! Arginine						
ANT	CO	NC	RS1	0.7	2	180
CO	NC	RS1	RS2	0.7	2	0
COO1	NC	RS1	RS2	0.7	1	180
NC	RS1	RS2	RS3	0.7	2	0
! Cysteine						
CS	NCT	CO	NC	1.0	1	0
NCT	CO	NC	CS	0.6	2	0
! Glutamic Acid						
COO1	C2M	NCT	CO	1.0	2	0
C2M	NCT	CO	NC	0.7	2	0
NCT	CO	NC	C2M	0.7	2	0
CO	NC	C2M	COO1	1.0	1	90
COO1	C2M	NC	COO1	1.0	1	90
! Glutamine						
CON1	C2M	NCT	CO	0.5	2	0
C2M	NCT	CO	NC	0.7	2	0
NCT	CO	NC	C2M	0.7	2	0
CO	NC	C2M	CON1	0.5	2	0
CON1	C2M	NC	COO1	0.5	1	180
! Glycine						
GNT	CO	GNC	COO1	0.5	2	0

! Histidine

NCR1	CCR1	NCT	CO	0.1	3	180
NCR2	CCR1	NCT	CO	0.1	1	180
CCR1	NCT	CO	NC	0.1	2	81
NCT	CO	NC	CCR1	0.5	1	60
CO	NC	CCR1	NCR2	0.5	1	180
CO	NC	CCR1	NCR1	0.5	1	180
COO1	NC	CCR1	NCR2	0.8	1	0
COO1	NC	CCR1	NCR1	0.5	1	0

! Isoleucine

C2E	C2E2	NCT	CO	0.5	1	0
C2E2	NCT	CO	NC	0.5	2	0
NCT	CO	NC	C2E2	0.5	1	90
CO	NC	C2E2	C2E	0.5	1	270
C2E	C2E2	NC	COO1	0.5	2	180

! Leucine

C41	NCT	CO	NC	0.5	1	0
NCT	CO	NC	C41	0.6	1	90

! Lysine

NC2	C3M	NCT	CO	0.2	2	76
C3M	NCT	CO	NC	0.7	1	90
NCT	CO	NC	C3M	2.0	1	120
CO	NC	C3M	NC2	0.2	3	180
NC2	C3M	NC	COO1	0.1	3	0

! Methionine

CS	C2M	NCT	CO	0.5	1	0
C2M	NCT	CO	NC	0.7	2	0
NCT	CO	NC	C2M	0.7	2	0
CO	NC	C2M	CS	0.5	1	0
CS	C2M	NC	COO1	0.4	1	60

! Phenylalanine

BZF1	TL2F	NCT	CO	0.3	2	0
BZF2	TL2F	NCT	CO	0.4	2	0
TL2F	NCT	CO	NC	0.3	2	76
NCT	CO	NC	TL2F	1	1	90
CO	NC	TL2F	BZF1	0.5	1	270
CO	NC	TL2F	BZF2	0.3	1	120
BZF1	TL2F	NC	COO1	0.3	1	270

BZF2	TL2F	NC	COO1	0.5	2	180

! Proline

C3E2	NCT	CO	PNC	1.0	2	0
NCT	CO	PNC	PS	0.3	1	60
NCT	CO	PNC	CO	0.3	1	60
CO	PNC	CO	NC	0.5	1	180
PS	PNC	CO	NC	0.3	2	0
PNC	CO	NC	C41	0.5	1	60
PNC	CO	NC	COO1	1.7	1	0

! Serine

COH1	NCT	CO	NC	0.5	2	0
NCT	CO	NC	COH1	0.8	1	0

! Threonine

CCOH	NCT	CO	NC	0.3	2	0
NCT	CO	NC	CCOH	1	1	90

! Tryptophan

TOL4	TL2W	NCT	CO	0.8	2	180
PL1	TL2W	NCT	CO	0.7	2	0
TL2W	NCT	CO	NC	1.0	2	0
NCT	CO	NC	TL2W	1.0	1	180
CO	NC	TL2W	PL1	1.0	1	180
CO	NC	TL2W	TOL4	1.0	1	180
COO1	NC	TL2W	PL1	1.2	1	0
COO1	NC	TL2W	TOL4	1.3	1	0

! Tyrosine

COH3	TL2Y	NCT	CO	0.5	1	180
BZY	TL2Y	NCT	CO	0.5	1	180
TL2Y	NCT	CO	NC	0.5	1	0
NCT	CO	NC	TL2Y	0.5	2	180
CO	NC	TL2Y	COH3	0.7	1	180
CO	NC	TL2Y	BZY	0.5	1	180
COH3	TL2Y	NC	COO1	1.0	1	0
BZY	TL2Y	NC	COO1	0.5	1	0

! Valine

C3E2	NCT	CO	NC	1.3	2	90
NCT	CO	NC	C3E2	2.0	1	43

! Backbone

NCT	CO	NC	COO1	0.5	1	0	
NCT	CO	NC	CO	2.0	1	0	! from valylvalylvaline (t8)
CO	NC	CO	NC	0.1	2	0	! from valylvalylvaline (t8)
NC	CO	NC	COO1	2.5	1	0	! from valylvalylvaline (t8)
X	X	X	X	0	0	0	! sets all others not listed to 0

! So all dihedrals like LS1-NC-CO-NC and C3M-C3M-C3M are set to 0!

NONBONDED

!

!V(Lennard-Jones) = Eps,i,j[(Rmin,i,j/ri,j)**12 - 2(Rmin,i,j/ri,j)**6]

!

!epsilon: kcal/mole, Eps,i,j = sqrt(eps,i * eps,j)

!Rmin/2: A, Rmin,i,j = Rmin/2,i + Rmin/2,j

!

!bead	ignored	epsilon	Rmin/2	
CCR1	0.000000	-0.3058	2.4725	
NCR1	0.000000	-0.7506	1.8759	
NCR2	0.000000	-0.4417	2.1546	
NCT	0.000000	-0.6000	2.1650	! same as NC2 (Lys bead)
CO	0.000000	-0.6690	2.4361	
NC	0.000000	-0.6765	2.1436	
COO1	0.000000	-1.2516	2.2088	! from Yaxin's propionic acid bead
NCC1	0.000000	-0.8699	2.4014	
NCC2	0.000000	-0.8699	2.4014	
C41	0.000000	-0.7120	2.8515	
NC2	0.000000	-0.6000	2.1650	
CON2	0.000000	-1.1332	2.4343	
COO2	0.000000	-0.8714	2.3950	
RS1	0.000000	-0.3420	2.4343	! Arg side chain; same as C2M
RS2	0.000000	-0.6765	2.1436	! Arg side chain; same as NC bead
RS3	0.000000	-1.0881	2.2635	! Arg side chain
SC	0.000000	-0.8638	2.3606	
CON1	0.000000	-1.0876	2.0609	
GNT	0.000000	-0.6000	2.1650	! same as NCT
GNC	0.000000	-0.6765	2.1436	! same as NC
C2E2	0.000000	-0.3544	2.4146	
CS	0.000000	-0.7001	2.3969	
BZF1	0.000000	-0.3205	2.2332	! from Karteek's benzene model
BZF2	0.000000	-0.3205	2.2332	! from Karteek's benzene model

TOL2	0.000000	-0.5003	2.5318	
TL2F	0.000000	-0.5003	2.5318	
TL2W	0.000000	-0.5003	2.5318	
TL2Y	0.000000	-0.5003	2.5318	
PNC	0.000000	-0.6000	2.1650	! Same as NC2
PS	0.000000	-0.5545	2.6010	! Same as C3M
COH1	0.000000	-0.7152	2.0909	
CCOH	0.000000	-0.7999	2.3741	
TOL4	0.000000	-0.5839	2.6920	
PL1	0.000000	-1.0688	1.8769	
COH3	0.000000	-0.7359	2.0730	
BZY	0.000000	-0.3205	2.2332	! from Karteek's benzene model
C3E2	0.000000	-0.4599	2.7807	
C2E	0.000000	-0.3710	2.4343	! Yaxin's parameters
C2M	0.000000	-0.3420	2.4343	! Yaxin's parameters
C3M	0.000000	-0.5545	2.6010	! Yaxin's parameters
C3E	0.000000	-0.5927	2.6010	! Yaxin's parameters
H2O1	0.000000	-1.1425	2.1170	

NBFIX

! Bead		Emin	Rmin
! Types		(kcal/mol)	(Å)
C2E	H2O1	-0.5130	4.2359
C2M	H2O1	-0.4400	4.2359
C3E	H2O1	-0.6202	4.5000
C3M	H2O1	-0.5434	4.5000

END

D3: NAMD All-Atom Simulation Configuration File Example

```
##############################################################
## ADJUSTABLE PARAMETERS                                    ##
##############################################################

    structure         NAMD_input.psf
    coordinates       NAMD_input.pdb
    outputName        run

    set temperature   298

# Continuing a job from the restart files
if {0} {
set inputname         run
binCoordinates        $inputname.restart.coor
binVelocities         $inputname.restart.vel  ;# remove the "temperature" entry if you use this!
extendedSystem        $inputname.restart.xsc
}

firsttimestep         0

##############################################################
## SIMULATION PARAMETERS                                    ##
##############################################################

# Input
paraTypeCharmm        on
parameters            par_all36m_prot.prm

# NOTE: Do not set the initial velocity temperature if you have also specified a .vel restart file!
temperature           $temperature

# Periodic Boundary conditions
# NOTE: Do not set the periodic cell basis if you have also specified an .xsc restart file!
if {1} {
cellBasisVector1      66.0  0      0
cellBasisVector2       0   66.0   0
cellBasisVector3       0    0     66.0
cellOrigin             0    0      0
}
wrapWater             on
wrapAll               on
```

```
# Force-Field Parameters
exclude              scaled1-4
1-4scaling           1.0
cutoff               12.0
switching            on
switchdist           10.0
pairlistdist         14.0

# Integrator Parameters
timestep             1.0 ;# 1fs/step
nonbondedFreq        1
fullElectFrequency   1
stepspercycle        10

#PME (for full-system periodic electrostatics)
if {1} {
PME                  yes
PMEGridSpacing       1.2
}

# Constant Temperature Control
langevin             on    ;# do langevin dynamics
langevinDamping      1      ;# damping coefficient (gamma) of 5/ps
langevinTemp         $temperature
langevinHydrogen     yes   ;# don't couple langevin bath to hydrogens

# Constant Pressure Control (variable volume)
if {1} {
useGroupPressure     yes ;# needed for 2fs steps
useFlexibleCell      no   ;# no for water box, yes for membrane
useConstantArea      no    ;# no for water box, yes for membrane
langevinPiston           on
langevinPistonTarget   1.01325 ;#  in bar -> 1 atm
langevinPistonPeriod  100.0
langevinPistonDecay   50.0
langevinPistonTemp    $temperature
}

restartfreq          1000    ;# 500steps = every 1ps
dcdfreq              1000
xstFreq              1000
outputEnergies       100
outputPressure       100
```

```
###########################################################
## EXECUTION SCRIPT                                      ##
###########################################################

# Minimization
if {1} {
minimize            100
reinitvels          $temperature
}

run 5000000 ;# 1 million femtoseconds = 1 ns
```