Understanding Small Molecule Conformations using Statistical Machine Learning

Leung Sing Chan

Wolfson College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2020

Declaration

I declare that this thesis is entirely my own work, and except where otherwise stated, describes my own research. This thesis contains fewer than 65,000 words, including appendices, bibliography, footnotes, tables and equations, and has fewer than 150 figures.

Leung Sing Chan
October 2020

Acknowledgements

I would like to thank everyone who has supported my DPhil study. In particular, I would like to give a special mention to my supervisor Professor Garrett M. Morris for his guidance, patience, and enthusiasm. I am always impressed by his insights and his insistence of using Oxford comma, inspiring me throughout my studies and improving readability of my works and this thesis. I am thankful to have him as my supervisor. I would also like to thank my scientific beacon and collaborator, Professor Geoffrey R. Hutchison, for his advice and sharing the computational resources. There are many colleagues and friends helped me during my study, and I would to give a huge thanks to: Leon Law Ho Chung, Jean-Francois Ton, Jin Xu, Edwin Fong, Catherine Wing Ki Wong, Susan Leung, Anne Nierobisch, James Wilsenach, Carlos Outeiral, Fergus Imrie, Fergus Boyles, and other OPIGlets. I am very thankful for the enlightening discussions that we had.

A special thanks to Professor Geoffrey R. Hutchison and Dr. Greg Landrum for the development of Open Babel and RDKit toolkit. Cheminformatics is hard, but Open Babel and RDKit make it easier.

Outside of my academic work, I would like to thank my partner Catherine Wing Ki Wong for her unconditional support, as well as my dear friends: Jacky Lei, Alan Chau, Michael Li, Alex Tsui, Angel Wong, Timothy Wong, Matthew Mak, Richard Tse, and many others. Without them,, my life would have been less interesting.

Finally, and most of all, thanks to my family for their endless supports and encouragement. Without all of them, my experience at Oxford would not have been complete.

Abstract

The generation of conformations for small molecules is one of the cornerstones of computational chemistry. Identifying diverse low energy conformers and, in particular, the lowest energy conformer are essential for applications, such as molecular docking and molecular property predictions. The large conformational space of flexible molecules and the high computational cost of accurate energy evaluation with methods such as quantum mechanics are two key challenges.

This thesis explores the use of statistical techniques to (i) understand the factors governing the conformational preferences of small molecules and their population, and (ii) improve the efficiency in finding the lowest energy conformer of a molecule. I first provide an overview of conformer sampling and Bayesian optimisation, followed by an introduction to circular data analysis for analyzing torsional distribution in small molecule conformations.

I demonstrate the effectiveness of Bayesian optimisation algorithm in finding the lowest energy conformation of molecules, which requires orders of magnitude fewer energy evaluation to find the lowest energy conformation. I also show how sampling efficiency can be further improved by biasing the search towards low energy regions through a knowledge-based acquisition function. To extend the sampling framework, I explore the use of Cremer Pople puckering parameters to characterise complex ring geometries, and study the resulting ring puckering preferences extensively.

Finally, I investigate the factors contributing to the conformational entropies of small molecules, and develop linear models that predict the conformational entropies of small molecules accurately and rapidly.

In summary, this thesis contributes an improved understanding of small molecule conformational preferences, and introduces new methods to improve the efficiency of sampling conformers and entropy calculations.

Contents

List of Figures

List of Tables

Abbreviations

BNN Bayesian Neural Network

BO Bayesian Optimisation

BOA Bayesian Optimisation Algorithm

BOKEI Bayesian Optimisation with Knowledge-based Expected Improvement

CHARMM Chemistry at HARvard MolecularMechanics

CP Cyclic Peptide

CSD Cambridge Structural Database

CTP Cyclic Tetrapeptide

DFT Density Functional Theory

DFTB Density Functional Tight Binding

DG Distance Geometry

ECFP Extended Connectivity Fingerprints

EI Expected Improvement

ETKDG Experimental-Torsion Distance Geometry with basic Knowledge

GA Genetic Algorithm

GAFF General AMBER Force Field

GFN-xTB Geometry Frequency Non-covalent eXtended TB

GP Gaussian Process

HF Hartree–Fock

InChI International Chemical Identifier

KEI Knowledge-based Expected Improvement

KRR Kernel Ridge Regression

LASSO Least Absolute Shrinkage and Selection Operator

LCB Lower Confidence Bound

MACCS Molecular ACCess System

MD Molecular Dynamics

MMFF94 Merck Molecule Force Fields

NN Neural Networks

PDB Protein Data Bank

QM Quantum Mechanics

RBF Radial Basis Function

ReLu Rectified Linear Unit

RMSD Root Mean Square Deviation

RR Rigid Rotors

SBDD Structure-Based Drug Design

SLN SYBYL Line Notation

SMARTS SMILES Arbitrary Target Specification

SMILES Simplified Molecular-Input Line-Entry System

TF Torsion Fingerprints

TFD Torsion Fingerprint Deviation

UFF Universal Force Fields

WLN Wiswesser Line Notation

Chapter 1

Introduction

1.1 Motivation and Contribution

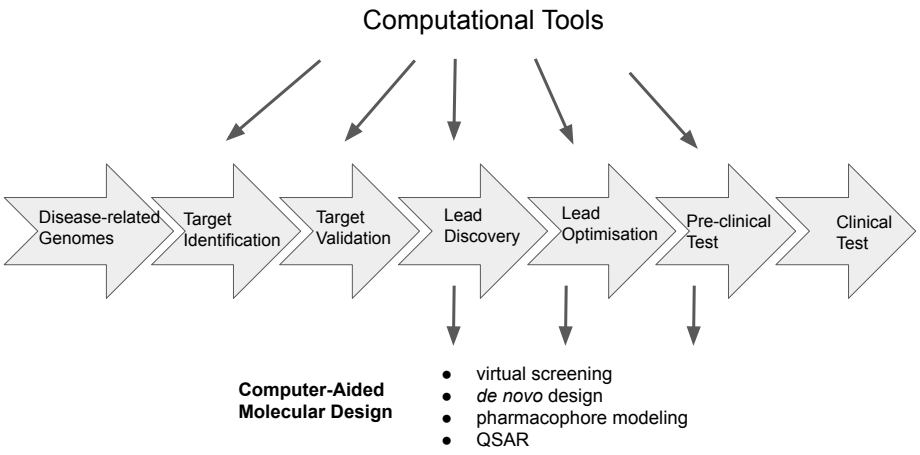


Figure 1.1: Drug discovery pipeline and applications of computer-aided molecular design in the discovery process.

Material and drug discovery process are time consuming and expensive. A new material or drug can cost millions of dollars and can take twenty years or longer to reach the market (Morgan et al., 2011). Recent development in computational tools and the improvement in hardware help accelerate the discovery process, and reduce the overall cost. These computational tools can be applied in different stages, as illustrated in Figure 1.1. For instance, various ligand-based (Ballester and Richards, 2007; Armstrong et al., 2010; Lee et al., 2016) and structure-based virtual screening meth-

ods (Morris et al., 2009; Trott and Olson, 2010; Ragoza et al., 2017) were proposed for lead discovery, and different quantitative structure-activity models (QSARs) (Ma et al., 2015; Wenzel et al., 2019) were used to predict molecular properties.

Computer-aided molecular design methods have played a major role in the lead discovery and optimisation stages over the last decades. Sampling conformers, *i.e.* distinct spatial arrangements of atoms of a molecule typically at a minimum on its potential energy surface (Moss, 1996), is one of the key tasks in the computational discovery pipelines. The goals are to: (i) sample *diverse* low-energy conformers, and (ii) find the *lowest energy* conformation (Hawkins, 2017). Generating distinct low energy conformers is particularly useful in Structure-Based Drug Design (SBDD). It provides inputs for different types of calculation in lead discovery and lead optimization, for example molecular docking (Morris et al., 2009; Trott and Olson, 2010), shape similarity (Ballester and Richards, 2007; Armstrong et al., 2010), pharmacophore searching (Schwab, 2010), and 3D-QSARs (Verma, 2010). On the other hand, finding the lowest energy conformation and other low energy conformations help determine the Boltzmann distribution of conformers, and predict the molecular properties, such as solubility (Hyttinen and Prisle, 2020), standard entropy (Ghahremanpour et al., 2016), and whether it will crystallise (Wicker and Cooper, 2015). Accurate prediction of molecular properties, especially thermochemistry, is crucial for designing chemicals with new functionality.

Most small molecules are flexible and can adopt multiple low-energy conformations. The number of possible conformers increases exponentially with the number of degrees of freedom, in particular, free rotors in the molecule. It is therefore time-consuming to search for relevant conformers in an enormous conformational space. Certain approximations, such as the rigid rotor (RR) hypothesis, in which the bond length and bond angles are kept fixed, have been applied to simplify the problem, and reduce the search space. Still, even in medium-sized molecules, *i.e.* molecules with six or more rotatable bonds, there can be thousands or millions of possible conformations. However, only a small fraction of low-energy conformers are usually thermally accessible, and relevant to the downstream applications. Hence, an efficient sampling scheme is required for the search of all relevant low-energy conformations and the lowest energy conformation.

Directly related problem to effective sampling is evaluating the energies of the candidate conformations. Ideally, quantum mechanics (QM) method (Slater, 1951; Becke, 1988; Lee et al., 1988; Petersson and Al-Laham, 1991; Stephens et al., 1994) should be

used for the energy evaluations and to rank the candidate conformations, however, it is computationally expensive. Different semi-empirical approximations (Dewar et al., 1985; Stewart, 1989; Grimme et al., 2017; Bannwarth et al., 2019), force fields (Rappe et al., 1992; Halgren, 1996; Halgren and Nachbar, 1996; Wang et al., 2004; Brooks et al., 2009) based on molecular mechanics and machine learning potentials (Smith et al., 2017, 2019; Devereux et al., 2020) have been developed to rank the top candidate conformations. These approximations typically improve the run time, but decrease accuracy as a trade-off (Folmsbee and Hutchison, 2020).

To overcome this computational challenge, I present the use of a sampling scheme, namely Bayesian optimisation (Brochu et al., 2010; Snoek et al., 2012), to search for the lowest energy conformation efficiently (Chan et al., 2019, 2020a). This sampling scheme learns the most likely torsion angles in a molecule from previously available energy evaluations, and samples new conformations based on the existing model's uncertainty. This approach balances *exploration* and *exploitation*, to avoid being trapped in local minima of the potential energy surface, and also attempt to find globally optimal solutions. More importantly, it is independent of the choice of energy function, and generally requires orders of magnitude fewer energy evaluations than other search methods to reach the top candidates. It also does not assume any functional forms of the objective function of interest.

Beyond efficient sampling of the low energy conformations, I also develop methods to analyse and model the conformational preferences of small molecules, including correlated acyclic torsions and ring puckerings in cyclic molecules. Inspired by Ramachandran et al. (1963), where the correlation of adjacent acyclic torsion angles in polypeptide chains can be depicted by Ramachandran plots, I extended these ideas to small molecules (Chan et al., 2020a). This helps to provide a better understanding of conformational preferences of small molecules, including the effects of intramolecular interactions such as hydrogen bonds and π-π stacking. Previous work has involved large scale analysis of the torsional preferences of small molecules, and these preferences have been embedded into different sampling schemes (Hawkins et al., 2010; Riniker and Landrum, 2015; Cole et al., 2018; Wang et al., 2020).

Additionally, I conduct an extensive conformational analysis on ring conformations to better understand the geometrical constraints in rings (Chan et al., 2020b). Cyclisation imposes additional constraints on a molecule, and significantly restricts its conformational space. While the conformational preference of small rings, *i.e.* 5- and 6-membered rings, have been widely studied, there are relatively few studies

on medium-sized cyclic molecules and macrocycles, due to their high flexibility and complexity, with the notable exception of cycloalkanes (Bocian et al., 1975; Anet and Cheng, 1975; Pakes et al., 1981; Pawar et al., 1998; Dragojlovic, 2015) and some families of macrocycles (Gutsche and Bauer, 1985; Al-Jallal et al., 2005; El-Azhary and Al-Kahtani, 2005; Gong et al., 2009; Begel et al., 2014). In macrocycles, *i.e.* rings with 12 or more atoms, small local structural changes usually result in notable changes in conformation through transannular repulsion and intramolecular interactions, and the effect sometimes propagates via ring strain to distal structural features (Appavoo et al., 2019). The framework I propose here will provide chemical insights into the factors governing the conformational preferences in rings, which will in turn enhance the computational sampling of ring conformations.

In addition to conformational sampling, it is crucial to determine the conformer population (or distribution of conformers). Boltzmann weighting of conformers is used to determine many molecular properties. In particular, the calculation of conformational entropy can help us understand the stability of a molecule. Again, such calculations can be computationally prohibitive when standard quantum mechanics methods are used. To estimate the conformational entropy for a molecule, S_{conf}, the Boltzmann equation (Equation 1.1) is used:

$$S_{\text{conf}} = k \log(\omega) \tag{1.1}$$

where ω is the number of conformers of a molecule and k is the gas constant. Since the number of conformers is associated with the number of degrees of freedom in a molecule, a simple approximation (Equation 1.2) based on the number of rotatable bonds (N_{rotor}) is commonly used, assuming each rotatable bond correspond to exactly three conformations (Ghahremanpour et al., 2016).

$$\omega \approx 3^{N_{rotor}} \tag{1.2}$$

Hence, the empirical approximation of conformational entropy becomes:

$$S_{\text{conf}} \approx C N_{\text{rotor}} \tag{1.3}$$

where C is a constant.

Since the delocalization of electrons within functional groups and formation of intramolecular interactions such as hydrogen bonds typically reduce the conformational flexibility of a molecule, this approximation will tend to overestimate the conformational entropy. I thus investigate the effects of these factors on the conformer population of a molecule, as well as the resulting conformational entropy. I present a set

of novel descriptors that improve the prediction of the conformational entropies of small molecules.

In brief, the main contributions of the thesis can be summarised as follows:

- Provide a better understanding of the conformational preferences of small molecules.
- Develop a novel sampling framework for low-energy molecular conformations using Bayesian optimisation.
- Investigate the conformational preferences of rings.
- Develop a novel knowledge-based sampling framework for ring conformations.
- Gain chemical insights into the components that contribute to the conformational entropy of a molecule.
- Build a model to predict the conformational entropies of small molecules rapidly and accurately.

1.2 Thesis Outline

This thesis contains six chapters. In Chapter 2, I provide the background material and prior work upon which this work builds. In Chapter 3, I investigate the use of Bayesian optimisation in sampling low-energy molecular conformations. I also propose a new acquisition function that incorporate our prior knowledge about correlated adjacent torsion angles to enhance sampling performance. In Chapter 4, I study the conformational preferences of flexible molecular rings, and discuss the roles of substituents and the effects of intramolecular interactions on ring geometries. In Chapter 5, I investigate the components that contribute to the conformational entropy of small molecules. I also build physically motivated statistical models to predict conformational entropies on a wide range of molecules. Finally, in Chapter 6, I summarise my works and provide a discussion of future work.

Chapter 2

Background

This chapter describes the background and related work on which this thesis builds. Sections 2.1 and 2.2 discuss different representations of molecules and their geometries. Sections 2.3 and 2.4 provide the definition of conformers and a summary of conformer sampling methods. Section 2.5 introduces various metrics to assess the quality of the sampled conformations. Section 2.6 explains different methods for molecular energy evaluation. Section 2.7 and 2.8 provide the basis of kernel methods, Gaussian Process (GP) in machine learning, and an overview of Bayesian optimisation. Section 2.9 provides a brief summary of circular data analysis. Section 2.10 discusses various regression techniques, including linear models and different machine learning models.

2.1 Molecular Representation

2.1.1 String Representations

There are multiple ways to represent a molecule, including a molecular formula, a string or a graph. The simplest string representation is the simplified molecular-input line-entry system (SMILES) (Weininger, 1988). In the SMILES language, there are fundamental types of symbols for *atoms*, *bonds*, charge and stereochemistry. For example, the 2-butene can be represented by "CC=CC". The SMILES strings also encode information about the molecular graph, where the atoms are the nodes and bonds are the edge of the graph. Stereochemistry can be captured using the "@" and "@@" symbols, while *cis* and *trans* double bond isomerism can be described using "\" and "/" symbols.

An extension of SMILES, namely SMILES arbitrary target specification (SMARTS) had been developed. It is a language that allows us to specify substructures using rules based on the extension of SMILES. For example, to search for cyclohexane-containing structures in a database, one could use the SMARTS string [C1CCCCC1]. In addition to atoms and bonds, logical operators (see Table 2.1) can be included, using special atomic and bond symbols. The comprehensive descriptions of SMARTS can be found in Daylight's SMARTS theory (Daylight, 2011).

Table 2.1: SMARTS logical operators; e is an atom or bond SMARTS expression.

Symbol	Expression	Meaning
exclamation mark	!e1	not e1
comma	e1,e2	e1 or e2
semicolon	e1;e2	e1 and e2

Other linear representations, such as Wiswesser line notation (WLN) (Wiswesser, 1954), SYBYL line notation (SLN) (Ash et al., 1997), and the International Chemical Identifier (InChI) (Heller et al., 2013) have also been developed. These unique and unambiguous representations provide a standard way to encode molecular information and facilitate the search of molecule in database. For example, the InChI algorithm converts the input structure into a unique InChI identifier in three steps: (i) normalization, (ii) canonicalization, and (iii) serialization. Redundant information is removed in first step, and unique labels are generated for each atom in second step. The serialization produce a string of characters. The hashed version of InChI representation, also known as InChIKey, is frequently used in my work, to remove duplicate molecules in the datasets.

2.1.2 Molecular Fingerprints

Besides string representations, molecular fingerprints are an alternative way of encoding structural features of a molecule. Typically, a series of binary digits (bits) are used to represent the presence or absence of particular substructures in the molecule. For example, the Molecular ACCess System (MACCS) keys (Dalke, 2014; Landrum, 2018) and the extended connectivity fingerprints (ECFP) (Rogers and Hahn, 2010) are two widely used molecular fingerprints. The MACCS keys are fingerprints representing 166 substructures defined by SMARTS patterns, originally designed for substructure searching with similarity calculation, while ECFP is a circular fingerprint generated

by considering the bonded environment of each atom up to a given radius or diameter measured in bonds. The ECFPs are usually generated using a variant of the Morgan algorithm. Recent advancements in machine learning facilitate the development of molecular fingerprints, and neural fingerprints have been proposed (Duvenaud et al., 2015).

These molecular fingerprints allows us to search for similar molecules by comparing fingerprint similarity, where Tanimoto or Jaccard similarity (Equation 2.1) (Ralaivola et al., 2005) is the most frequently measure:

$$\text{similarity}(A, B) = \frac{A \cap B}{A \cup B} \tag{2.1}$$

where $A \cap B$ represents the common bits shared between two molecules A and B, and $A \cup B$ are the bits set in the fingerprint for molecule A and B. Fingerprints have also been used as molecular features to predict molecular properties (Chan et al., 2020c) and reaction predictions (Schneider et al., 2015).

2.2 Coordinate Systems

There are two common ways for representing positions of atoms in a molecule. The most straightforward approach is to use Cartesian (x, y, z) coordinates. The alternative is to use internal coordinates, *i.e.* bond lengths, bond angles and torsion angles (or dihedral angles). This describes an atom's position relative to other atoms in a molecule. It is always possible to convert internal coordinate to Cartesian coordinates and *vice versa*. Note that there are a total of $3N - 6$ coordinates in internal coordinates, as translation and rotation along the x-, y-, and z-axes do not change the relative positions of the atoms. It is thus preferred for conformational analysis of small molecules.

2.3 Conformers

Conformer is an isomer of a molecule that differs from another form of the same molecule by rotation of one or more single bonds. For example, there are three conformers of a acyclic molecule butane, namely anti-, eclipsed- and gauche- conformations, as illustrated in Figure 2.1. The energy changes during rotation and the Merck molecular force field (MMFF94) (Halgren, 1996; Halgren and Nachbar, 1996)

potential energy of butane is shown in Figure 2.2. The functional form of MMFF94 energy will be discussed in Section 2.6. For cyclic molecules, rotation about a ring bond lead to subsequent changes of torsion angles inside a ring, and results in different conformations. In cyclohexane, it can adopt multiple conformations, including chair, half chair, boat and twist boat conformations, as shown in Figure 2.3. Again, these conformations give different energies, and the chair conformation gives the lowest energy. From these examples, one can expect the number of conformers will increase exponentially with the number of degrees of freedom (rotatable bonds) in a molecule. However, there are other geometrical constraints in larger molecular systems reducing their flexibility and the number of conformers. I will explain it in detail in Chapter 5.

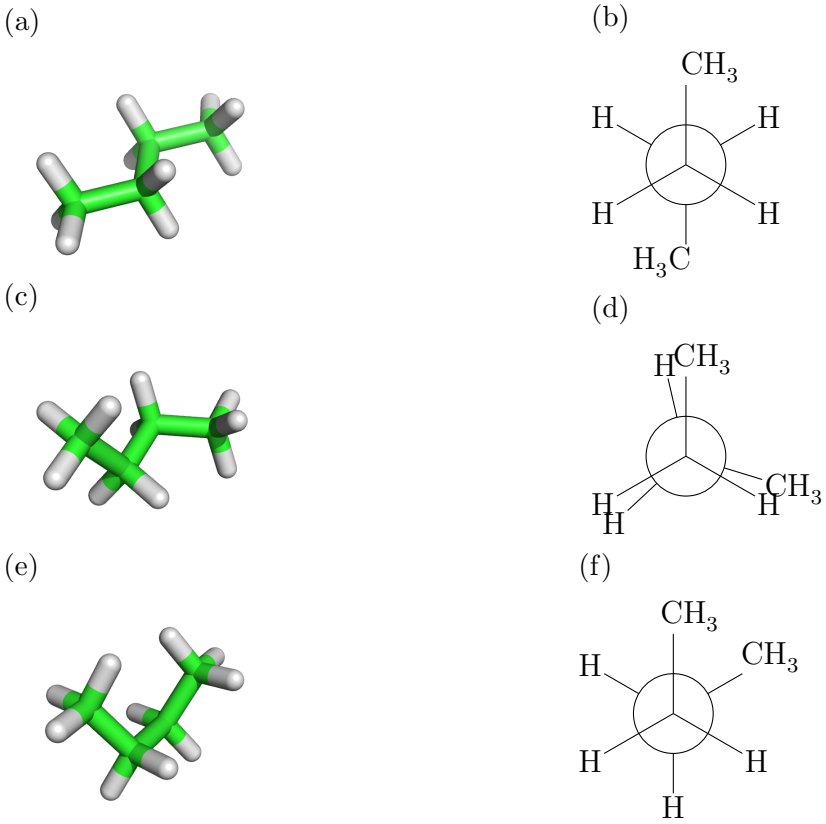


Figure 2.1: Conformations of butane and their corresponding Newman projections: (a) & (b) anti conformation (torsion $\theta = \pm\pi$); (c) & (d) eclipsed conformation (torsion $\theta = \frac{2\pi}{3}$); (e) & (f) gauche conformation (torsion $\theta = \pm\frac{\pi}{3}$).

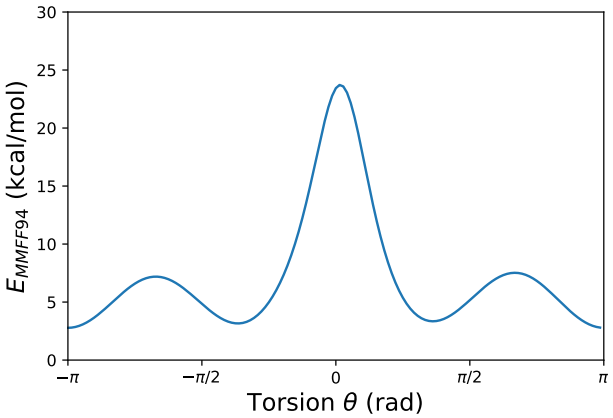


Figure 2.2: MMFF94 potential energy surface of butane.

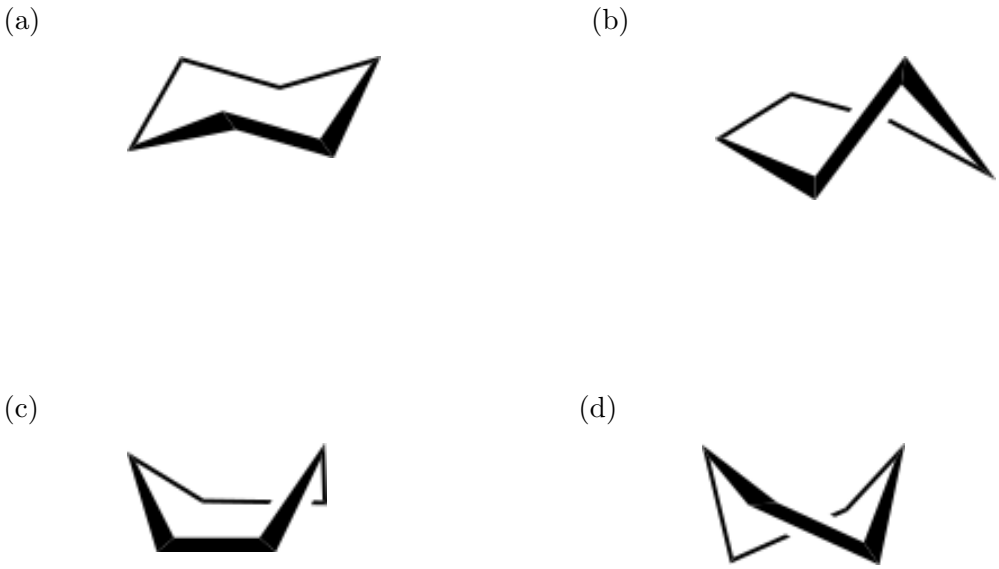


Figure 2.3: Conformations of cyclohexane: (a) chair; (b) half chair; (c) boat; and (d) twist boat.

2.4 Conformer Sampling

Conformer sampling methods can be broadly classified as either systematic or stochastic. Systematic sampling methods typically keep bond lengths and bond angles fixed,

and deterministically enumerates all of the allowed low energy torsion angles for each rotatable bond in a molecule with predefined increment angles, typically 30° or 60°. This brute-force approach is impractical for molecules with a large number of rotatable bonds, because of the combinatorial explosion of candidate conformations. To overcome this issue, various knowledge-based methods have been proposed, for example Confab (O'Boyle et al., 2011b), OMEGA (Hawkins et al., 2010) and CSD Conformer Generator (Cole et al., 2018). They use predefined libraries for torsion angles and ring conformations, and possibly 3D fragment libraries to generate conformers. These libraries are typically created from experimentally determined structures in databases such as the Cambridge Structural Database (CSD) (Groom et al., 2016) or the Protein Data Bank (PDB) (Berman et al., 2000).

On the other hand, various stochastic methods are widely used in the community, including molecular dynamics simulation (MD) (Tsujishita and Hirono, 1997), Monte Carlo-simulated annealing (MC) (Chang et al., 1989; Wilson et al., 1991; Sperandio et al., 2009), distance geometry (DG) (Havel et al., 1985; Spellmeyer et al., 1997; Riniker and Landrum, 2015; Wang et al., 2020) and genetic algorithms (GA) (Mekenyan et al., 1999; Vainio and Johnson, 2007; Brain and Addicoat, 2011; Supady et al., 2015). MD simulation is one of the most complex and time-consuming stochastic methods. It applies Newtonian mechanics to study the evolution of conformations over time, using a molecular mechanics force field to estimate energetics. This approach samples the conformational space of a molecule, and a Boltzmann-weighted conformers ensemble can be obtained if the adequate sampling is achieved. This allows the calculation of different molecular properties of an isolated molecule. The molecular dynamic simulation is also used for conformational sampling of biomolecular systems, such as protein (Minary et al., 2004).

Computationally less expensive methods based on low-mode MD methods (Labute, 2010; CCG, 2018) and MC simulated annealing have also been developed. Low mode MD simulation samples conformers by running a brief MD simulation, with velocities initialized to low-frequency vibrational models. The output conformers are followed by energy minimisation. In simulated annealing, a rejection technique, Metropolis algorithm (Metropolis et al., 1953), is used to sample low-energy conformers, *i.e.* the new conformation generated by random walk is accepted if it satisfies the acceptance criterion, otherwise, it is rejected. The acceptance criterion is defined by the energy difference between the new conformation and current conformation (ΔE), and the annealing temperature, T. The probability that a new sampled conformation with

positive ΔE being accepted is defined as follows:

$$P(\Delta E) = \exp\left(-\frac{\Delta E}{RT}\right) \tag{2.2}$$

where R is the gas constant, and ΔE and T are defined as above.

In distance geometry (DG), a matrix representing the lower and upper bounds of all pairwise distances in a molecule is created. Triangle inequality smoothing of bounds is then applied to refine the matrix. A set of random coordinates that satisfies the bound matrix is generated, followed by energy minimisation to generate a candidate conformer that satisfies these distance constraints. A knowledge-based DG, namely Experimental-Torsion Distance Geometry with basic Knowledge (ETKDG) (Riniker and Landrum, 2015; Wang et al., 2020), has recently been introduced. This method applies an additional minimisation step with knowledge-based torsion potentials, and has shown great promise in reproducing conformations observed in crystal structures. In the ETKDG framework, the torsional potentials take the following forms:

$$V(\theta) = K[1 + \cos(d)\cos(m\theta)] \tag{2.3}$$

$$V(\theta) = \sum_{i=1}^{6} K_i[1 + \cos(d_i)\cos(m_i\theta)] \tag{2.4}$$

where θ is the torsion angle, K is the force constant, d is the phase shift, and m is the multiplicity. For both potentials, the phase shift is restricted to 0 or π. The multiplicity in Equation 2.3 can take values from 1 to 6. Torsion potentials of 479 substructures defined by SMARTS patterns, including 105 aliphatic ring torsions, were fitted. To improve the sampling performance of macrocycles, custom pairwise Coulombic interactions and eccentricity constraints were introduced to bias the sampling towards more experimentally relevant structures. For full description of the method, I refer interested reader to Wang et al. (2020).

Genetic algorithms (Mitchell, 1998) use ideas based on the natural genetics and biological evolution. They typically begin with a random population of random individuals. A new set of individuals is generated by either *mutation* and/or *crossover*. A new generation of individuals is selected according to a fitness criterion to form the next generation. This is an iterative process, and is typically terminated when a maximum number of generations have been produced, or a satisfactory fitness level has been reached for the population. In conformer sampling, the selection is based on either

energy or conformational diversity criteria. Variants of the GA are used to generate conformers in Open Babel (O'Boyle et al., 2011a) and to perform protein-ligand docking in AutoDock (Morris et al., 1998).

For an extensive review of conformer sampling, I refer the interested readers to Ebejer et al. (2012) and Hawkins (2017).

2.5 Metrics for Conformer Sampling Performance

Three metrics are frequently used to assess conformational sampling performance: heavy atom root-mean-square-deviation (RMSD), torsion fingerprint deviation (TFD) (Schulz-Gasch et al., 2012), and the energy difference (ΔE) between sampled conformation and the reference conformation.

The RMSD is defined as follows:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_{i,\text{sample}} - r_{i,\text{ref}})} \tag{2.5}$$

where N is the number of heavy atoms in a molecule, $r_{i,\text{sample}}$ and $r_{i,\text{ref}}$ are the coordinates of the i-th atom of the sampled conformation and the reference conformation respectively. The reference conformation is usually the experimentally determined structure or the lowest energy conformation. In conformational sampling, the RMSD calculation requires alignment of the two conformations, and the RMSD values differ slightly between alignment methods.

The Torsion Fingerprint (TF) of a molecule is a fingerprint encoding the torsion angles of a molecule. The TFD between a molecule in a reference conformation and one of its sampled conformations is calculated as follows:

1. The difference of all torsion angles, including ring torsions, are calculated.
2. The deviation of each torsion angle is normalized to a number between 0 (no deviation) and 1 (maximal deviation).
3. The deviation at topologically central bond or rings are heavily weighted, using a Gaussian function.
4. The TFD is the sum of the weighted deviations.

The TFD values range from 0 to 1, where 0 indicates perfect match and 1 indicates maximal deviation for every single torsion angles and ring torsions. In contrast to RMSD, the TFD is an alignment-free method.

The energy difference ΔE is defined as follows:

$$\Delta E = E_{\text{sample}} - E_{\text{ref}} \tag{2.6}$$

where E_{sample} and E_{ref} are the energy of the sampled conformation and the reference conformation respectively.

2.6 Quantum Mechanical, Semi-Empirical Quantum Mechanical, Molecular Mechanics Methods, and Machine Learning Potentials

Energy evaluations are typically used to rank candidate conformations, and only the top ranked conformers are retained for downstream applications. The energy of a molecule at a particular conformation can be obtained by solving the Schrödinger equation (Equation 2.7) below:

$$\{-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{r})\}\Psi(\mathbf{r}) = E\Psi(\mathbf{r}) \tag{2.7}$$

where $\hbar$ is the Planck's constant divided by 2π; the position of a single particle is given by a vector $\mathbf{r}$, E is the energy of the particle, Ψ is the wave function which characterises the particle's spatial distribution, and ∇^2 is defined as follows:

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{2.8}$$

This equation can be solved exactly for only the simplest systems such as the hydrogen atom. For general molecular systems, approximation is required, and there are different levels of approximation. Quantum mechanical methods (QM) such as Hartree–Fock (HF) methods (Slater, 1951; Petersson and Al-Laham, 1991) and Density Functional Theory (DFT) methods (Becke, 1988; Lee et al., 1988; Becke, 1992; Stephens et al., 1994) provide an accurate estimate of the conformational energies, however, it is computational expensive. Semi-empirical quantum mechanical methods such as Geometry, Frequency, Noncovalent, eXtended Tight Binding (GFN-xTB, GFN for short) (Grimme et al., 2017; Bannwarth et al., 2019), molecular mechanics

methods (MM) such as universal force fields (UFF) (Rappe et al., 1992), Merck molecular force field (MMFF94) (Halgren, 1996; Halgren and Nachbar, 1996), and general AMBER force field (GAFF) (Wang et al., 2004), Chemistry at HARvard Molecular Mechanics (CHARMM) (Brooks et al., 2009), and machine learning potential (Smith et al., 2017, 2019) were also developed to approximate the energies. These approximations improve the run time, and decrease in accuracy as a trade-off (Folmsbee and Hutchison, 2020). I provide a brief summary of these methods below.

2.6.1 *Ab Initio* Quantum Mechanical Methods

The Hartree-Fork (HF) method is a mean-field and wavefunction-based approach to solve the Schrödinger equation. The Born-Oppenheimer approximation (Born and Oppenheimer, 1927) is assumed in the calculation, *i.e.* the motion of atomic nuclei and electrons in a molecule can be treated independently. Hence the total wavefunction for the molecule, Ψ, can be expressed in the following form:

$$\Psi(\text{nuclei}, \text{electrons}) = \Psi(\text{electrons})\Psi(\text{nuclei}) \tag{2.9}$$

The total energy equals the sum of nuclear energy and the electronic energy. A Slater determinant (Slater, 1951) is the functional form of the wavefunction used in Hartree Fork methods. Popular basis functions include 6-31G (Petersson and Al-Laham, 1991).

Similarly, density functional theory (DFT) considers single electron functions. The major difference is that it only attempts to calculate the total electronic energy and the overall electron density distribution. Thanks to the Honhenberg and Kohn theorems (Hohenberg and Kohn, 1964), the ground state energy of the system can be uniquely defined by the electron density, *i.e.* the total energy is a functional of the electron density. The variational principle is applied to minimise the energy with respect to the electron density. A first level approximation, the local density approximation, can be used. This approach is based upon a model called the uniform electron gas, in which the electron density is constant throughout the space. Other methods, including generalized gradient approximations (GGA) (Becke, 1988, 1992) and hybrid functional methods (Lee et al., 1988; Stephens et al., 1994), have also been proposed.

Quantum mechanical methods typically have a high computational cost, with $\mathcal{O}(n^3)$ for DFT methods and $\mathcal{O}(n^4)$ for HF methods, where n is the number of basis functions. Hence, different approximations have been developed.

2.6.2 Semi-Empirical Quantum Mechanical Method

Semi-empirical methods (Thiel, 2014; Christensen et al., 2016) are derived from either HF or DFT theory by applying systematic approximations. Such approximation lead to efficient computational schemes orders of magnitude faster than conventional *ab initio* calculations.

Semi-empirical methods such as AM1 (Dewar et al., 1985) and PM3 (Stewart, 1989) are derived from HF theory and these methods neglect a varying degree of the differential overlap between atomic basis functions. On the other hand, the density functional tight binding approach (DFTB) is derived in the framework of DFT based on a Taylor expansion of the energy with respect to a reference density. Both approaches use minimal basis sets and various approximations of the electron integrals. Methods based on the DFTB approach includes GFN-xtb (Grimme et al., 2017) (GFN for short) and GFN2-xtb (Bannwarth et al., 2019) (GFN2 for short). GFN2 is heavily used in my analysis, so I provide a brief summary of the energy calculation method below.

In GFN2, the total energy of the system is given by:

$$E_{GFN2} = E_{rep} + E_{disp} + E_{EHT} + E_{IES+IXC} + E_{AES} + E_{AXC} + G_{Fermi} \tag{2.10}$$

where E_{rep} is the classical repulsion term; E_{disp} is the pairwise London dispersion contribution; E_{EHT} is the energy term derived by extended Hückel theory; $E_{IES+IXC}$ is the isotropic electrostatic and isotropic exchange correlation; E_{AES} and E_{AXC} refer to the anisotropic electrostatic and anisotropic exchange correlation respectively; the last term, G_{Fermi}, refers to the entropic contribution of an electronic free energy at finite electronic temperature due to Fermi smearing. The inclusion of the anisotropic second order approximation leads to an increase in accuracy, without a noticeable increase in computational cost, compared to its predecessor, GFN (Bannwarth et al., 2019).

2.6.3 Molecular Mechanics

Unlike quantum methods, molecular mechanics methods use classical mechanics to model the molecular system. Force fields are commonly used to estimate the forces between atoms within and between molecules, and the calculation only depends on the atomic positions. This highly simplifying assumption allows us to study larger and more complex molecular systems with many thousands to millions of atoms. Hence it is widely used for screening and filtering large number of organic molecular structures for atomistic properties, for example computational drug design (Kitchen et al., 2004; Sliwoski et al., 2014) and/or conformational search (Kaminský and Jensen, 2007; O'Boyle et al., 2011b).

In molecular mechanics, the potential energy is determined by the interaction energies of bonded atoms and non-bonded atoms respectively. The interactions energies of bonded atoms includes bond stretching, angle bending and torsional terms, while the energies of non-bonded atoms includes electrostatic, and van der Waals (vdW) terms. The parameters in each energy term are typically derived from either experimental or theoretical data. The total energy in molecular mechanics is simply the sum of all the interaction energies, and takes the following general form:

$$\begin{aligned} E_{\text{Total}} &= E_{\text{bonded}} + E_{\text{non-bonded}} \\ &= E_{\text{bond}} + E_{\text{angle}} + E_{\text{torsion}} + E_{electrostatic} + E_{vdW} \end{aligned} \tag{2.11}$$

The parameterisation of each energy term depend on the force fields, and the bond and angle terms are commonly modelled by quadratic or cubic energy functions. For example, the parameterisation of MMFF94 (Halgren, 1996) is shown below:

Bond Stretching

The bond stretching term describes the change in energy when the bond lengths change, and it has the following form:

$$EB_{ij} = -143.9325\frac{kb_{ij}}{2}\Delta r_{ij}^2 \times (1 + cs\Delta r_{ij} + \frac{7}{12}cs^2\Delta r_{ij}^2) \tag{2.12}$$

where Δr_{ij} is the difference between actual and reference bond lengths, kb_{ij} is the force constant, and cs is the stretching constant.

Angle Bending

The angle bending term describes the change in energy as the bond angles change.

$$EA_{ijk} = 0.043844 \frac{ka_{ijk}}{2} \Delta\theta_{ijk}^2 (1 + cb\Delta\theta_{ijk}) \tag{2.13}$$

where ka_{ijk} is the force constant, $\Delta\theta_{ijk}$ is the difference between actual and reference bond angles, and cb is the bending constant. A special set of parameters is used for angles involved in delocalized bonds and/or in small rings.

Torsion Interactions

The torsion interaction term describes the energy change upon internal rotation about single bonds.

$$ET_{ijkl} = 0.5(V_1(1 + \cos\ \Phi) + V_2(1 - \cos\ 2\Phi) + V_3(1 + \cos\ 3\Phi)) \tag{2.14}$$

where V_1, V_2 and V_3 are constants depending on the atom types I, J, K and L for atoms i, j, k, l, where i-j, j-k and k-l are bonded pairs. Φ is torsion angle defined by four atoms ($i,\ j,\ k,\ l$).

Van der Waals Interactions

The van der Waals interaction term provides an estimate of the induced electrical interactions between two or more atoms or molecules that are close to each other. In particular, "Buffered-14-7" form (Halgren, 1992) is used in MMFF94 parameterisation, which describes the 14-th and 7-th power dependencies for repulsive and attractive terms. Alternative functional form for van der Waals interaction term is the Lennard-Jones potential (Jones, 1924), which 12-th and 6-th power terms are used.

$$E_{vdW_{ij}} = \epsilon_{ij} \left(\frac{1.07R_{ij}^*}{R_{ij} + 0.07R_{ij}^*}\right)^7 \left(\frac{1.12R_{ij}^{*7}}{R_{ij}^7 + 0.12R_{ij}^{*7}} - 2\right) \tag{2.15}$$

where R_{ij}^* is the buffering constants, and ϵ_{ij} is used to describe hydrogen bonding interactions.

Electrostatic Interactions

The electrostatic interaction term describes the attractive or repulsive interaction between atoms with electric charges. It takes the following form:

$$EQ_{ij} = \frac{332.0716 q_i q_j}{D(R_{ij} + \delta)^n} \tag{2.16}$$

where q_i, q_j are the partial atomic charges, R_{ij} is the internuclear separation, δ is the electrostatic buffering constant, and D is the dielectric constant. The exponent n is taken as 1 normally.

Two additional terms, stretch-bend interactions and out-of-plane bending at tricoordinate centers are included in MMFF94 energy.

Stretch-Bend Interactions

To model the coupling between bond stretching and angle bending in adjacent bonds, and the associated change in energy, addition stretch-bend interaction is introduced. It takes the following form:

$$EBA_{ijk} = 2.51210(kba_{ijk}\Delta r_{ij} + kba_{ijk}\Delta r_{kj})\Delta\theta_{ijk} \tag{2.17}$$

where kba_{ijk} and kba_{kjl} are force constatns that couple the $i-j$ and $k-j$ stretches to the $i-j-k$ bend, and Δr and $\Delta\theta$ are defined above.

Out-of-Plane Bending at Tricoordinate Centers

The out-of-plane bending term describes the potential of a displacement of the trigonal center atom bonded to three other out-of-plane atoms.

$$EOOP_{ijk;l} = 0.043844\frac{koop_{ijk;l}}{2}\chi^2_{ijk;l} \tag{2.18}$$

where $koop_{ijk;l}$ is the force constant and $\chi_{ijk;l}$ is the Wilson angle between bond $j-l$ and the plane $i-j-k$.

2.6.4 Machine Learning Potential

Recent advances in machine learning and the improvements in general-purpose computing on graphics processing units (GPGPU) and the introduction of machine learning capable silicon chips such as Google's tensor processing units (TPUs) enable us to develop machine learning models to calculate the energies at a smaller trade-off between accuracy and computational cost. Recently, ANAKIN-ME (Accurate NeurAl networK engINe for Molecular Energies) (Smith et al., 2017), or "ANI" for short, and its variants (Smith et al., 2019; Devereux et al., 2020) have been developed. The atom coordinates are taken as input and a set of features describing the atomic environment are generated. The generated features and the molecular energies as computed with

QM (DFT) methods are used to train a fully-connected deep neural network (NN). These deep learning models are able to predict the energies of unseen molecules accurately, and dramatically reduce the computational cost to a semi-empirical or force field level (Folmsbee and Hutchison, 2020).

2.7 Gaussian Process

Gaussian process (GP) (Williams and Rasmussen, 2006) is a Bayesian non-parametric approach that treats f, the function of interest, as a random variable in an infinite dimensional space of functions. It allows us to place a prior on functions, and updating it to a posterior on functions with available observations (evaluations of the function at a set of points). For example, we can treat the function, f, as the potential energy surface of a molecule, and the points are the evaluated energies of a set of conformers. In particular, a Gaussian process can be fully specified by its mean and covariance function, *i.e.*:

$$m(x) = \mathbb{E}[f(x)] \qquad k(x, x') = \mathbb{E}[f(x) - m(x))(f(x') - m(x')]$$

for some observed data x. Let $\mathbf{x} = \{x_i\}_{i=1}^{N}$, and by definition, $\mathbf{f} = [f(x_1), ...f(x_N)]^T$ follows a multivariate normal distribution $\mathcal{N}(\mathbf{m}, \mathbf{K})$, with $\mathbf{m}_i = m(x_i)$, and $\mathbf{K}_{ij} = k(x_i, x_j)$. For simplicity, we can set $m(x)$ as a constant function. Different kernels can be used, including the radial basis function (RBF) kernel(k_{RBF}), Equation 2.19, periodic kernel (k_{PER}), Equation 2.20, Matérn kernel (k_{Matern}), Equation 2.21 and Jaccard (Tanimoto) kernel ($k_{Tanimoto}$) (Ralaivola et al., 2005), Equation 2.22. For a comprehensive review of kernels, I refer the interested readers to Duvenaud (2014).

$$k_{RBF}(x_i, x_j) = \sigma^2 \exp(\frac{-(x_i - x_j)^2}{2l^2}) \tag{2.19}$$

$$k_{PER}(x_i, x_j) = \sigma^2 \exp(\frac{-2\sin^2(\pi(x_i - x_j)/p)}{l^2}) \tag{2.20}$$

$$k_{Matern}(x_i, x_j) = \frac{2^{1-v}}{\Gamma(v)}(\frac{\sqrt{2v}}{l}||x_i - x_j||)^v K_v(\frac{\sqrt{2v}}{l}||x_i - x_j||) \quad l, v > 0 \tag{2.21}$$

$$k_{Tanimoto}(x_i, x_j) = \frac{k_t(x_i, x_j)}{k_t(x_i, x_i) + k_t(x_j, x_j) - k_t(x_i, x_j)} \tag{2.22}$$

where l, p, σ^2 are the length scale, period and the variance respectively; K_v is the modified Bessel function of the second kind of order v; $\Gamma(\cdot)$ is the gamma function; $k_t(x_i, x_j)$ is defined as $\sum_{p \in P} \mathbb{1}(p, x_i)\mathbb{1}(p, x_j)$ and $\mathbb{1}$ is an indicator function. Note that the Tanimoto kernel is simply the Tanimoto similarity between two vectors with binary inputs. These kernels determine the characteristic of the functions, *e.g.* smoothness of the function. Prior knowledge on the distribution family of function f can be incorporated via appropriate kernel functions, *e.g.* Chan et al. (2019, 2020a).

Suppose we have a GP prior on function f, with observed data $\{x_i, y_i\}_{i=1}^N$, and denote $\mathbf{x} = \{x_i\}_{i=1}^N$, $\mathbf{y} = \{y_i\}_{i=1}^N$ $\mathbf{f} = [f(x_1), ..., f(x_N)]$, then we can define the following regression model:

$$\begin{aligned} f &\sim GP(m(\cdot), k(\cdot, \cdot)) \\ \mathbf{f} &\sim \mathcal{N}(\mathbf{m}, \mathbf{K_{xx}}) \\ \mathbf{y}|\mathbf{f} &\sim \mathcal{N}(\mathbf{f}, \sigma^2\mathbf{I}) \end{aligned} \tag{2.23}$$

where $\mathbf{m}$ is the mean vector, $\mathbf{K_{xx}}$ denotes the kernel matrix on inputs $\mathbf{x}$; σ^2 is the variance of the noise, and $\mathbf{I}$ is the identity matrix. Using the standard Gaussian conditioning, the posterior distribution, $\mathbf{f}|\mathbf{y}$, is given below:

$$\mathbf{f}|\mathbf{y} \sim \mathcal{N}(\mu_{post}, \Sigma_{post}) \tag{2.24}$$

where

$$\begin{aligned} \mu_{post} &= \mathbf{m} + \mathbf{K_{xx}}(\mathbf{K_{xx}} + \sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{m}) \\ \Sigma_{post} &= \mathbf{K_{xx}} - \mathbf{K_{xx}}(\mathbf{K_{xx}} + \sigma^2\mathbf{I})^{-1}\mathbf{K_{xx}} \end{aligned}$$

μ_{post} and Σ_{post} are the posterior mean and variance respectively.

In addition, the posterior prediction distribution with unseen observation $\mathbf{x}' = \{x_j\}_{j=1}^m$, and $\mathbf{f}' = [f(x'_1), .., f(x'_m)]$ given below:

$$\mathbf{f}'|\mathbf{y} \sim \mathcal{N}(\mathbf{m}' + \mathbf{K_{x'x}}(\mathbf{K_{xx}} + \sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{m}), \mathbf{K_{x'x}} - \mathbf{K_{x'x}}(\mathbf{K_{xx}} + \sigma^2\mathbf{I})^{-1}\mathbf{K_{xx'}}) \tag{2.25}$$

where $\mathbf{m}'$ denotes the mean function evaluation on $\mathbf{x}'$, $\mathbf{K_{x'x}}$ denotes the kernel matrix between $\mathbf{x}'$ and $\mathbf{x}$ and *vice versa* for $\mathbf{K_{xx'}}$. It should be noted that Gaussian process incurs a computational cost of $\mathcal{O}(n^3)$, where n is the number of observations, due to the matrix inversion. Two approximation methods, namely Nyström approximations (Williams and Seeger, 2001) and random Fourier features (Rahimi and Recht, 2008), are frequently used.

The Nyström approximation is a low-rank approximation to an $n \times n$ Gram matrix, *i.e.* the kernel matrix. This approximation is commonly used for the numerical solutions of eigenproblems. It is achieved by carrying out an eigendecomposition on a smaller system of size $m < n$, and then expanding the results back up to n dimensions. The computational complexity is reduced to $\mathcal{O}(m^2 n)$. Random sampling or weighted random sub-sampling methods can be used to construct a smaller $(m \times m)$ matrix. Note that this approach is data-dependent, and it yields a good approximation for kernel matrix with rapidly decaying eigenvalues.

In contrast, random Fourier features is a data-independent approximation. Instead of matrix approximation, it attempts to approximate the feature maps with random Fourier bases $\cos(w^T x + b)$, where $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are random variables, and $x \in \mathbb{R}^d$ is observed data point. These mappings project data points on to a randomly chosen line, and then pass the resulting scalar through a sinusoidal function. Using classical theorems from harmonic analysis, the kernel $k(x_i, x_j)$ can be approximated by the inner product of the approximated feature maps $z(\cdot)$, that is:

$$k(x_i, x_j) \approx z(x_i)^T z(x_j) \tag{2.26}$$

Note that this approximation is widely applied to shift-invariant kernels such as RBF kernel.

2.8 Bayesian Optimisation

In Bayesian optimisation (BO), we have a function, $f(\theta)$, that we want to optimise with respect to $\theta \in \Theta$, *i.e.* the goal is to find $\theta^* = \text{argmin}_{\theta \in \Theta} f(\theta)$. For example, we might want to determine accurately the Boltzmann weights of conformers of a flexible molecule, so we would have to search for the lowest energy conformation. Bayesian optimisation can be applied in this scenario, with the function, f, as the potential energy surface of a molecule of interest, and parameters, θ, are the torsion angles in a molecule. In the BO setting, the function, f, can be non-differentiable and non-convex with respect to θ. The only requirement is that we can evaluate the function pointwise. While conventional optimisation methods such as basin hopping (Wales and Doye, 1997) and evolutionary algorithms (Mitchell, 1998; Hansen, 2006) can solve the problem, the evaluation of f is often computationally expensive, *e.g.* energy evaluation with standard quantum mechanics methods. Hence, we are interested in

obtaining the set of torsion angles around rotatable bonds that give the lowest energy conformations, θ^*, in as few evaluations as possible.

Typically, as function, f, is smoothly varying with respect to θ. By considering previous evaluations of f, one can employ a sequential strategy to select θ. In particular, we can model the function f using a Gaussian Process (GP). The general procedure of Bayesian optimisation is described in Algorithm 1 (see Figure 2.4).

Algorithm 1 Bayesian Optimization

- **Input**: Data $\mathcal{D}_{t-1}$

1. Choose θ_t by optimizing the acquisition function, a, over the Gaussian Process (GP) such that:
$$\theta_t = \text{argmax}_\theta \; \alpha(\theta|\mathcal{D}_{t-1}) \tag{2.27}$$
2. Sample the objective function: $y_t = f(\theta_t) + \epsilon_t$
3. Augment the data $\mathcal{D}_t = \{\mathcal{D}_{1:t-1} \cup (\theta_t, y_t)\}$
4. Repeat until the maximum number of iterations is reached.

The acquisition function takes the predicted values and model uncertainty into account to strike a balance between *exploration* and *exploitation*. Here, exploration means seeking locations with high posterior variance, while exploitation means seeking locations with low posterior mean. There are different acquisition functions, including Probability of Improvement (PI), Expected Improvement (EI) (Mockus et al., 1978) and Gaussian Process Lower Confidence Bound (GP-LCB) (Srinivas et al., 2009):

$$\text{PI}(\theta) = \Phi(z(\theta)) \tag{2.28}$$

$$\text{EI}(\theta) = \sigma(\theta)(z(\theta)\Phi(z(\theta)) + \phi(z(\theta)) \tag{2.29}$$

$$\text{GP-LCB}(\theta) = \mu(\theta) - q_\alpha \sigma(\theta) \tag{2.30}$$

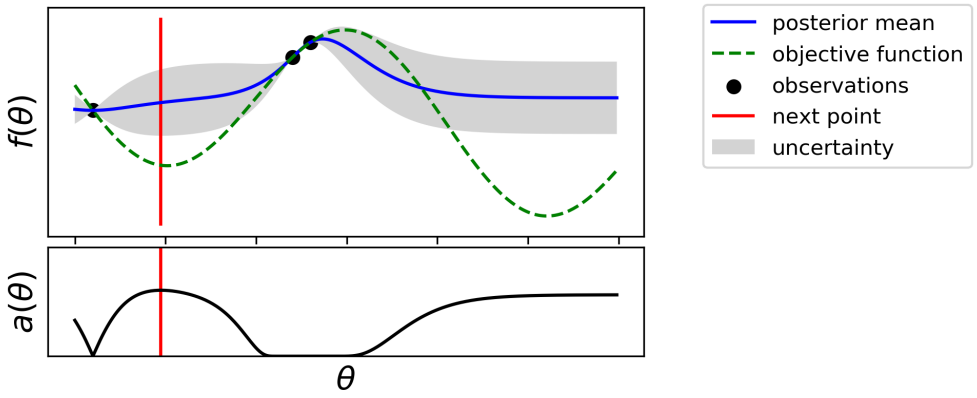
where $z(\theta) = \frac{f(\theta_{best}) - \mu(\theta)}{\sigma(\theta)}$, $\mu(\theta)$ and $\sigma^2(\theta)$ are the predictive mean and predictive variance respectively; $\Phi(\cdot)$, $\phi(\cdot)$, q_α are the cumulative distribution function, probability density function and the desired quantile of the standard normal distribution.

Note that the size of the improvement in the objective function is not taken into account in PI, but it is considered in EI. Hence exploitation is preferred in PI: locations that have a high probability of being better than the current best are drawn. It may offer larger improvement, but less certainty.

In addition to Gaussian Process, other surrogates such as Bayesian linear regression (Law et al., 2019) and Bayesian neural networks (BNN) (Snoek et al., 2015) are

also used. For a comprehensive review, we refer the interested readers to Brochu et al. (2010) and Snoek et al. (2012).

(a)



(b)

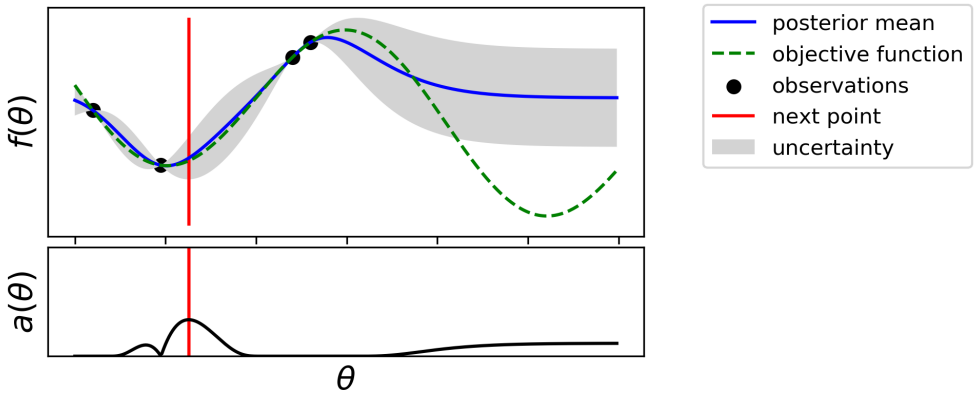


Figure 2.4: Illustration of Bayesian optimisation (without noise). The green dotted line represents the underlying unknown function, f, while the blue line refers to the Gaussian Process model (with uncertainty represented by the shaded grey region). The black line represents the acquisition function, $a(\theta)$. The red line indicates the next query point, *i.e.* θ that maximises the acquisition function $a(\theta)$. Bayesian optimisation at (a) time t; (b) at time $t+1$.

2.9 Statistical Analysis of Circular Data

Internal coordinates, *i.e.* bond lengths, bond angles, and torsion angles, are frequently used in conformational analysis of small molecules (McCabe et al., 2014) and macromolecules, *e.g.* proteins (Ramachandran et al., 1963; Mardia et al., 2007). Conventional statistical analysis, however, cannot be applied directly, due to the nature

of circular variables, such as bond angles and torsion angles. The summary statistics such as mean and standard deviation are not well defined, as they depend on the point where the circle is cut. To address these issues, I will describe the basic concepts of circular data analysis below. For a comprehensive review, I refer the interested reader to Jammalamadaka and Sengupta (2001) and Mardia and Jupp (2009).

2.9.1 Preliminaries and Notation

Each point x on a circle can be represented by an angle θ (in radians), *i.e.*,

$$x = (\cos\theta, \sin\theta) \tag{2.31}$$

2.9.2 Measures of Location, Concentration and Dispersion

For given unit vectors x_1,..., x_n with corresponding angles θ_i, for $i = 1, ..., n$. The mean direction $\bar{\theta}$ of θ_1,.., θ_n is the direction of resultant $x_1 + ... + x_n$ of $x_1, ..., x_n$. The Cartesian coordinates of x_j are $(\cos\theta_j, \sin\theta_j)$, for $j = 1, ..., n$, and the center of mass are $(\bar{C}, \bar{S})$:

$$\bar{C} = \frac{1}{n}\sum_{j=1}^{n}\cos\theta_j, \qquad \bar{S} = \frac{1}{n}\sum_{j=1}^{n}\sin\theta_j \tag{2.32}$$

Therefore $\bar{\theta}$ is the solution of the equations:

$$\begin{aligned} \bar{C} &= \bar{R}\cos\bar{\theta} \\ \bar{S} &= \bar{R}\sin\bar{\theta} \end{aligned} \tag{2.33}$$

provided that $\bar{R} > 0$, and $\bar{R} = |\sqrt{(\bar{C}^2 + \bar{S}^2)}|$ is the mean resultant length. The $\bar{\theta}$ can be obtained by:

$$\bar{\theta} = \begin{cases} \tan^{-1}(\frac{\bar{S}}{\bar{C}}) & \text{if } \bar{C} \geq 0 \\ \tan^{-1}(\frac{\bar{S}}{\bar{C}}) + \pi & \text{if } \bar{C} < 0 \end{cases} \tag{2.34}$$

Since $x_1, ..., x_n$ are unit vectors, the resultant length $\bar{R}$ ranges from 0 to 1, where 1 indicates the directions $\theta_1, ..., \theta_n$ are tightly clustered, while 0 indicates the directions widely dispersed. The resultant length $\bar{R}$ is therefore a measure of concentration of a data set, and the sample circular variance is defined as :

$$V = 1 - \bar{R} \tag{2.35}$$

Alternatively, circular variance can be $2(1-\bar{R})$. It is sometimes useful to have an analogue for circular data of the standard deviation of data on the line. The calculation requires transformation of the summary statistics, and is given by:

$$v = |\sqrt{-2\log(1-V)}| = |\sqrt{-2\log\bar{R}}| \tag{2.36}$$

Note that v takes values in positive real, whereas V takes values in [0,1]

In addition to location and concentration, a measure of the distance between two angles θ^a and θ^b is required. Two circular distances are introduced below:

$$\begin{aligned} d(\theta^a, \theta^b) &= \min(\theta^a - \theta^b, 2\pi - (\theta^a - \theta^b)) \\ &= \pi - |\pi - |\theta^a - \theta^b|| \end{aligned} \tag{2.37}$$

$$d(\theta^a, \theta^b) = 1 - \cos(\theta^a - \theta^b) \tag{2.38}$$

The distance defined in Equation 2.37 is simply the smaller of the two arc-lengths between two points along the circumference. Equation 2.38 is a monotone increasing function of difference between two angles. It takes value 0 when the difference is 0, while it is 2 when the angular difference is π.

2.9.3 Circular Correlation

We are often interested in measuring the association between two circular variables, similar to that in linear analysis. Following the definition in (Fisher and Lee, 1983), the circular correlation coefficient $\rho_c(\theta^a, \theta^b)$ is defined as:

$$\rho_c(\theta^a, \theta^b) = \frac{\mathbb{E}((\sin(\theta^a - \mu)(\sin(\theta^b - \nu))}{|\sqrt{\mathrm{Var}(\sin(\theta^a - \mu))\mathrm{Var}(\sin(\theta^b - \nu))}|} \tag{2.39}$$

where μ and ν are the mean directions of θ^a and θ^b respectively.

This circular correlation coefficient, ρ_c satisfies the following properties:

- it does not depend on the zero direction used for either variable;
- it is symmetric, *i.e.* $\rho_c(\theta^a, \theta^b) = \rho_c(\theta^b, \theta^c)$
- $|\rho_c(\theta^a, \theta^b)| \leq 1$
- $\rho_c(\theta^a, \theta^b) = 0$ if θ^a, θ^b are independent, but the converse does not always hold;
- $\rho_c(\theta^a, \theta^b) = 1$ if and only if (iff) $\theta^a = \theta^b + \text{const} \pmod{2\pi}$, and $\rho_c(\theta^a, \theta^b) = -1$ iff $\theta^a + \theta^b = \text{const} \pmod{2\pi}$

2.9.4 von Mises Distribution

Similar to conventional statistics, we can build various models for circular data. Here, I discuss some important families of of distributions on circles. The most basic distribution on the circle is the uniform distribution. The von Mises distribution is another important distribution for circular data. It is an analogue to the normal distributions on a line. Other useful distributions include wrapped normal, wrapped Cauchy and projected normal distributions. I only provide a brief summary of von Mises distribution below. For a review of other distributions, I refer the interested readers to Mardia and Jupp (2009).

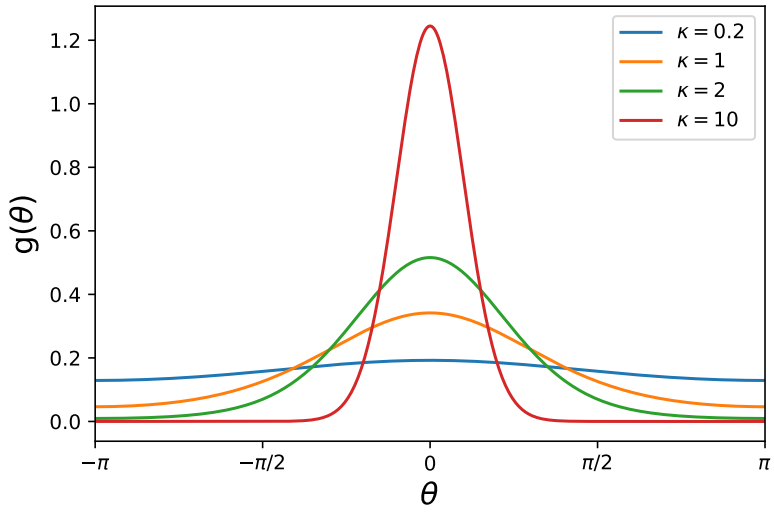


Figure 2.5: von Mises distribution with varying κ: 0.2 (blue), 1 (orange), 2 (green) and 10 (red). The mean direction is zero, *i.e.* $\mu = 0$.

The von Mises distribution $\mathcal{M}(\mu, \kappa)$ has the probability density function:

$$g(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp^{\kappa \cos(\theta - \mu)} \tag{2.40}$$

where I_0 denotes the modified Bessel function of the first kind and order 0. The Bessel function I_0 has power series expansion:

$$I_0(\kappa) = \sum_{r=0}^{\infty} \frac{1}{(r!)^2} (\frac{\kappa}{2})^{2r} \tag{2.41}$$

The parameter μ is the mean direction and the parameter κ is known as the concentration parameter. Typically, we take $\kappa \geq 0$. When $\kappa = 0$, $M(\mu, \kappa)$ is simply the

uniform distribution. For high concentration κ, it can be approximated by normal distribution.

The von Mises distribution is unimodel and is symmetrical about $\theta = \mu$, as illustrated in Figure 2.5. The mode is at $\theta = \mu$. and the antimode is at $\theta = \mu + \pi$. Since von Mises distribution belongs to the exponential family of distributions, its properties can be derived in a similar fashion to other exponential family distributions.

There are multiple ways in which the von Mises distribution can arise. I provide a brief summary of one of the approaches, which is the conditional distribution of bivariate normal distribution. I refer the interested readers to Mardia and Jupp (2009) for a comprehensive review.

Conditioning Normal Distribution

Let x have a bivariate normal distribution with mean $\mu = (\cos\mu, \sin\mu)^T$ and variance matrix $\kappa\mathbf{I}_2$, where $\mathbf{I_2}$ is a 2×2 identity matrix. Set $x = r(\cos\mu, \sin\mu)^T$. The probability function of (r, θ) becomes:

$$p(r, \theta) \propto r \exp^{-\frac{\kappa}{2}(r^2 - 2r\cos(\theta - \mu))} \tag{2.42}$$

and the conditional distribution of $\theta | r = 1$ is $M(\mu, \kappa)$.

2.9.5 Mixture Models

Often the data exhibits multimodality, and a single probability distribution is not sufficient to describe the data. The classical statistics technique, mixture models, can be applied to address this.

Mixture models assume that the dataset was created by sampling independently and identically from K distinct populations (called mixture components). In other words, data come from a mixture of several sources and the model for the data can be viewed as a combination of several distinct probability distributions, often modelled with a given parametric family, g, *e.g.* von Mises distribution:

$$g_M = \sum_{i=1}^{K} \omega_i g_i(\theta | \mu_i, \kappa_i) \tag{2.43}$$

where K is the number of components, g_i denote a von Mises distribution with mean μ_i and concentration κ_i, and ω_i is the weight of each component, with $\sum_i \omega_i = 1$.

To get the estimates of the parameter, μ_i, κ_i, and weights ω_i, a local optimisation technique called Expectation Maximisation (EM) (Dempster et al., 1977) can be used.

EM algorithm is a general iterative method for local maximisation of the likelihood under missing data or hidden variables. I provide a brief explanation of the EM algorithm below.

Let (x, z) be a pair of observed variables x, and latent variables z. The probabilistic model is given by $p(x, z|\Gamma)$, for some parameters, Γ, and we do not have access to z. Direct optimisation of the observed data log-likelihood (marginal log-likelihood) $l(\Gamma) = \log p(x|\Gamma) = \log \int p(x, z|\Gamma)dz$ over parameter x, is typically not feasible due to presense of many local optima. The EM algorithm seeks to find the estimates of the marginal likelihood by alternating maximisation. It first updates the distribution of latent variable $q(z)$ (E-step), with full conditional of z using the current estimates of Γ, followed by updates of parameters Γ, by maximising the expected value of the log-likelihood function (see Algorithm 2).

Algorithm 2 Expectation Maximisation (EM)

At time $t > 0$, given Γ^t

- E-step: set $q^t = p(z|x, \Gamma^{(t)})$ and compute the expected value of the log-likelihood functio $E_{z\sim q^t} \log p(x, z|\Gamma^t)$
- M-step: $\Gamma^{t+1} = \text{argmax}_\Gamma \; E_{z\sim q^t} \log p(x, z|\Gamma)$

2.10 Linear Models and Machine Learning Models

Often we are interested in modelling the relationships between a dependent variable (molecular property) and one or more independent variables (molecular features). Linear models are commonly used. In particular, I discuss the case where the dependent variable (response) is continuous and real-valued. I refer interested readers to (Nelder and Wedderburn, 1972) for the general modelling framework with different response variable types.

2.10.1 Linear Models

Given data $(\mathbf{x}_i, y_i)_{i=1}^n$, where $\mathbf{x}_i$ is a p dimensional vector with all independent variables, and y_i is dependent variable. The linear model takes the following form:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i \tag{2.44}$$

where $\beta_0, ..., \beta_p$ are the model parameters and ϵ (error term) is random variables normally distributed with mean 0, and with constant variance σ^2. Alternatively, it can be written in matrix form.

$$Y = X\beta + \epsilon \tag{2.45}$$

where Y is the response vector $(y_1, ..., y_n)^T$; X is the data matrix $(X_1, ..., X_p)$; β is the parameter vector; and ϵ is defined above.

To estimate the β, least-squares method is commonly used, *i.e.* minimising the squared error between Y and $X\beta$. The resulted estimator of $\hat{\beta}$ take the following form:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{2.46}$$

In standard linear regression, a number of assumptions are made: (i) linearity which the dependent variable is a linear combination of the parameters (regression coefficient, β) and the independent variables; (ii) homoscedasticity which means constant variance of the error terms; (iii) independence of errors, which the error terms are uncorrelated with each other; and (iv) normality of the error terms. If any of the assumption fails to hold, the resulting model may not be useful and the results in hypothesis tests are not valid.

Note that one can always increase the number of independent variables to achieve lower error and match the observations, however, the complex model may not generalise to new samples. To prevent overfitting, *regularisation* is commonly used. It can be achieved by adding penalties to the large values of parameters to the risk function, *i.e.* the expected loss. Hence the objective is as follows:

$$\min_\beta \frac{1}{n} \sum_{i=1}^{n} L(y_i, f_\beta(\mathbf{x_i})) + \lambda ||\beta||_p^p \tag{2.47}$$

where $p \geq 1$; $L(y_i, f_\beta(\mathbf{x_i})$ is the loss function; f_β is a model with parameter β; $|| \cdot ||_p^p$ is the L_p norm; λ is a tunable penalty term; y and $\mathbf{x}$ are defined as above.

2.10.2 Least Absolute Shrinkage and Selection Operator

The least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) is a regression method with regularization, and it takes $p = 1$ in Equation 2.47. The

optimal value of penalty λ can be selected by cross validation over a parameter grid. For given penalty λ, the solution of β can be obtained using least angle regression (LARS) (Efron et al., 2004), and coordinate descent (Tibshirani and Taylor, 2011).

2.10.3 Ridge Regression and Kernel Ridge Regression

The ridge regression, which sets $p = 2$ in Equation 2.47, is another form of regularisation. Similarly, the optimal penalty λ is selected by cross validation over a parameter grid. The solution to β under squared loss function is as follows:

$$\hat{\beta} = (X^T X + \lambda \mathbf{I})^{-1} X^T Y \tag{2.48}$$

where I is an identity matrix.

Often non-linear relationship is observed between dependent variable and independent variables, and the data vector $\mathbf{x} = (x_1, ...x_p)$ is transformed by some function Φ, *i.e.* $\mathbf{x} \mapsto \Phi = \Phi(\mathbf{x})$. I denote the transformed matrix $\Phi(X)$ as Φ_X below. The solution to β under quadratic loss becomes:

$$\hat{\beta} = (\Phi_X^T \Phi_X + \lambda \mathbf{I})^{-1} \Phi_X^T Y \tag{2.49}$$

By using the matrix identity below:

$$(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = P B^T (B P B^T + R)^{-1} \tag{2.50}$$

for some matrix P, B and R.

The new predicted value of a new test point, $\mathbf{x}'$ is as follows:

$$\begin{aligned} y' &= (\Phi_{\mathbf{x}'} \Phi_X^T)(\Phi_X \Phi_X^T + \lambda \mathbf{I})^{-1} Y \\ &= \kappa(\mathbf{x}', X)(\mathbf{K}_{XX} + \lambda \mathbf{I})^{-1} Y \end{aligned} \tag{2.51}$$

where $\kappa(\mathbf{x}', X) = \Phi(\mathbf{x}')\Phi_X^T$ and $\mathbf{K}_{XX} = \Phi_X \Phi_X^T$. It shows that only the kernel (inner product between data points) is required in the calculation, instead of the explicit feature mapping. It is so called the *kernel trick*.

The kernel ridge regression is also closely connected to the Gaussian Process introduced before (Kanagawa et al., 2018). For a comprehensive review, I refer interested readers to Hofmann et al. (2008).

2.10.4 Neural Network

Recent advancement in technology enable the use of deep neural network to model complex biological systems and predict chemical properties accurately. For example, Ragoza et al. (2017) developed a convolutional neural network based scoring function to rank and predict protein-ligand binding affinities and poses. Imrie et al. (2020) developed a graph-based deep generative model that utilise structural knowledge to generate molecular linkers between two fragments. Ma et al. (2015) and Wenzel et al. (2019) also developed different neural network models to predict molecular properties. The flexible modelling framework allows it to learn the non-linear relatinoship effectively. Here, I provide a basic idea behind neural network. For a detailed introduction to neural networks, I refer interested readers to Goodfellow et al. (2016).

The typical artificial neural networks (ANN) are computing systems inspired by the biological neural networks in human brain. A neural network (NN) consists of many simple and connected nodes called neurons. The input and output nodes are connected by a set of weight links, see illustration in Figure 2.6. The nodes can be connected in various ways, and give rise to different network architectures. For instance, there are no cycles in the feed-forward networks, which allow a direct feedback. In contrast, the formation of cycles between nodes in recurrent networks creates feedback loops, and thus allows the modelling of dynamical changes over time (Salehinejad et al., 2017).

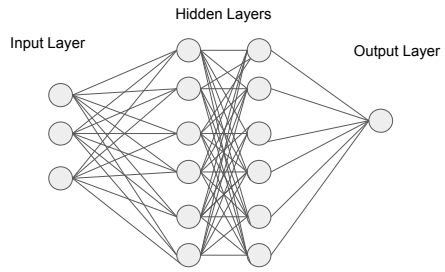


Figure 2.6: Schematic of a two hidden layers, feed-forward neural network.

In addition to network architecture, the activation function plays an important role in determining the non-linear characteristics of the function. It first maps weighted

inputs to feature space through one or more mapping(s), ϕ(s), which can be linear or non-linear. At the final layer (output layer), another transformation is used to map the transformed feature vectors to the outputs. Hence the multi-layer neural networks output, y, can be written as composite transformations of inputs, x, *i.e.*

$$y = \phi^{M+1}(\phi^{M}(...\phi^{(2)}(\phi^{(1)}(x)))))) \tag{2.52}$$

where M is the number of hidden layers, and $\phi^{(l)}(x) = \phi(w_l^T x + b_l)$, where y is the b_l denotes a bias term and w_l denotes the weight.

Commonly used activation functions include rectified linear unit (ReLu) (Nair and Hinton, 2010), leaky rectified linear unit (Leaky ReLU) (Maas et al., 2013), tanh, sigmoid, and radial basis function.

The solution to the weight, w, is typically calculated by backpropagation (Rumelhart et al., 1986a,b). It is an algorithm calculating the gradient of the loss function with respect to the weights by chain rule. The gradient of one layer is computed at a time, and is iterated backward from the last layer. Gradient descent methods, for example stochastic gradient descent, are frequently used. For an overview of gradient descent methods, I refer interested readers to Ruder (2016).

Typically, the multi-layer networks suffer from overfitting due to its high complexity. To overcome this issue, different techniques are proposed, including L_1 and L_2 regularization, dropout (Srivastava et al., 2014), early stopping and data augmentation.

The L_1 and L_2 regularizations work identically as above. In dropout, some nodes along with their connections from neural networks are randomly dropped to prevent it from co-adapting. This reduces overfitting and improve the generalising ability significantly. In early stopping, a hold-out validation set is used to assess the performance of the model. The training is stopped when the performance deteriorates. Lastly, the data augmentation inflates the training data via data warping or oversampling. Geometric and color transformations are typically used in data warping, while the label is preserved. Oversampling creates synthetic instances into the training sets. The data augmentation technique is frequently used in image problem (Shorten and Khoshgoftaar, 2019).

Chapter 3

Bayesian Optimisation for Conformer Generation

Most of the work in this chapter has been reproduced from the following publications:
(i) L. Chan, G. R. Hutchison, and G. M. Morris. Bayesian Optimization for Conformer Generation. *Journal of Cheminformatics* (2019) 11 32 and
(ii) L. Chan, G. R. Hutchison, and G. M. Morris. BOKEI: Bayesian Optimization Using Knowledge of Correlated Torsions and Expected Improvement for Conformer Generation. *Physical Chemistry Chemical Physics* (2020) 22 5211-5219

3.1 Background

Most small molecules are flexible and can adopt multiple energetically-accessible conformations. Even in medium-sized molecules, *e.g.* molecules with six or more rotatable bonds, there may be thousands or millions of possibilities. The multi-dimensional energy landscape and presence of huge number of local minima make finding low-energy conformations to be one of the key challenges in molecular modelling and cheminformatics.

There are two goals in conformer sampling: (i) generate geometrically diverse conformations and (ii) search for the lowest energy and other low energy conformations. Generating diverse conformations is important to many applications including molecular docking (Morris et al., 2009; Trott and Olson, 2010), pharmacophore modeling (Schwab, 2010), and generate 3D quantitative structure-activity relationships (3D-QSAR) (Verma, 2010). A variety of tools have been developed with the purpose

of generating diverse conformers. For a comprehensive review, I refer the interested readers to Ebejer et al. (2012) and Hawkins (2017).

On the other hand, finding low energy and the lowest energy conformations efficiently help determine the molecular physiochemical properties, such as solubility (Hyttinen and Prisle, 2020) and standard entropy (Ghahremanpour et al., 2016). To accurately predict the molecular properties, standard quantum mechanics methods is required to evaluate the conformer energies. However, it is computational prohibitive for flexible molecules. To overcome this challenge, I propose to use Bayesian optimisation algorithm (BOA) (Chan et al., 2019), as introduced in Section 2.8, to search for the lowest energy conformation efficiently. This technique learns the most likely torsion angles for an arbitrary molecule by sampling new conformers from the multi-dimensional potential energy surface, regardless of the energy function used.

In addition, structural information about the correlated torsion is crucial for conformer generation, as adjacent torsion angles are naturally constrained, in order to reduce steric clashes, retain π-conjugation, align intramolecular hydrogen bonds or other similar non-covalent interactions (Chan et al., 2020a). For instance, Figure 3.1b shows the computed Merck molecular force fields (MMFF94) (Halgren, 1996; Halgren and Nachbar, 1996) potential energy surface for 5-phenylthioquinazoline-2,4-diamine, with light blue indicating the low-energy feasible regions. Due to steric clashes, the neighboring dihedral angles are clearly correlated and thus the conformational search can be greatly reduced by incorporating this information. Even in a simpler molecule such as *ortho*-1,1':2',1"-terphenyl, Figure 3.1d indicates correlation between non-neighboring dihedral angles due to steric clashes. Clearly, it is necessary to understand the underlying correlation between adjacent torsion angles and possibly the non-nearest neighbors correlation to improve the sampling efficiency.

In this chapter, I first examine my new sampling approach based on Bayesian optimisation to search for the lowest energy conformation. In the second part, I study the distributions of correlated torsions in X-ray crystal structures, the lowest energy conformations simulated under force fields MMFF94 and a semi-empirical energy function, GFN2 (Bannwarth et al., 2019), separately. I will also develop a modified acquisition function that encapsulate the correlated torsions preferences to enhance the sampling performance (Chan et al., 2020a).

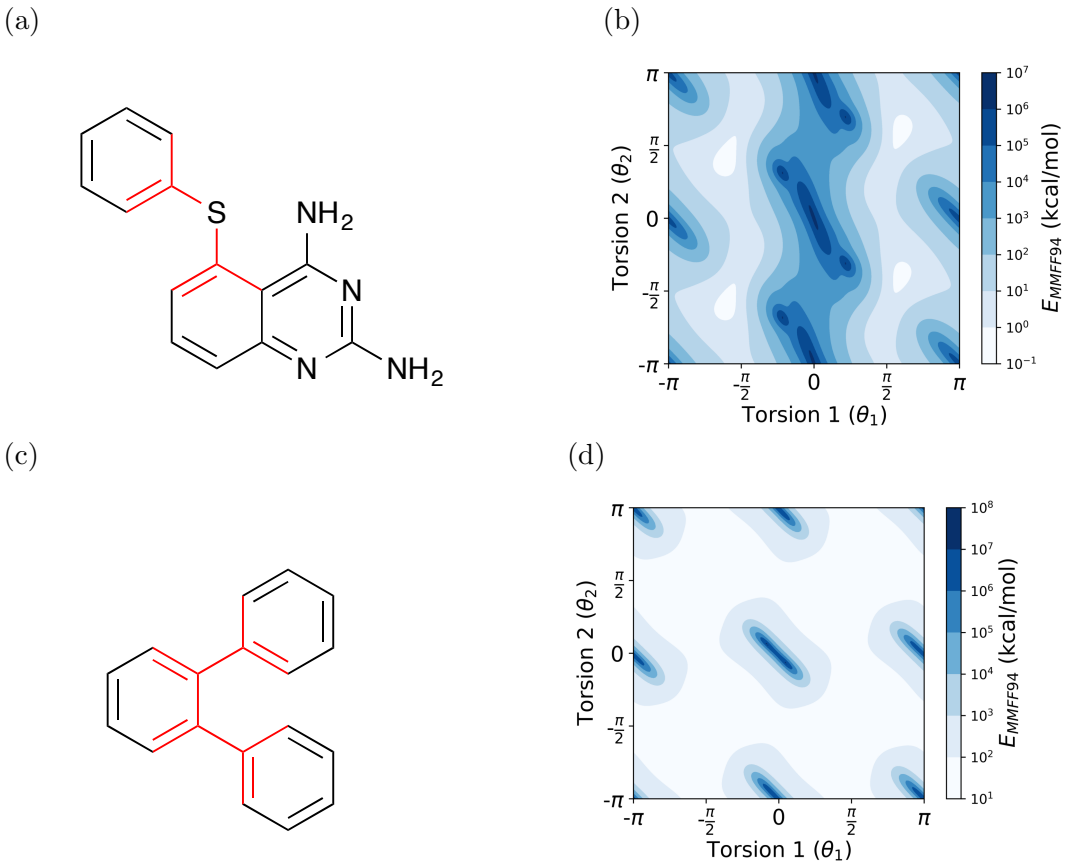


Figure 3.1: (a) 5-Phenylthioquinazoline-2,4-diamine. (b) MMFF94 potential energy landscape for 5-Phenylthioquinazoline-2,4-diamine. (c) *ortho*-1,1':2',1"-Terphenyl. (d) MMFF94 potential energy landscape for *ortho*-1,1':2',1"-terphenyl. The areas in light blue show the lowest-energy regions. The correlated torsions in the molecules are highlighted in red.

3.2 Methods and Data

In this work, rigid rotor approximation was used, *i.e.* bond lengths and bond angles were kept fixed. Hence, only the torsion angles about rotatable bonds were sampled. The torsion angles about the rotatable bonds are denoted by θ. Here, I first introduce the method, namely Bayesian Optimisation Algorithm (BOA) (Chan et al., 2019), followed by Bayesian Optimisation with Knowledge-based Expeceted Improvement (BOKEI) (Chan et al., 2020a). The underlying algorithms are almost identical, except the choice of acquisition function. The details are given below.

3.2.1 Bayesian Optimisation Algorithm (BOA)

There are two components in Bayesian optimisation: (i) surrogate model and (ii) acquisition function, as discussed in Chapter 2.8. Gaussian Process (GP) was used as the surrogate model, with a locally periodic kernel (Duvenaud, 2014; Chan et al., 2019, 2020a), as defined in Equation 3.1.

$$\begin{aligned} k_{LP} &= k_{RBF} k_{PER} \\ &= \sigma^2 \exp(-\frac{(\|\theta^a - \theta^b\|)^2}{2l^2}) \exp(\frac{-2\sin^2(\pi|\theta^a - \theta^b|/p}{l^2}) \end{aligned} \tag{3.1}$$

where l, p, σ^2 are the length scale, period and variance respectively. The θ^a and θ^b are two set of torsion angles in a molecule. This is simply the product of a RBF kernel and periodic kernel. Intuitively, the periodic kernel is used to capture the periodicity of the potential energy function, while the RBF kernel increases the flexibility of the GP, by allowing it to model the torsional potentials with varying amplitudes, as well as different local minima and maxima. Note that the Euclidean distance between two circular variables may not be a meaningful measure, as discussed in Chapter 2.9. Potential solutions are discussed in Section 3.4.

The periodicity was determined by torsion potentials corresponding to the 364 rotatable bond SMARTS patterns (Guba et al., 2016), see Appendix A, Table A.1. Note that this set of substructures were derived from commonly-occurring types of rotatable bonds observed in small molecules and protein-ligand X-ray crystal structures, and are also used in RDKit's (Landrum, 2018) Experimental-Torsion Distance Geometry with basic Knowledge (ETKDG) algorithm (Riniker and Landrum, 2015; Wang et al., 2020). When the list of patterns did not cover a specific type of rotatable bond, I assigned general values for the periodicity parameter based on the atomic hybridization of the two atoms in the rotatable bond, *i.e.* $sp^2 - sp^2$, $sp^2 - sp^3$, and $sp^3 - sp^3$.

To consider the adjacent correlated torsions, I enumerated all possible pairs of the 364 rotatable bond SMARTS patterns, and counted the frequencies of the corresponding pairs of torsion angles observed in small molecules having five or fewer rotatable bonds in the Crystallography Open Database (COD) (Gražulis et al., 2009, 2012). Pattern pairs with fewer than 100 observations were excluded, resulting in 19 adjacent correlated torsion patterns for the BOKEI framework, including one pattern from Cole et al. (2018), see Appendix A, Table A.2.

3.2.1.1 Acquisition Functions

In the BOA framework, two acquisition functions were used, namely Expected Improvement (Mockus et al., 1978) and Gaussian Process Lower Confidence Bound (GP-LCB) (Srinivas et al., 2009). In the BOKEI framework, a novel knowledge-based Expected Improvement (KEI) was used. The detail is discussed below.

Expected Improvement (EI)

$$\mathrm{EI}(\theta) = \sigma(\theta)(z(\theta)\Phi(z(\theta)) + \phi(z(\theta)) \tag{3.2}$$

Gaussian Process Lower Confidence Bound (GP-LCB)

$$\text{GP-LCB}(\theta) = \mu(\theta) - q_\alpha \sigma(\theta) \tag{3.3}$$

where $z(\theta) = \frac{f(\theta_{best}) - \mu(\theta)}{\sigma(\theta)}$, $\mu(\theta)$ and $\sigma^2(\theta)$ are the predictive mean and predictive variance respectively; $\Phi(\cdot)$, $\phi(\cdot)$, q_α are the cumulative distribution function, probability density function and the desired quantile of the standard normal distribution.

Knowledge-based Expected Improvement (KEI)

Knowledge-based Expected Improvement (KEI) can be considered as a modified EI algorithm that offers improvement only when a set of torsion constraints are satisfied:

$$a_{KEI}(\theta) = EI(\theta) \prod_{m=1}^{M} P_m(\theta_{m,1}, \theta_{m,2}) \tag{3.4}$$

where M is the total number of correlated torsions found in the molecule; $P_m(\theta_{m,1}, \theta_{m,2})$ is the mixture model of the torsion angle pairs in pattern m. Independence between each pair of correlated torsions were assumed. The idea of KEI is similar to the method of EI with Boolean constraints suggested in (Gelbart et al., 2014; Griffiths and Hernández-Lobato, 2020), with a user-specified minimum confidence of the cosntraints. Instead of Boolean constraints, I derived separate distributions of the correlated torsions from the lowest energy conformations found by MMFF94, and GFN2 (see simulation detail in Section 3.2.6). These were encapsulated by bivariate von Mises mixture models (see Section 3.2.2), and the resulting models were used in Equation 3.4.

3.2.2 Bivariate von Mises Distribution and Mixture Models

The bivariate von Mises distribution is a probability distribution that can be used to jointly model two angular variables, *i.e.* two torsion angles (θ_1, θ_2). Multiple bivariate von Mises models were developed, including the Sine model (Singh et al., 2002) and Cosine model (Mardia et al., 2007). The Cosine models (Equation 3.5 and 3.6 respectively) were used in my implementation.

Cosine density with positive interaction

$$g(\theta_1, \theta_2) = c(\kappa_1, \kappa_2, \kappa_3) \exp\{\kappa_1 \cos(\theta_1 - \mu) + \kappa_2 \cos(\theta_2 - \nu) - \kappa_3 \cos(\theta_1 - \mu - \theta_2 + \nu)\} \quad (3.5)$$

Cosine density with negative interaction

$$g(\theta_1, \theta_2) = c(\kappa_1, \kappa_2, \kappa_3) \exp\{\kappa_1 \cos(\theta_1 - \mu) + \kappa_2 \cos(\theta_2 - \nu) - \kappa_3 \cos(\theta_1 - \mu + \theta_2 - \nu)\} \quad (3.6)$$

where $c(\kappa_1, \kappa_2, \kappa_3)^{-1}$ is the normalizing constant with the following form:

$$c(\kappa_1, \kappa_2, \kappa_3)^{-1} = (2\pi)^2 \{I_0(\kappa_1) I_0(\kappa_2) I_0(\kappa_3) + 2 \textstyle\sum_{p=1}^{\infty} I_p(\kappa_1) I_p(\kappa_2) I_p(\kappa_3)\}$$

$I_r(\cdot)$ denotes the modified Bessels function of the first kind and order r; the parameters in the model represent the mean direction $(\mu,\ \nu)$, concentrations $(\kappa_1,\ \kappa_2)$ and a parameter (κ_3) controlling the correlation.

Unlike univariate von Mises distribution, as discussed in Chapter 2.9, the cosine densities can be unimodal $(\kappa_3 < \frac{\kappa_1 \kappa_2}{\kappa_1 + \kappa_2})$ or bimodal $(\kappa_3 > \frac{\kappa_1 \kappa_2}{\kappa_1 + \kappa_2})$. It is approximately bivariate normally distributed if and only if $\kappa_3 < \frac{\kappa_1 \kappa_2}{\kappa_1 + \kappa_2}$. Furthermore, the cosine density is flexible which allows us to consider transformation of all the data when estimating the model parameters. In particular, the cosine density with negative interaction can be obtained by transforming $(\theta_1, \theta_2) \mapsto (\theta_1, -\theta_2)$ in the model of cosine density with positive interaction (Mardia and Frellsen, 2012). In practice, the likelihood for original data and the transformed data are compared, and the one with larger value is selected.

Typically, there are multiple modes in the torsional space, and a single bivariate von Mises distribution is not sufficient to describe the correlated torsion. Therefore, a mixture model (Equation 3.7) was used.

$$g_M = \sum_{j=1}^{K} \omega_j g_j(\theta_1, \theta_2) \quad (3.7)$$

where K is the number of components, g_j denotes a cosine density, and ω_i is the weight of each component (with $\sum_i \omega_i = 1$).

Expectation Maximisation (EM), as discussed in Chapter 2.9, can be used to estimated the model parameters. It is well-known that the EM algorithm can easily get stuck in the local optimal. Hence, I performed the EM algorithm multiple times with different initialisation, and chose the best final solution. I also excluded any solutions with extremely high concentration. In the M-step, gradient ascent algorithm was used to update the model parameters. Note that the mixture models were applied to the 19 SMARTS patterns with adjacent correlated torsions, as defined in Appendix A, Table A.2. The resulting models were used as $P_m(\theta_1, \theta_2)$ in KEI acquisition function, Equation 3.4. The model parameters for all 19 SMARTS patterns are shown in Appendix A, Figure A.1.

We should note that the bivariate von Mises mixture model requires sufficient data to accurately describe torsional preferences, so cases with small numbers of observations were excluded in this work. This limited the current performance of my algorithm, and I will discuss some potential solutions in Section 3.4

3.2.3 Search Space

All methods explored the same search space for each molecule, as determined by the set of freely rotatable bonds in each. The search space of the algorithms was thus defined by a hypercube $[0, 2\pi)^d$, in the BOA framework, and $[-\pi, \pi)^d$ in the BOKEI framework, where d is the number of rotatable bonds in the molecule.

3.2.4 Comparison

3.2.4.1 Comparison between BOA, Confab and Uniform Search

To assess the effectiveness of the BOA algorithm, I compared it with two other conformational search algorithms, including a systematic search method, Confab (O'Boyle et al., 2011b) and a uniform random search. The expected improvement (EI) and Gaussian process lower confidence bound (LCB) acquisition functions were used in my BOA algorithm.

An energy cutoff of 500 kcal/mol was used in Confab, with up to one million conformers and a root mean square deviation clustering threshold of 0.05 Å; all other Confab

parameters were left as their default values. The RMSD cutoff of 0.05 Å was used to eliminate duplicate conformers with identical geometry to existing conformers. Note that only one compound (cochliodinol, a molecule with six rotatable bonds) would have generated more than a million conformers (1,327,104).

To ensure fair comparison between search algorithms, I used the same number of iterations, K, *i.e.* the number of energy evaluations, for all of the stochastic search methods. The number of energy evaluations depend on the number of rotatable bonds in a molecule, see Table 3.1. Note that by the nature of the algorithm, BOA requires initial observations of the energy landscape in order to fit a Gaussian Process. For each molecule, five initial observations were obtained by random sampling, and only $K - 5$ conformers were evaluated after initial sampling in BOA. Five runs were performed for all stochastic search algorithm, *i.e.* Uniform, BOA with EI and LCB.

Table 3.1: Number of simulated conformations versus number of rotatable bonds

Number of rotatable bonds	Number of conformers
1-3	50
4-6	100

To assess the effect of increasing the number of energy evaluations, I doubled the number of energy evaluations, *i.e.* $K = 200$, in BOA search for a subset of molecules with five rotatable bonds. I performed four runs for each molecule in the set.

A force field MMFF94 was used to evaluate the energy of the molecule throughout first assessment.

3.2.4.2 Comparison between BOKEI, BOA-EI and Genetic Algorithm

To evaluate the usefulness of the correlated adjacent torsions prior on conformer sampling, I compared BOA with my proposed knowledge-based expected improvement (KEI) acquisition function with standard EI, and a genetic algorithm (GA). The implementation of Bayesian optimisation was identical as above.

In GA, the search was terminated when either maximum number of energy evaluations was reached or three identical generations were observed; all other GA parameters were left as their default values.

Similarly, I fixed the same number of energy evaluations throughout the search, and I also included more flexible molecules, *i.e.* more than six rotatable bonds, for comparison. The number of energy evaluations depend on the number of rotatable bonds in a molecule, see Table 3.2. Likewise, only $K-5$ conformations were sampled after initial sampling in Bayesian optimisation algorithm. For accurate statistical comparisons of these methods, five runs were performed on each algorithm.

Table 3.2: Number of simulated conformations versus number of rotatable bonds

Number of rotatable bonds	Number of conformers
2-3	25
4-5	50
6-7	100
≥ 8	200

Two energy functions: (i) geometry-optimized MMFF94 and (ii) GFN2,were used. Instead of a single point MMFF94 energy evaluation, I performed constrained geometry optimisation with MMFF94, which the torsion angles in a molecule were kept fixed and other degrees of freedoms such as bond lengths and bond angles were being optimised. On the other hand, torsion constraints was not added in GA search, as surrogate models were not required.

3.2.5 Performance Metrics

As discussed in Chapter 2.5, the energy difference, heavy atom root-mean-square-deviation (RMSD) and torsion fingerprint deviation (TFD) are frequently used to assess the quality of the sampled conformations. Since my goal is searching for the lowest energy conformation, the energy difference should be used. RMSD and TFD were included for reference in the first comparison, *i.e.* comparison between BOA, Confab and uniform random search, while they were not reported in the second comparison.

In the first comparison, the molecule with the lowest energy conformation obtained from Confab was used as the reference conformation. Negative value indicates a better conformation was found by the search than by Confab. Note that symmetry was considered in the RMSD calculation. In the second comparison, *i.e.* comparison between BOKEI, BOA-EI and GA, the molecule with the lowest energy conformation obtained from BOA-EI was used as reference conformation. Similarly, negative value

indicates a better conformation was found by the search than by BOA-EI. Throughout my work, I denote the MMFF94 energy difference as ΔE_{MMFF94}, and GFN2 energy difference as ΔE_{GNF2}.

3.2.6 Data

In the first comparison, the datasets assembled by Ebejer et al. (2012) was used, which consists of 708 distinct small molecules and includes ligands from the Astex diverse set (Hartshorn et al., 2007). I filtered this set of molecules with six or fewer rotatable bonds, giving a subset of 572 molecules, excluding 4 rigid molecules.

In the second assessment, Platinum dataset (Friedrich et al., 2017) and a dataset assembled by Ebejer et al. (2012) were used to benchmark the performance of the search algorithms. The duplicated molecules in two datasets were removed based on their InChiKey (Heller et al., 2013). Molecules with 2 to 18 rotatable bonds, and consist of two adjacent rotatable bonds (correlated torsion) that match the 19 correlated torsion SMARTS patterns in Appendix A, Table A.2 were selected for the study, giving a subset of 533 unique molecules. Furthermore, I extracted molecules with correlated torsions from organic small molecules in the Crystallography Open Database (COD) (Gražulis et al., 2009, 2012) to derive the probabilistic constraints, as discussed above. The simulation of the lowest energy conformation is discussed next. Note that the overlapping molecules between the validation set and the COD were removed from the COD set based on their InChiKey.

3.2.6.1 Simulations

I calculated the lowest energy conformation of the molecules from the COD set, and used it to derive correlated torsions distribution. I only considered molecules with five or fewer rotatable bonds in this calculation. Under this setting, I could find the lowest energy conformation for both GFN2 and MMFF94 with high probability. The sampling schemes are described below.

MMFF94

I simulated diverse conformers by RDKit's ETKDG (Riniker and Landrum, 2015), followed by energy minimisation, and calculated the lowest energy conformation. Note that this is a basin-hopping style optimisation (Wales and Doye, 1997). Table 3.3 shows the number of initial conformers used in the simulation.

Table 3.3: Number of simulated conformations versus number of rotatable bonds

Number of rotatable bonds	Number of conformers
2	50
3	100
4	250
5	500

GFN2

I calculated the lowest energy conformation using the Conformer-Rotamer Ensemble Sampling Tool (CREST) (Pracht et al., 2020) based on GFN2 method. The CREST sampling consists of three parts: (i) meta-dynamic (MTD) simulation , (ii) regular MD sampling and (iii) a genetic structure crossing algorithms. In MTD, the atomic RMSDs are included as collective variables, and a Gaussian potential on RMSDs is introduced. The additional bias Gaussian potential helps generate diverse geometries. The GFN2 energy function is used throughout the entire process. Geometry optimisations are performed in between and in the final step. This iterative process is termed iMTD-GC, where the lowercase i indicates an iterative strategy within the algorithm. iMTD-GC workflow was used in the search. Note that CREST may break the molecule into fragments in the outputs and I discarded those observations in my analysis.

3.2.7 Statistical Tests

The Wilcoxon signed-rank test was used to test whether the distributions of the lowest energy conformations found by each pair of search methods was statistically significantly different from one another. In the first assessment, I compared three pairs of methods, including (BOA-EI, uniform), (BOA-LCB, uniform) and (BOA-EI, BOA-LCB). In the second assessment, I tested whether BOKEI framework found lower average energy conformations than BOA-EI and GA.

3.2.8 Implementation

In both comparisons, a Python package, GPyOpt (GPyOpt, 2016) was used for the Bayesian optimisation algorithm variants. Pybel (O'Boyle et al., 2008) was used to drive the torsion angles of the molecules. The MMFF94 energy and the RMSD

calculation that implemented in Open Babel 2.4.1 (O'Boyle et al., 2011a) were used. Numpy (van der Walt et al., 2011; Harris et al., 2020) was used to generate random numbers between 0 to 2π for the uniform search in the first comparison. The genetic algorithm (GA) implemented in Open Babel 2.4.1 was used in the second comparison.

3.2.9 Run Time Analysis

Run time analysis on my proposed algorithms was performed on a desktop running Fedora 28 with an Intel Core i7-6700 operating at 3.40 GHz, and 32 GB of RAM. A single core was used for energy evaluation and driving the torsion angles. All cores were used in the GPyOpt operations.

3.3 Results and Discussions

I first present the results in the first comparison, *i.e.* comparison between BOA, Confab and uniform random search, followed by the results in the second assessment. Note that due to occasional numerical instabilities, GPyOpt terminated early before reaching the maximum number of iterations requested. This was manifested by a non-positive definite kernel error. I separated out these molecules with "early stopping". The excluded molecules in comparison 1 are listed in Appendix A, Tables A.3, and A.4 and A.5 for comparison 2.

3.3.1 Comparison between BOA, Confab and Uniform Search

3.3.1.1 Number of Conformers Generated in Systematic Search

I first analyzed the number of conformers sampled by systematic search, Confab. Figure 3.2 shows that up to 10^6 conformers were explored for molecules with six rotatable bonds. For molecules with four or more rotatable bonds, the median number of conformers generated was approximately 10^3 to 10^4. Cochliodinol has six rotatable bonds and had generated over one million conformers, with 750,402 conformers retained, and the lowest energy of 146.04 kcal/mol. Bayesian optimisation algorithm, on the other hand, required only 10^2 evaluations to obtain low energy conformations, and the best conformation out of five trials had an almost identical energy of 146.13

kcal/mol. This highlights the power of BOA, and it gives good performance in general despite using orders of magnitude fewer energy evaluations.

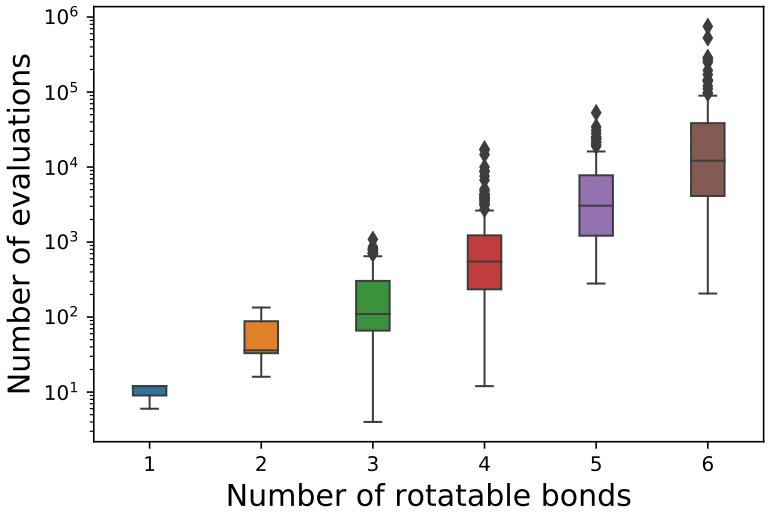


Figure 3.2: Distributions of the number of energy evaluations by Confab versus the number of rotatable bonds.

3.3.1.2 Search Performance

As expected, the uniform random search performed the worst of all search methods. It gave higher median energy differences and larger ranges in energy difference than BOA search across all sets of rotatable bonds, as shown in Figure 3.3. The distributions of the energy differences were very similar for BOA search, with both acquisition functions, EI and LCB. When constrained by a maximum number of energy evaluations, uniform random search suffered more in higher dimensions than BOA, and the median of the energy differences increased rapidly. On the other hand, the median of the energy differences in BOA search increased slowly and reaches approximately 9 kcal/mol for molecules with six rotatable bonds.

Confab was used to enumerate systematically all conformers for each molecule using the 'torsion driving approach'. Confab iterates systematically through a set of allowed torsion angles for each rotatable bond in the molecule. Being a systematic search, Confab was thus expected to identify all the low energy conformations for each molecule. However, the best torsion angles may not be covered by the set of discrete torsion angles used in Confab. On the other hand, BOA samples torsion angles freely in the space and learns from the observed conformations, which enables it to recover

conformations with lower energies using orders of magnitude fewer evaluations than Confab.

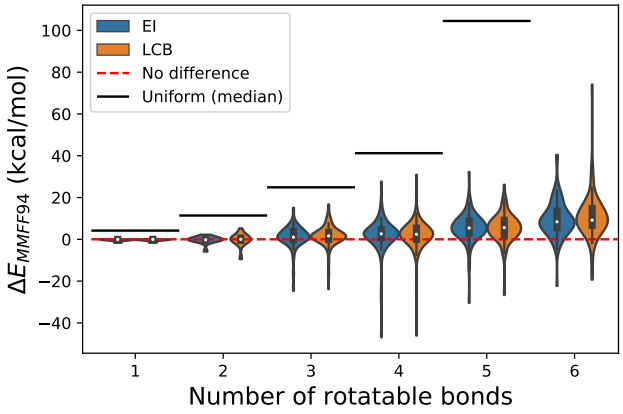


Figure 3.3: MMFF94 energy difference versus number of rotatable bonds: BOA and uniform random search. The red dotted line indicates no energy difference between the lowest energy conformation found by the search algorithm and that found by Confab. The black line indicates the median energy difference of the uniform random search. Note that the median (213 kcal/mol) for molecules with six rotatable bonds is not shown in figure. The blue and orange groups show the result for BOA with EI, and BOA with LCB, respectively

I define the lowest energy conformer (LEC) for a given molecule as the lowest energy conformation found by any search method in our experiments. I computed the frequency that each method (Confab, BOA, and Uniform) was able to find each molecule's LEC. Since the number of energy evaluations used in the search varies between stochastic and systematic methods, I also computed the average efficiency of the Bayesian optimisation and Confab. The average efficiency takes the number of energy evaluations into account, and is defined as $\frac{\sum_i^{N_{\text{mol}}} \frac{\mathbb{1}_{\text{lowest}}}{N_{\text{eval}}}}{N_{\text{mol}}}$, where $\mathbb{1}_{\text{lowest}}$ is 1 if the search method found the lowest energy conformation, N_{eval} is the number of energy evaluation, and N_{mol} is the number of molecules. Figure 3.4 shows that BOA recovers the most LECs for molecules with three or fewer rotatable bonds. This suggests that the geometries of the LECs deviate slightly from those with ideal torsion angles used in Confab. It should be noted that these non-ideal conformers cannot be generated by Confab. Although Confab found more lower energy conformations for molecules with four or more rotatable bonds, it used more energy evaluations in the search and the average efficiency was lower across all rotatable bonds, as shown in Figure 3.4. Examples of conformations found by BOA that have significantly lower energies than those found by Confab are shown in Figure 3.5.

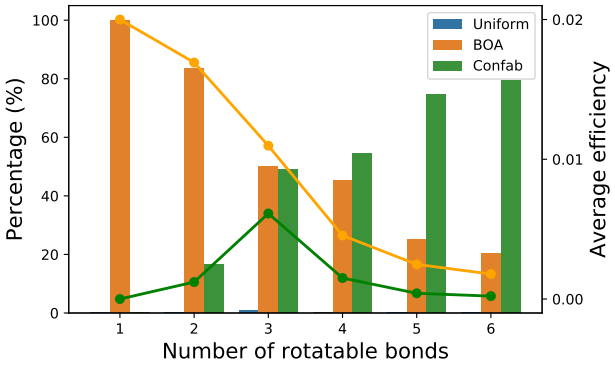


Figure 3.4: Percentage of the lowest possible energy conformation found by any of the search methods, from five independent trials, and the average efficiency of search methods. Confab found more lower energy conformations than BOA for molecules with four or more rotatable bonds. However, Confab (green line) has a lower average efficiency than BOA (orange line), as it used orders of magnitude more energy evaluations for the search, as shown in Figure 3.2.

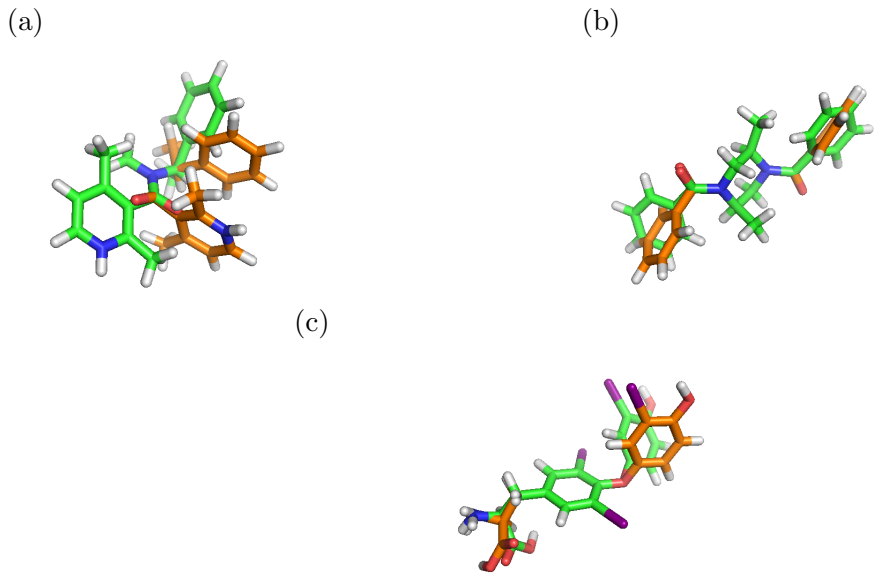


Figure 3.5: Examples where BOA found lower energies than Confab: (a) for omegacsd-FUPFIF, the lowest energy Confab found was 117 kcal/mol, while for BOA, it was 70 kcal/mol; (b) for omegacsd-CDBMPI10, the lowest energy Confab found was 150 kcal/mol, while for BOA: 118 kcal/mol; (c) for omegapdb-1SN5, the lowest energy Confab found was 131 kcal/mol, while for BOA, it was: 99 kcal/mol. The lowest energy conformations found by Confab and BOA are shown in green and orange respectively. Figures are generated by PyMOL (Schrödinger, LLC, 2015).

I assessed the "champion" rate' of all search methods in each trial, *i.e.* the percentage

of molecules that the search algorithm found better conformations than Confab in a single trial. Figure 3.6 shows that uniform random search had the lowest champion rate. I observed a similar rate in BOA search with both acquisition functions, EI and LCB. BOA search had a very high champion rate of 100% and 55% in molecules with one and two rotatable bonds, respectively. As expected, it decreased as the number of rotatable bonds increased. The champion rate was approximately 25% for molecules with three and four rotatable bonds, and 10% for molecules with five or more rotatable bonds. A key question here is how many samples are required to recover a better conformation and thus achieve a high recovery rate. I addressed the influence of the maximum number of evaluations by doubling the number of energy evaluations, see Section 3.3.1.3.

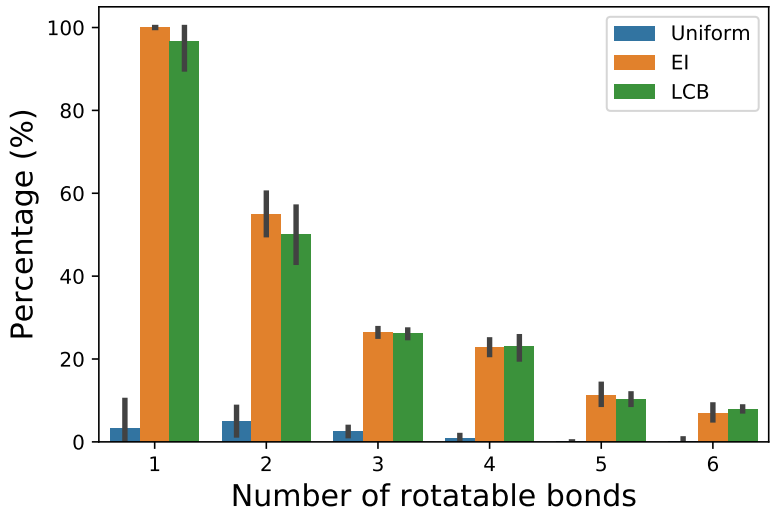


Figure 3.6: Champion rate: percentage of conformers with lower energy than Confab recovered by uniform random search, BOA with EI and LCB. The error bars show the variation in the trials. It can be seen that both variants of BOA recovered a lower energy conformer than Confab much more often than uniform random search, although the overall recovery rate dropped as the number of rotatable bonds increased. A maximum of 100 iterations was used for molecules with four or more rotatable bonds and 50 iterations otherwise.

The Wilcoxon signed-rank test of energy difference distributions in Table 3.4 shows that uniform random search is significantly different from BOA with EI and LCB (p-value $\ll 0.01$) for all numbers of rotatable bonds. Since the sample sizes of the sets of molecules with one and two rotatable bonds were small, I combined these with molecules having three rotatable bonds for the statistical test to give more reliable test results. For the EI-LCB pair, large p-values were obtained, except for molecules

with one to three rotatable bond (p-value of 0.02). Thus I found no evidence to reject the null hypothesis that the results for EI and LCB come from the same distribution.

Table 3.4: Energy difference: Wilcoxon signed-rank test on each method pair. Molecules with three or fewer rotatable bonds ($N_{\text{rotor}} : 1, 2, 3$) are grouped together due to small sample size. The p-values are rounded to 2 significant figures.

Method-Pair \ N_{rotor}	1,2,3	4	5	6
Uniform-EI	$\mathbf{8.1 \times 10^{-24}}$	$\mathbf{4.5 \times 10^{-23}}$	$\mathbf{3.5 \times 10^{-17}}$	$\mathbf{2.8 \times 10^{-15}}$
Uniform-LCB	$\mathbf{4.5 \times 10^{-24}}$	$\mathbf{4.5 \times 10^{-23}}$	$\mathbf{3.7 \times 10^{-17}}$	$\mathbf{2.6 \times 10^{-15}}$
EI-LCB	**0.02**	0.44	0.89	0.09

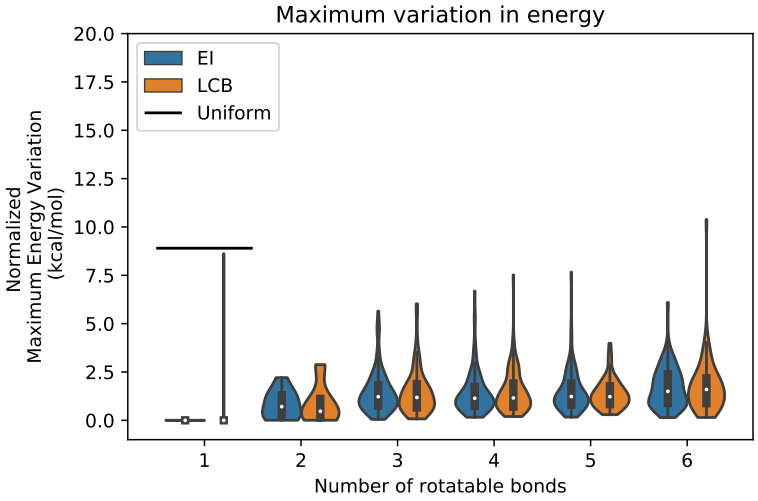


Figure 3.7: Normalized maximum energy variation versus number of rotatable bonds in BOA. The black horizontal line indicates the median of normalized maximum energy variation in uniform random search. The median for molecules with two or more rotatable bonds are greater than 20 kcal/mol, and they are not shown in the figure.

Furthermore, I assessed the variation in energies found by BOA. In particular, the normalized maximum energy variation for each molecule, *i.e.* maximum energy difference between computed conformation and the lowest energy conformation found in all trial divided by the number of rotatable bonds, was computed as shown in Figure 3.7. The variation in BOA was smaller than that in uniform search. The variation increased increased gradually as the number of rotatable bonds increased in BOA search.

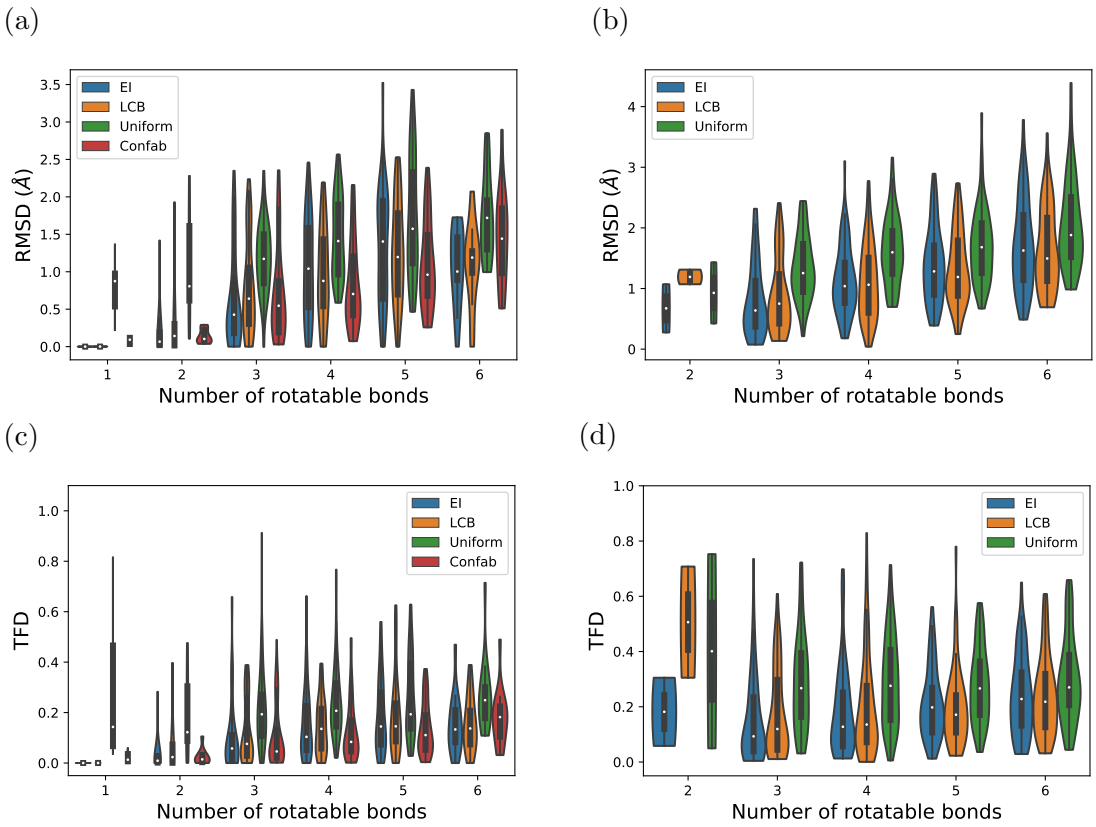


Figure 3.8: RMSD and TFD by varying number of rotatable bond. (a) RMSD case 1: the lowest energy conformations found by either BOA or uniform random search were used as reference conformations. (b) RMSD case 2: the lowest energy conformation found by Confab were used as reference conformations. (c) TFD case 1: the lowest energy conformations found by either BOA or uniform random search were used a reference conformations. (d) TFD case 2: the lowest energy conformations found by Confab were used as reference conformations. It should also be noted that there were 6 and 10 molecules with one and two rotatable bonds in panel (a) and (c). There were only two molecules with two rotatable bonds in panel (b) and (d). It can be seen that BOA with either EI or LCB outperformed uniform search. It is notable that BOA performed as well as Confab despite tending to use orders of magnitude fewer iterations.

Both RMSD and TFD were used to measure the distance between reference conformer and that obtained by various search methods. The lowest energy conformation across all methods was used as the reference conformer, and two scenarios were considered. Case 1 considered the lowest energy conformation found by either BOA or uniform random search from all trials for each molecule, while case 2 considered the lowest energy conformation found by Confab.

Figure 3.8 shows that the conformers generated by uniform random search had higher

RMSD and TFD values than those generated by BOA. The conformations found by BOA with both acquisition functions had similar distributions in RMSD and TFD values in case 2, while EI and LCB slightly vary in case 1.

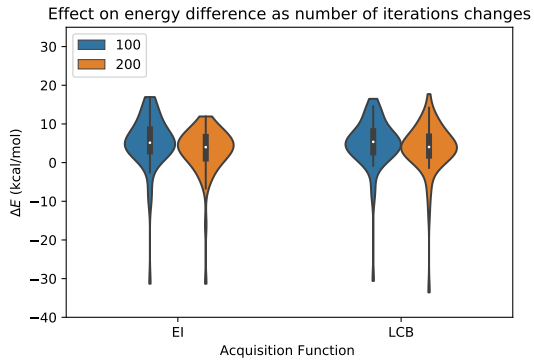
Similarly, I grouped molecules with three or fewer rotatable bonds together. In addition, I combined molecules with five or more rotatable bonds together in case (1) due to the small sample size in molecules with six rotatable bonds. Wilcoxon signed-rank tests for the RMSD and TFD distributions showed consistent results, see Appendix A, Tables A.6 and A.7: the distribution of conformers generated by uniform random search was significantly different from that generated by BOA (p-value $\ll 0.01$). The conformers generated by BOA with both acquisition functions, EI and LCB, were not statistically different from each other.

3.3.1.3 Doubling number of energy evaluations

I investigated the effect of doubling the maximum number of energy evaluations to 200 on the BOA, for the set of molecules with five rotatable bonds. I found that the results were more robust and had smaller ranges of energetic differences than were found with 100 iterations. Figure 3.9 shows that the median of the energy difference distributions decreased by 1.1 kcal/mol for EI, and 1.3 kcal/mol for LCB. The maximum variation also decreased, by 1.5 kcal/mol for EI, and 1.7 kcal/mol for LCB. Thus, increasing the maximum number of iterations improves the likelihood of finding low-energy minima, and decreases the stochastic variance between multiple runs.

Performance in terms of finding the lowest energy was improved by increasing the maximum number of energy evaluations. However, the computational cost also increased, in a rate between $n^{2.3}$ and $n^{2.7}$, where n is the number of energy evaluations, as shown in Figure 3.10. We should note that the computational complexity of the Gaussian process regression is $\mathcal{O}(n^3)$ in theory, due to the matrix inversion. The time shown in Figure 3.10 included computational time for energy evaluation ,T_{energy}, and Bayesian optimisation, T_{BOA}, and time to read input molecules or write the conformers to disk was not included. In this comparison, a relatively fast energy function was used, which only took about 7 ms to update the torsions and evaluate the energy 100 times in a molecule with six rotatable bonds. Hence, the relative contribution of energy evaluation time to the total computational time was small, *i.e.* $T_{\text{energy}} << T_{\text{BOA}}$.

(a)



(b)

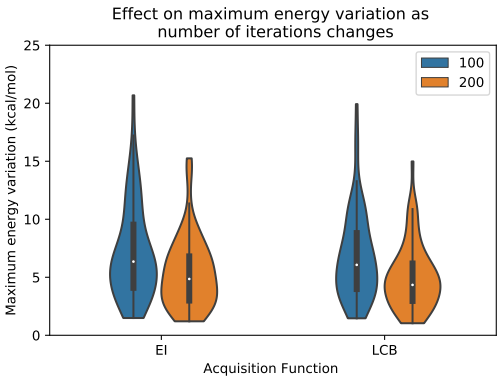


Figure 3.9: Effect of doubling the maximum number of energy evaluations: (a) energy difference; (b) maximum variation in energy. Note that only molecules with five rotatable bonds were tested. BOA with 100 iterations and 200 iterations are shown in blue and orange respectively. As expected, the energy difference and variation in energy decreased as number of energy valuations increased.

If the energy function was replaced by a more accurate but computationally more expensive method, such as quantum mechanical methods, as discussed in Chapter 2.6, which can take hours for the evaluations, the relative contribution of computational time of Bayesian optimisation would be small in such scenario, *i.e.* $T_{\text{energy}} >> T_{\text{BOA}}$. The Bayesian optimisation algorithm would become a more competitive search strategy than those used by other stochastic search methods.

In fact, other surrogates model can be used to reduce computational cost, for instance Bayesian neural network (BNN) (Snoek et al., 2015; Häse et al., 2018). Alternatively, one can incorporate more accurate prior knowledge to enhance the search, which will be further discussed in the Section 3.3.2.

3.3.1.4 Limitations

In addition to high computational complexity, poor parameter initialization leads to numerical instabilities. To address it, one would be to place priors on the parameters; alternatively, it may be possible to set boundary constraints on the parameters. The former approach would require more computational power, but would give a more robust estimation of the parameters.

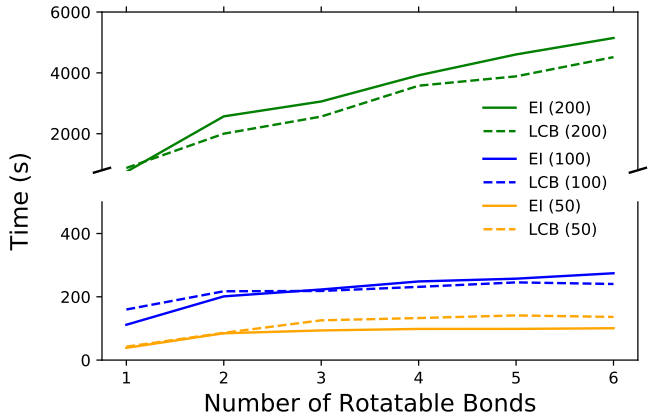


Figure 3.10: Average computational time for BOA with two difference acquisition function (EI or LCB) and different number of evaluations (50, 100 and 200). Five molecules were randomly sampled for each rotor. The average computational time increased as number of rotatable bonds increased, but was primarily dominated by the number of conformers generated.

3.3.1.5 Summary

In summary, I demonstrated the effectiveness of Bayesian optimisation algorithm (BOA) in searching the lowest energy conformation. I incorporated prior knowledge about torsion angle preferences to accelerate the search. This strategy even finds lower energy minima than those generated by systematic enumeration, using orders of magnitude fewer energy evaluations.

3.3.2 Bayesian optimisation Algorithm with Knowledge-based Expected Improvement (BOKEI)

In this section, I will evaluate the usefulness of the correlated adjacent torsions prior on conformer sampling. As discussed above, the knowledge of correlated adjacent torsion priors can be embedded in the acquisition function, *i.e.* KEI. I revisited the example, 5-phenylthioquinazoline-2,4-diamine, and the sampling behavoir of BOA-EI and BOKEI are shown in Figure 3.11. Figure 3.11b shows the mixture models of the adjacent correlated torsions in the molecule, it exhibits reflectional dependence between torsion angles. Using KEI, the sampled conformations were more likely observed in the region where the mixture models with high density values. In both cases, the posterior mean showed a coarse boundary between high and low energy regions.

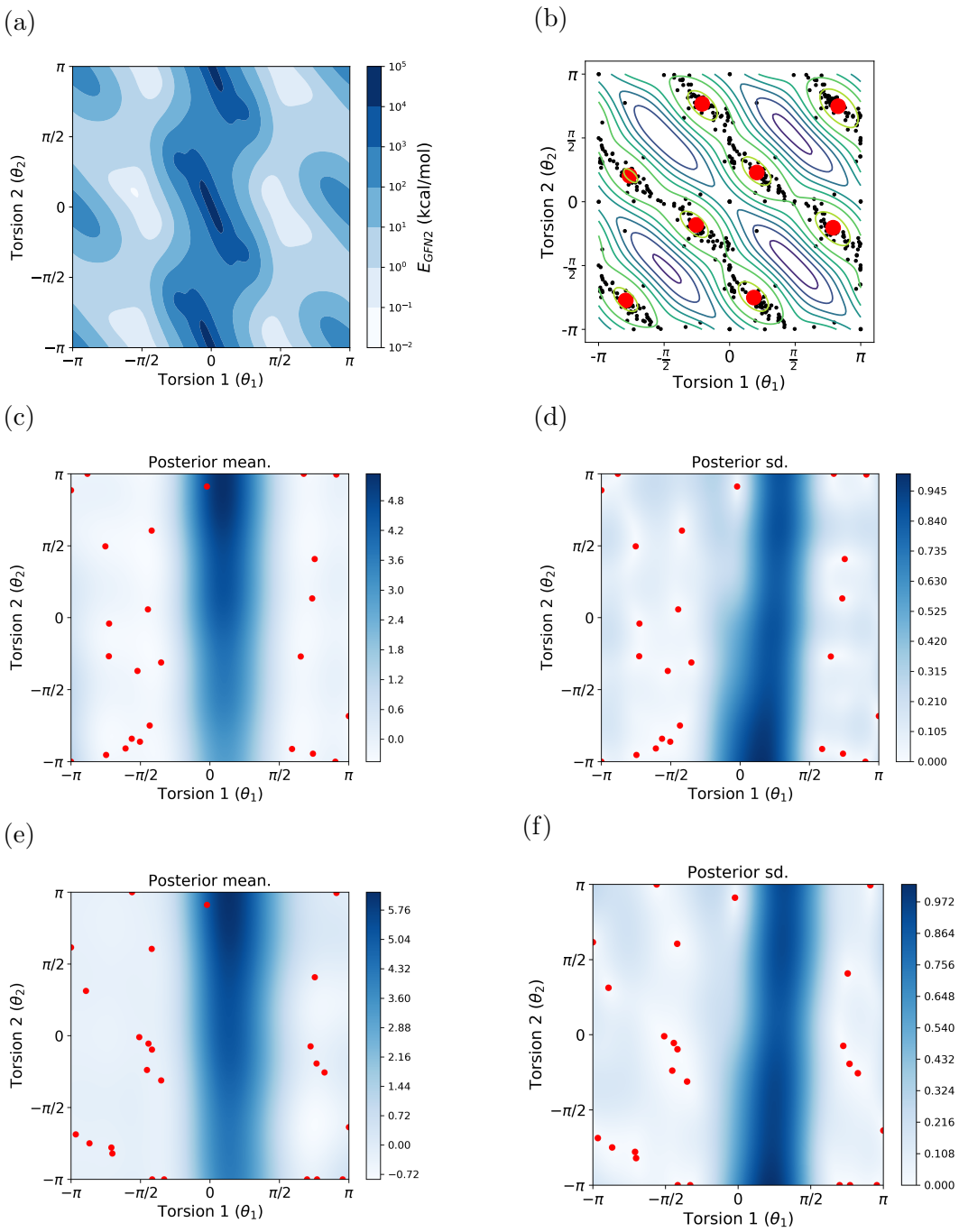


Figure 3.11: (a) GFN2 potentail energy surface of 5-phenylthioquinazoline-2,4-diamine. (b) Mixture models for the correlated torsion. The contour plot indicates the log density of the mixture model and the points (in red) mark the mean location for the components. (c) Posterior mean (normalized) energy landscape and (d) posterior standard deviation in BOA-EI. (e) Posterior mean (normalized) energy landscape and (f) posterior standard deviation in BOKEI. The samples were more concentrated in the low energy regions for BOKEI than that in BOA-EI.

Furthermore, two molecules were sampled to illustrate the strengths and the weaknesses of the BOKEI algorithm, using the geometry-optimized MMFF94 energy. Ten runs were performed for each molecule. Figure 3.12a shows that BOKEI was able to find lower energy conformation consistently using same number of iterations. The energy gap between BOKEI and BOA-EI decreased as the number of evaluations increased, since both methods should converge to the same global minimum. Unusually, the performance of BOKEI was worse than BOA-EI, which was a result of under-estimation of the correlated torsion distribution, as illustrated in Figure 3.12b. Insufficient sampling or biased selection of molecules gave rise to the incomplete prior information, and led to the inferior performance. Using additional data to improve the density estimates, even in this case, the performances became similar, as discussed in Section 3.3.2.5.

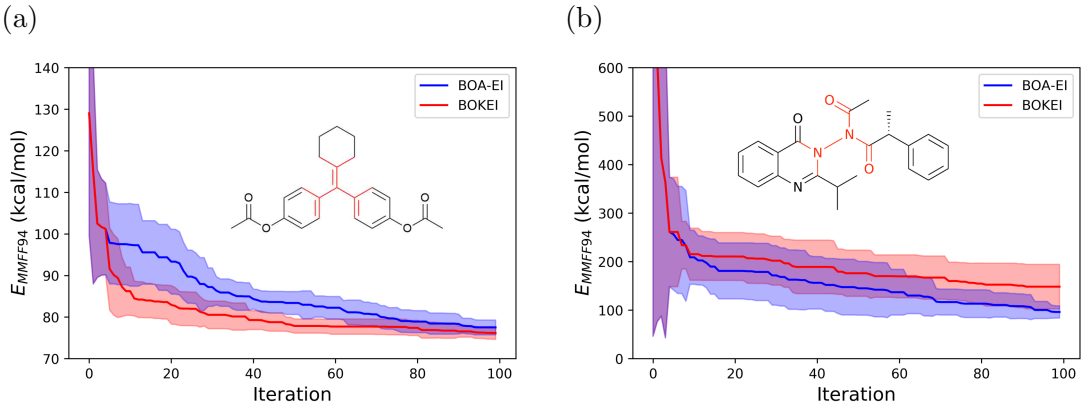
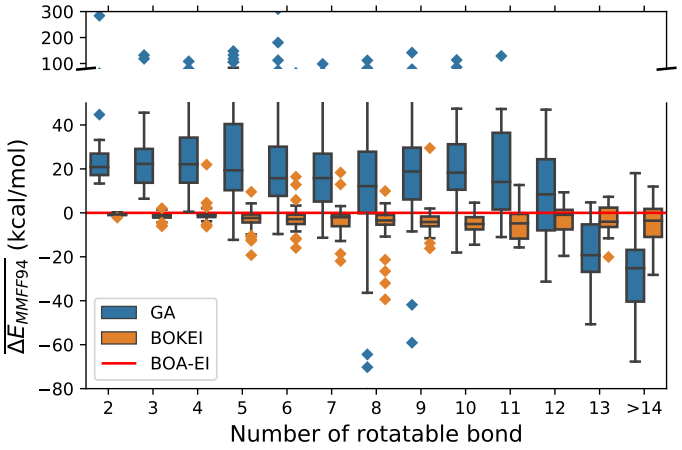


Figure 3.12: The red line and the blue line in the convergence plots represents the average rate of the BOKEI and BOA-EI in finding lower energy conformations respectively, with ±1 sample standard deviation (shaded region). The correlated torsion in corresponding molecule is highlighted in red. Geometry-optimized MMFF94 energy function was used in both cases. (a) BOKEI consistently found lower energy conformations than BOA-EI in early stage and the energy gap reduced as the number of iterations increased. (b) Unusually, BOKEI performed worse than BOA-EI, which was a result of under-estimation of the correlated torsion.

For a broader comparison, I validated the performance on a set of 533 molecules containing 2 to 18 rotatable bonds, as mentioned above. In GFN2, I benchmarked the performance with molecules up to 13 rotatable bonds only. Note that there were nine molecules (five in MMFF94 and four in GFN2) excluded from analysis due to early stopping in the search, see Appendix A, Tables A.4 and A.5.

(a)



(b)

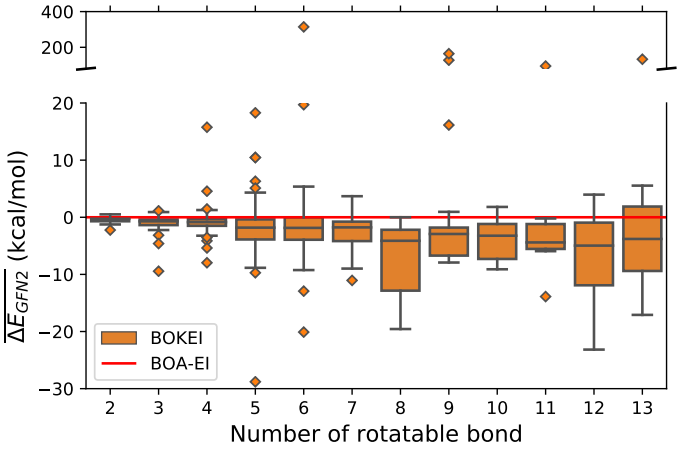


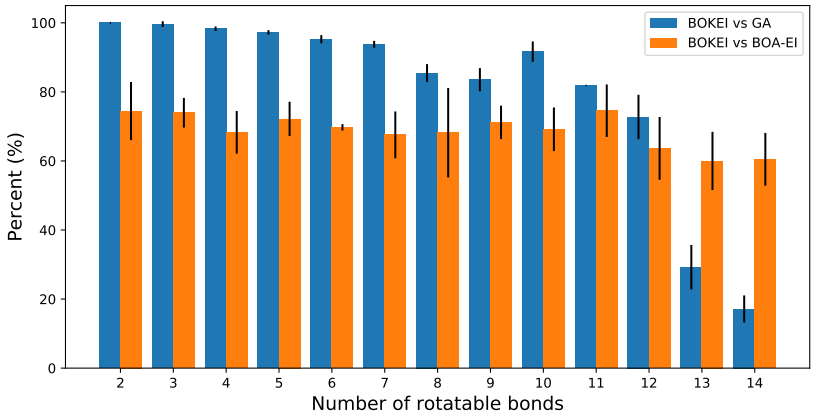
Figure 3.13: (a) MMFF94 average energy difference from five runs. (b) GFN2 average energy difference from five runs. The average energy of the outputs from all runs found by BOA-EI was used as the reference point (red line) in (a) and (b). BOKEI often found lower energy conformations than BOA-EI in both cases. The GA in MMFF94 outperformed BOKEI and BOA-EI for molecules with twelve or more rotatable bonds.

3.3.2.1 MMFF94

Figure 3.13a shows that BOKEI consistently found a lower energy conformation than BOA-EI and GA. A Wilcoxon signed rank test shows that energy difference between BOKEI and BOA-EI was statistically significant (p-value $\ll 0.01$, see Appendix A, Table A.8), across all rotatable bonds. On the other hand, one can see that the GA outperformed BOKEI and BOA-EI for molecules with more than twelve rotatable bonds. This is because the small number of samples (200 energy evaluations) may

not be sufficient for the BOA-EI or BOKEI models to learn the most likely dihedral angles in high dimensional problems. Figure 3.14a shows that BOKEI frequently found lower energy conformations than BOA-EI ($\approx 65 - 75\%$) and GA ($> 80\%$) for molecules with fewer than eleven rotatable bonds respectively. Figure 3.15a also shows that BOKEI gave a lower variation in energy than BOA-EI of the output conformations in all five runs. In both cases, the variation increased slowly to less than 20 kcal/mol for molecules with more than 14 rotatable bonds.

(a)



(b)

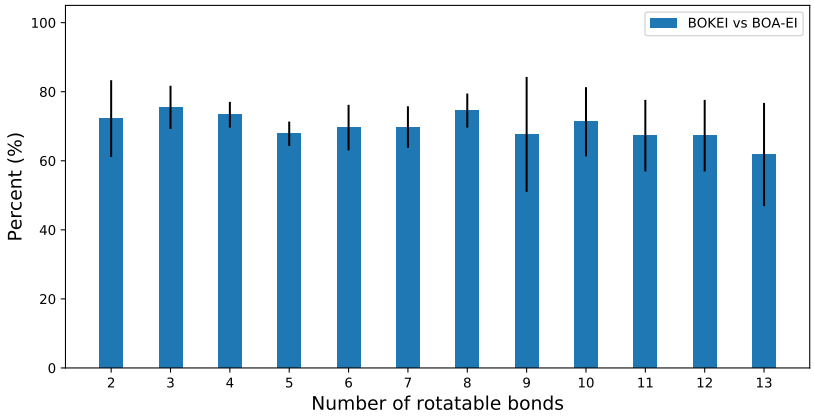
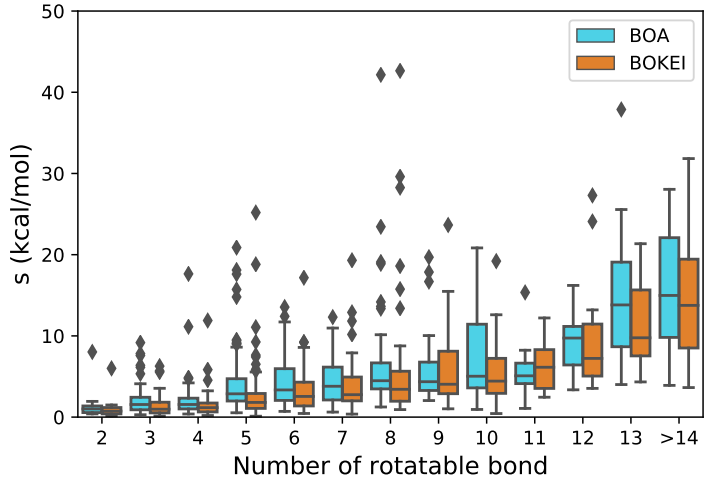


Figure 3.14: (a) Percentage of BOKEI found lower energy conformations than BOA-EI and GA, with geometry-optimized MMFF94 energy. (b) Percentage of BOKEI found lower energy conformations than BOA-EI, with GFN2 energy.

(a)



(b)

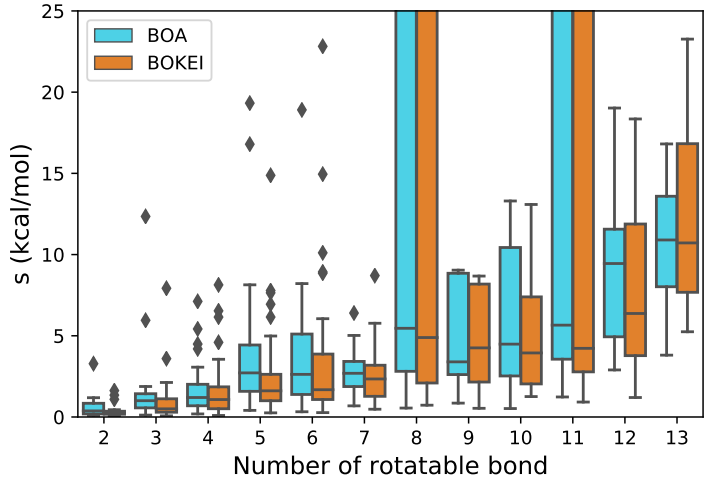


Figure 3.15: Sample standard deviation, s, of the energy of the output conformations in five independent runs: (a) MMFF94, and (b) GFN2. The sample standard deviation increased as the number of rotatable bonds increased in both cases. BOKEI had a lower median sample standard deviation than BOA in almost all cases. In (b), for a few molecules with eight and eleven rotatable bonds, one in five runs terminated with a high energy, which resulted in high sample standard deviations.

Furthermore, Figure 3.12a highlighted that BOKEI had greater benefit in the early stage of the search. Comparing the mean energy difference between BOKEI and BOA-EI at different stages (40%, 60% and 100%) of the maximum number of energy evaluations), the energy gap was indeed greater, and in favor of BOKEI, in the early stages, see Table 3.5. The gap diminished when more evaluations were used, since both methods converged to the same global optimum. These results suggest that the

information about correlated torsions greatly helped the search in the early stage, pointing the search towards favorable regions of the potential energy landscape.

Table 3.5: Average MMFF94 energy difference in different stages: 40%, 60% and 100% of the maximum number of energy evaluations. The value found by BOA-EI was used as reference. Negative value indicated BOKEI found lower energy conformations than BOA-EI. The median was reported. In general, BOKEI outperformed BOA-EI in the early stage, and the energy difference diminished as more evaluations were used.

No. of rotatable bonds	$\text{Median}_{40\%}$	$\text{Median}_{60\%}$	$\text{Median}_{100\%}$
2	-12.567	-8.482	-0.736
3	-17.910	-11.980	-1.151
4	-12.305	-8.897	-0.962
5	-24.361	-12.811	-2.324
6	-24.204	-15.089	-2.883
7	-16.496	-20.423	-1.995
8	-10.831	-15.354	-3.472
9	-30.170	-12.668	-4.177
10	-83.867	-18.585	-5.077
11	-87.031	-18.255	-4.801
12	-7.456	-66.339	-0.901
13	24.885	284.054	-4.037
14-18	68.935	58.545	-3.569

3.3.2.2 GFN2

In GFN2, I used a single-point energy calculation, and excluded GA in the analysis. Figure 3.13b shows that BOKEI consistently found lower energy conformations than BOA-EI. Similarly, a Wilcoxon signed rank test shows the average energy difference between BOKEI and BOA-EI was statistically significant (p-value $\ll 0.01$, see Appendix A, Table A.9). The energy gap was greater in the early stage and the gap diminished as more energy evaluations were used, see Table 3.6. Figure 3.14b also shows that BOKEI frequently ($> 60\%$) found lower energy conformations than BOA-EI across all rotatable bonds. Figure 3.15b shows that BOKEI gave a lower variation in energy than BOA of the output conformations in all five runs.

Table 3.6: Average GFN2 energy difference in different stages: 40%, 60% and 100% of the maximum number of energy evaluations. The value found by BOA-EI was used as reference. Negative value indicated BOKEI found a lower energy conformation than BOA-EI. The median was reported. In general, the BOKEI found lower energy conformations than BOA-EI in the early stage for molecules with seven or fewer rotatable bonds. The energy gap between BOKEI and BOA-EI diminished as more evaluations were used. The molecules with eight or more rotatable bonds had positive energy difference in the early stage.

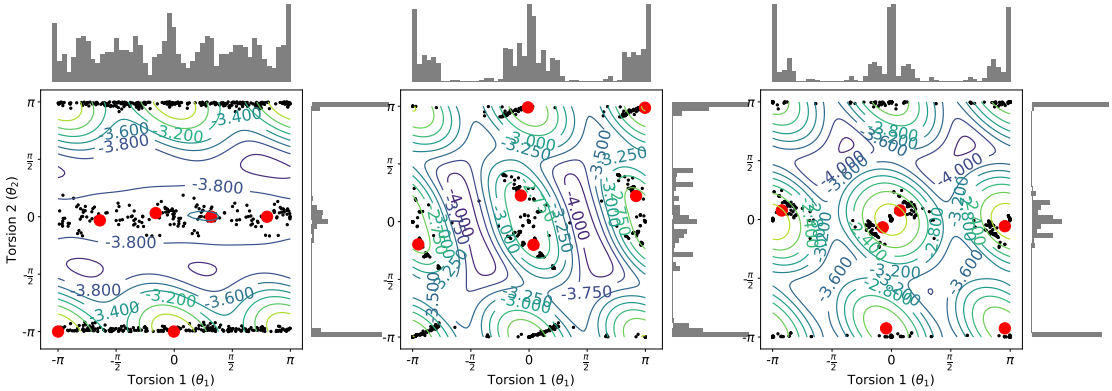
No. of rotatable bonds	$\text{Median}_{40\%}$	$\text{Median}_{60\%}$	$\text{Median}_{100\%}$
2	-14.143	-8.438	-0.276
3	-33.585	-13.870	-0.658
4	-8.817	-11.639	-0.856
5	-53.612	-15.612	-1.815
6	-14.688	-22.946	-1.857
7	-71.926	-50.814	-1.772
8	54.914	4.356	-4.117
9	119.435	11.961	-2.927
10	-57.236	-95.970	-3.228
11	88.074	-950.314	-4.405
12	-1075.099	-921.136	-4.961
13	1148.000	1937.425	-3.796

3.3.2.3 Correlated Torsions

Correlated torsion angles between adjacent rotatable bonds usually arise because of unfavorable steric interactions and favorable intramolecular interactions. Intramolecular hydrogen bonds and π-π stacking were observed in four of the nineteen patterns. For example, in pattern 2 and 16, the thioamide and thiourea functional groups play an important role in determining the torsion preferences. The delocalization of the nitrogen lone pairs contributes to its all planarity, and results in either *cis* ($\theta \approx 0$) or *trans* ($\theta \approx \pi$) configurations, see Figure 3.16. The *cis*/*trans* configuration preference depends on the atom environments. In particular, a specific thiourea derivative that is bonded to a carbonyl group was found to adopt the following conformations: (i) the C=S and C=O were oriented in "opposite" directions, while (ii) the thiourea adopted the syn-anti conformation (Sahu et al., 2011). This resulted in the formation of a pseudo six-membered ring that was stabilized by a C=O–H-N intramolecular hydrogen bond, see example Figure 3.18a. To align such intramolecular hydrogen bonds, the associated torsion angles were restricted, as illustrated in Figure 3.17. Similarly,

the formation of intramolecular hydrogen bonds between N-H or O-H groups and the adjacent carbonyl oxygen atoms in the esters can be easily observed in the substructure defined in pattern 15, see example Figure 3.18b.

(a)



(b)

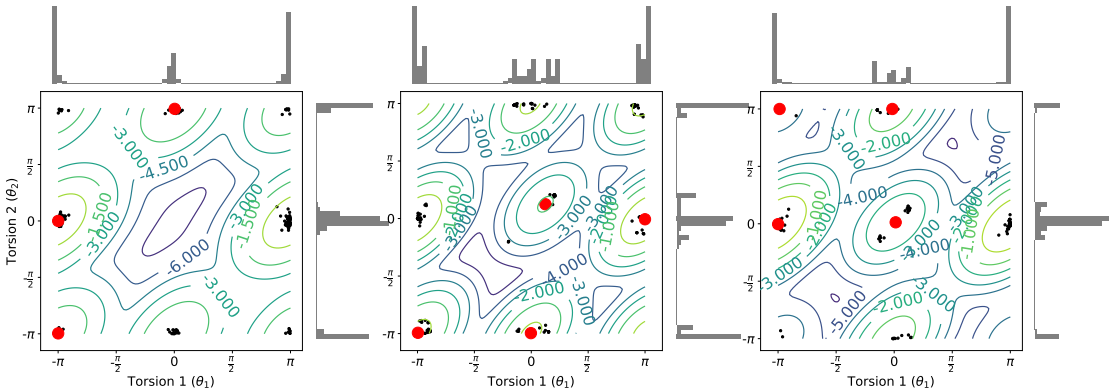


Figure 3.16: Mixture models for correlated torsions. The contour plot indicates the log density of the mixture model and the points (in red) mark the mean location for the components. The histogram indicates the marginal distribution of the torsion angles. The torsional preference of X-ray crystal structures, the lowest energy conformation from MMFF94 and GFN2 are on the left, middle and right respectively. (a) Pattern 2. (b) Pattern 16. The torsion 2 (θ_2) in pattern 2, and both torsions (θ_1,θ_2 were restricted around 0 or $\pm\pi$ due to the delocalisation of nitrogen lone pairs.

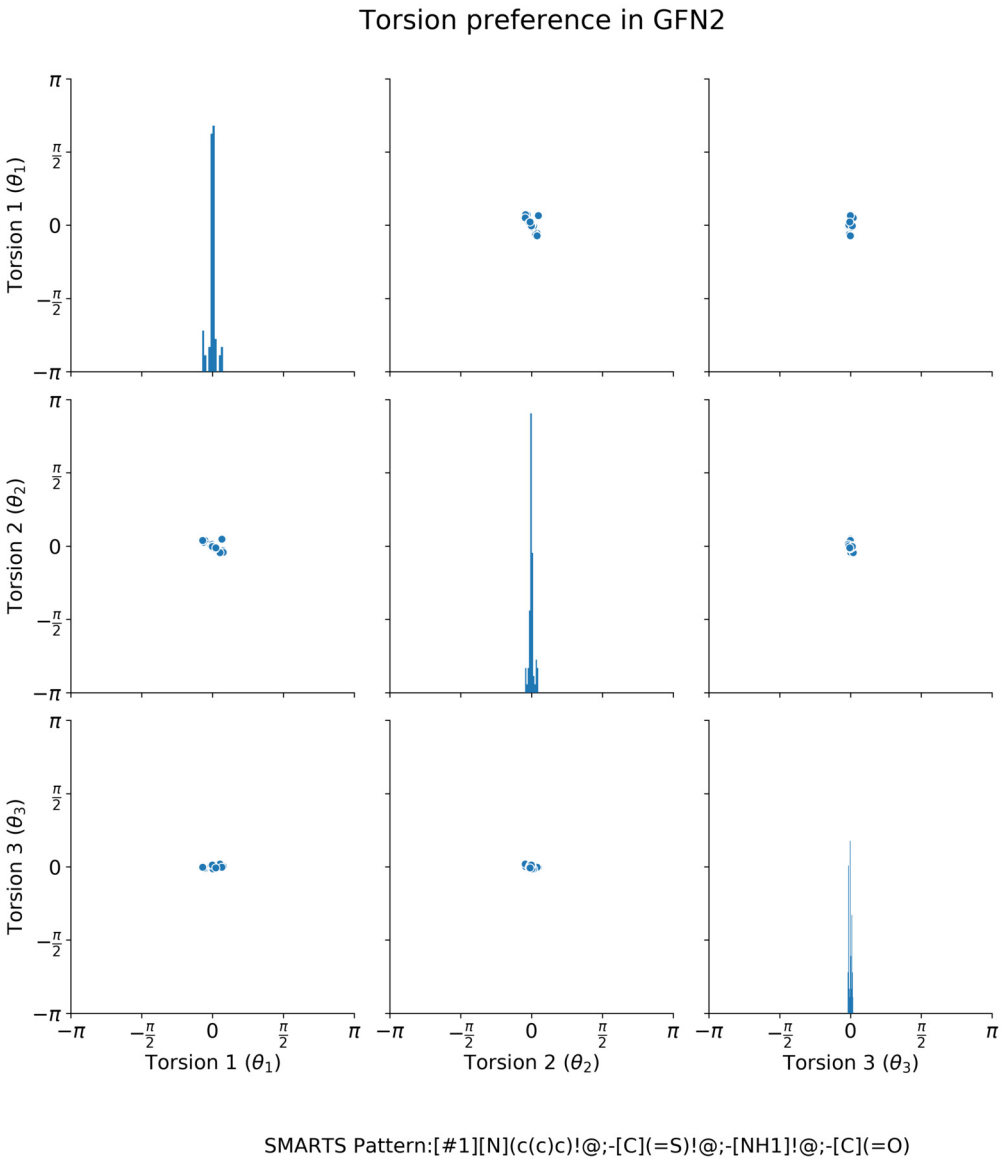
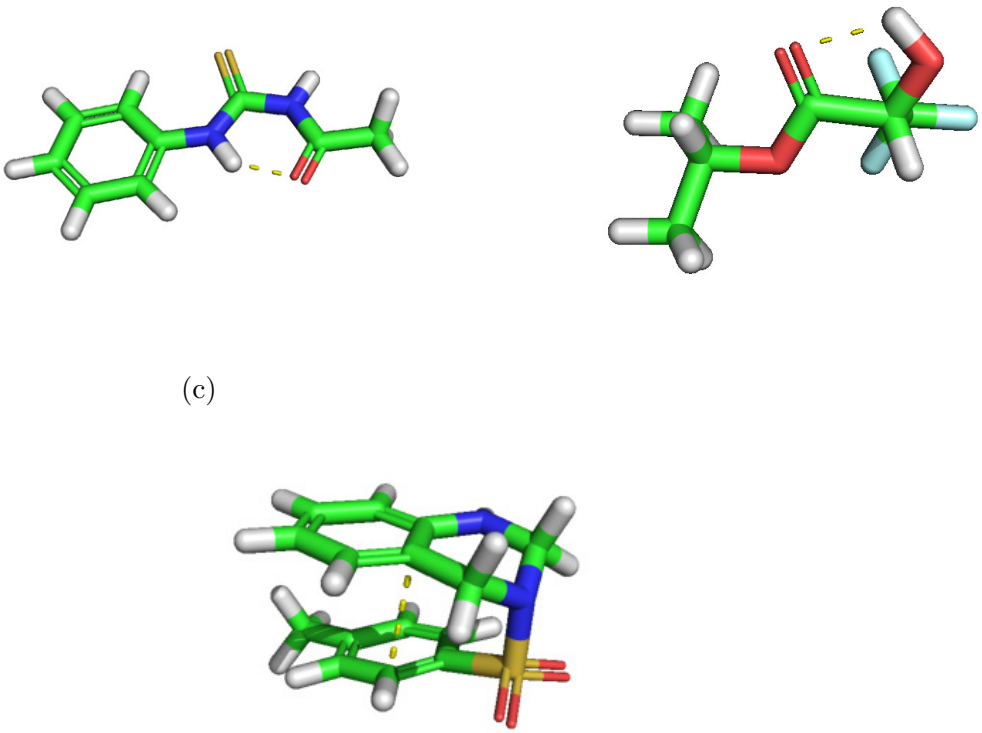


Figure 3.17: Higher order correlated torsion angles in the lowest energy (GFN2) conformation. Torsion angles were concentrated around 0, suggesting C=S and C=O were oriented in opposite directions and formed a pseudo six-membered ring. This conformation promoted C=O–H-N intramolecular hydrogen bond.

(a) (b)



(c)

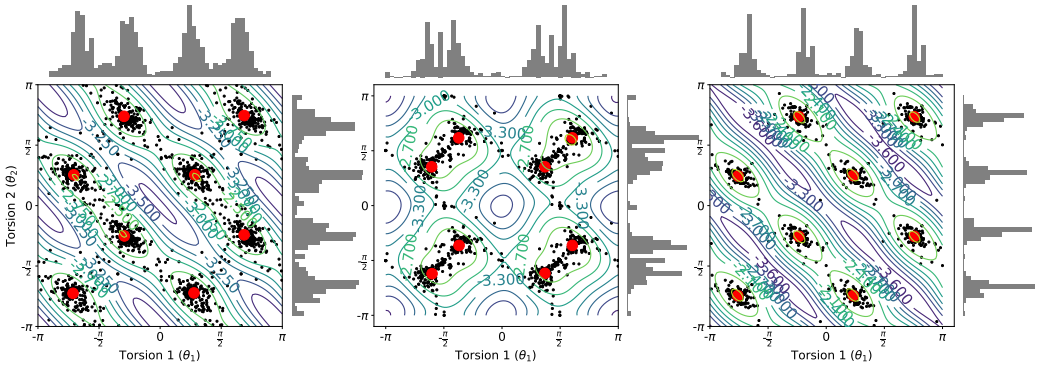
Figure 3.18: Intramolecular H-bonds are observed in patterns 2 and 16 (a) and pattern 15 (b); while intramolecular $\pi - \pi$ stacking is evident in pattern 17 (c).

For pattern 17, π–π stacking is evident: when aromatic rings are attached to both ends of the pattern, both rings prefer to interact with one another, see Figure 3.18c. It should be noted that the CSD Conformer Generator (Cole et al., 2018) considered 11 correlated torsions, but a simple clash term was used for all other interactions. Here, I used a more flexible approach that employs bivariate von Mises mixture models to fully describe the correlated torsions. Both favorable intramolecular interactions and unfavorable steric clashes can be described. It would also possible to expand this to a multivariate case (Mardia et al., 2008) in order to capture higher-order correlations as mentioned earlier.

I also noticed that the torsional preference of MMFF94 differed from the one in GFN2 and crystal structures in some of the patterns, as shown in Figure 3.19. It suggests that the classical force fields failed to capture all reflectional and rotational dependence between adjacent torsion angles. Moreover, in pattern 2, the *trans* configuration ($\theta_2 \approx \pm\pi$) were more frequently observed in crystal structures than the one in GFN2,

as illustrated in Figure 3.16a. Such discrepancy could be explained by formation of inter- and intramolecular hydrogen bonds. The *trans* configuration was in favor of the intermolecular hydrogen bonds of N-H and C=S group in crystal structures, while the *cis* configuration helped the formation of intramolecular hydrogen bonds.

(a)



(b)

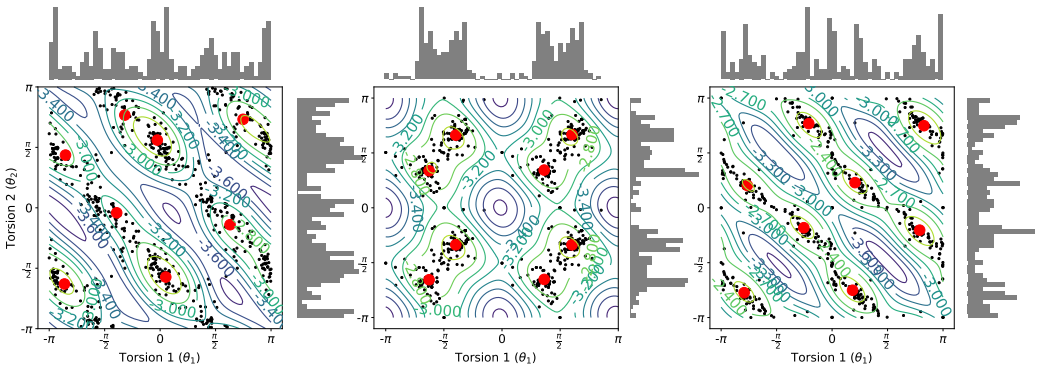


Figure 3.19: Mixture models for correlated torsions. The contour plot indicates the log density of the mixture model and the points (in red) mark the mean location for the components. The left, middle and right are the models derived from crystal structures, the lowest energy conformations from MMFF94, and GFN2 respectively. (a) Pattern 1; (b) Pattern 12. The conformations from MMFF94 failed the capture the reflectional dependence in crystal structures and GFN2 low energy structures.

Additionally, I computed the coverage of these nineteen correlated torsions patterns in different databases, including Platinum, COD and ChEMBL 25 (Gaulton et al., 2016). The SMARTS patterns library constructed in my analysis showed matches of 10%-15% organic molecules, which was noticeably higher than the patterns defined in CSD Conformer Generator (Cole et al., 2018) ($\sim 1\% - 4\%$). These results suggest that broader investigation of correlated torsion angles is warranted, despite the conventional assumption of each rotatable bond as an independent free rotor.

Table 3.7: Frequency of molecules with the presence of correlated torsion patterns, comparing this work to previous steric constraints (Cole et al., 2018) across various databases, including the Crystallographic Open Database (COD).

Dataset	Number of Molecules	% Matches (New)	% Matches (CSD)
Platinum	4,548	9.2	2.5
COD	110,623	13.5	1.6
ChEMBL 25	1,870,461	14.6	3.6

3.3.2.4 Computational Time

Similarly, I evaluated the computational time of the algorithm with proposed acquisition function. Figure 3.20 shows the average run time of BOA-EI and BOKEI with GFN2 energy function, varying the number of iterations (50, 100) and number of rotatable bonds (two to six rotatable bonds). Both computational cost increased as the number of rotatable bonds increased. The computational time also increased when BOKEI was used, but was primarily dominated by the number of conformers generated. Note that the current implementation can be further optimized by providing the gradient information of the acquisition function.

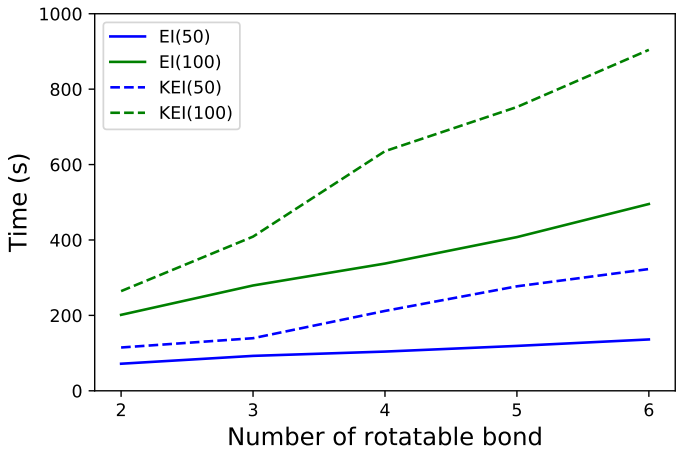


Figure 3.20: Average computational time for BOA-EI and BOKEI with GFN2 energy function, using different number of energy evaluations (50 and 100). The computational time increased as the number of rotatable bonds increased, but was dominated by the number of conformers sampled.

In theory, extra multiplication in the BOKEI acquisition function increases the computational complexity by $\mathcal{O}(mn)$, where m and n are the number of correlated torsion found in a molecule and the number of samples used in evaluating the acquisition

function respectively. The relative contribution to the computational time of the new acquisition function will be small, when a more accurate and computational expensive DFT or Hartree methods are used for energy evaluation.

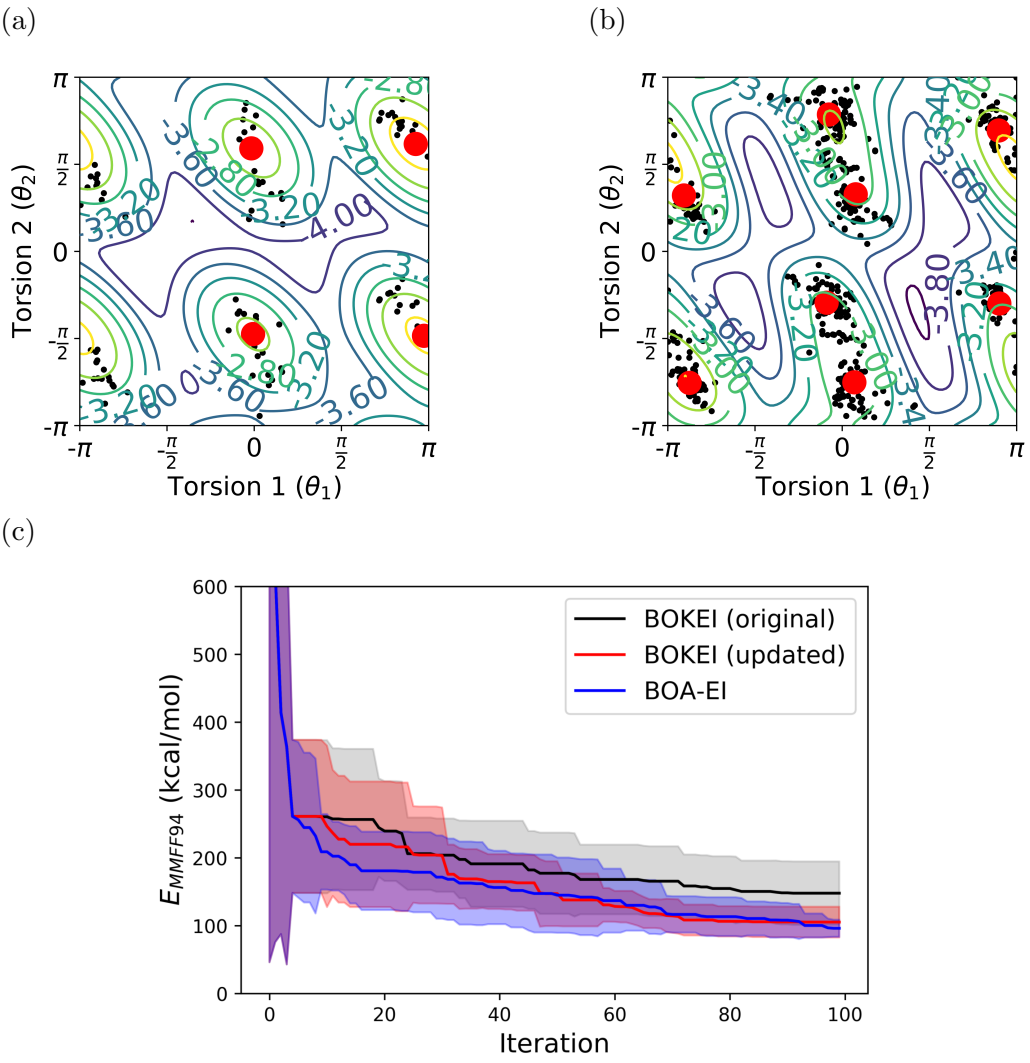


Figure 3.21: (a) Mixture model derived from the COD dataset. (b) Mixture model derived from COD and additional ChEMBL database. The contour plot indicates the log density of a mixture model and the points (in red) mark the mean location for the components. (c) Convergence plot. The search performance improved when updated prior was used.

3.3.2.5 Limitations

The lowest energy conformation of the molecules with up to five rotatable bonds in the COD datasets were considered, and the sampling of the low energy region of the

correlated torsion angles could be incomplete. This may lead to inferior performance of the algorithm, as illustrated in Figure 3.12b. It can be easily solved by increase sampling of the corresponding substructures from different databases, for example ChEMBL (Gaulton et al., 2016) and PubChem (Kim et al., 2018), and re-estimating the distribution. Figures 3.21a and 3.21b shows the difference between the original prior and the updated prior with additional observations from ChEMBL. The search performance improved with the update prior, as illustrated in Figure 3.21c.

3.4 Discussion

A local periodic kernel was used throughout the analysis. However, the L_2 norm used in the kernel is not an appropriate measure for angular variables, as discussed in Chapter 2.9. To overcome this issue, the L_2 norm can be replaced by an appropriate angular distance measure. Alternatively, one can project the torsion angle to the Cartesian coordinates, *i.e.* $\theta \mapsto (\cos\theta, \sin\theta)$, and the original kernel can be directly used. I anticipate the new kernel will fully capture the circular nature of the variables and lead to an improved surrogate model for the search.

On the other hand, the bivariate von Mises mixture models used in my analysis can be extended for general conformer sampling. For instance, the trained models (or statistical potentials) based on X-ray crystal structures can be used in the minimisation step in ETKDG framework, to better reproduce the conformations observed in a solid state. For substructures with small number of observations, the correlations between adjacent torsion angles may not be fully captured by the bivariate models. Alternatively, a meta-learning approach proposed by Ton et al. (2019) can be used. Instead of a joint density, it attempts to model the conditional distributions of a set of correlated torsions through a shared representation. By sharing information from other related pairs of correlated torsions through the representation, the approach can generalise the distribution with few observations. This meta-learning approach can potentially help discover more unexpected correlated torsions patterns for the sampling.

3.5 Summary

In this chapter, I demonstrated the effectiveness of Bayesian optimisation algorithm (BOA) in finding the lowest energy conformation of a molecule. It learned the most likely torsion angles by approximating the energy landscape with a surrogate model, Gaussian Process (GP), followed by an optimisation of an acquisition function to select next query conformation. Prior knowledge about the characteristic of the energy landscape, *i.e.* periodicity of the energy landscape, was encapsulated into the kernel function in GP. The sequential sampling strategy improved the approximation in each iteration, and returned the optimal solution(s) with high confidence. I showed that it required orders of magnitude fewer energy evaluations to reach the top candidates, compared to uniform random search and a systematic enumeration. The uncertainty of the optimal solutions reduced with the number of energy evaluations. The computational cost, however, increased at a cubic rate with the number of evaluations theoretically.

To address the high computational complexity, I incorporated the knowledge of the correlated adjacent torsion angles preferences through a modified acquisition function, namely Knowledge-based Expected Improvement (KEI), which is the product of a standard expected improvement and a collection of bivariate von Mises mixture models. The bivariate von Mises mixture models were used to capture the rotational and reflection dependencies between adjacent torsion angles, and biased the search towards low energy regions. This approach showed great improvement in finding the lowest energy conformation, especially in the early stage of the search, compared to the standard expected improvement. More importantly, this approach did not show substantial increase in the computational complexity, while improved the search performance. Additionally, I performed conformational analysis on a set of experimental determined structures, the lowest energy conformations from a force field MMFF94 and a semi-empirical energy GFN2, and I identified the deficiencies of the MMFF94 force filed. It failed to fully capture the reflection dependencies in some of the adjacent torsion angle pairs, as shown in experimental determined structures or the lowest energy conformations under GFN2 energy evaluations.

Lastly, I discussed some extensions of current works, including design of new kernels and density estimation techniques under low data regime.

Chapter 4

Understanding Ring Puckering in Small Molecules and Cyclic Peptides

Most of the work in this chapter has been reproduced from the following work:
(i) L. Chan, G. R. Hutchison, and G. M. Morris. Understanding Ring Puckering in Small Molecules and Cyclic Peptides. *ChemRxiv* (2020) doi:10.26434/chemrxiv.12999938

4.1 Background

Molecular rings play an important role in chemistry and biology, and their shapes are intimately linked to their physical and chemical properties. For instance, the glycosidases reactions heavily depend on their conformations (Davies et al., 2012). Beyond small rings, macrocycle conformations are crucial in host-guest chemistry and drug design. In host-guest chemistry, the conformational preferences of macrocyclic rings lead to selective complexation of organic ligands (Hancock and Martell, 1989; Gong et al., 2009; Begel et al., 2014). On the other hand, macrocycles including cyclic peptides have recently demonstrated their potential in modulating traditionally less druggable targets, *e.g.* mimicking protein-protein interactions (Driggers et al., 2008; Marsault and Peterson, 2011; Villar et al., 2014; Giordanetto and Kihlberg, 2014). The flexibility of cyclic molecules improves their chance to adopt favorable conformations that will bind to targets with flat surfaces. Despite the importance of ring conformations, most studies on ring conformations focus on small subsets, for example on carbohydrate rings (Altona and Sundaralingam, 1972; Ionescu et al.,

2005; Mayes et al., 2014; Alibay and Bryce, 2019), cycloalkanes (Anet and Cheng, 1975; Bocian et al., 1975; Pakes et al., 1981), and families of macrocycles involved in host-guest chemistry (Gutsche and Bauer, 1985; Al-Jallal et al., 2005; El-Azhary and Al-Kahtani, 2005), resulting in a lack of general understanding of ring conformational preferences, especially for medium-sized rings and macrocycles. I therefore carried out an extensive conformational analysis on a wide range of ring molecules, including cyclic peptides.

Flexible rings can adopt different conformations due to out-of-plane bending motions, caused by changes in the rotatable ring bonds, resulting in so-called ring *puckering*. Typically, the ring puckers can be classified into different *canonical forms*, and are usually low energy conformations; a classic example is the chair and boat conformations in 6-membered rings such as cyclohexane. These canonical forms are not "unique", as pseudo-rotation leads to multiple equivalent conformations, for example the $^4C_1$ and $^1C_4$ chair conformations in cyclohexane (Cremer and Pople, 1975; Ionescu et al., 2005). The pseudo-rotation and the coupled change in substituents orientation sometimes lead to diverse geometry, *i.e.* large root-mean-square-deviation (RMSD) in overall 3D conformations, asillustrated in Figure 4.1. It is therefore neceesary to sample ring conformations adequately to generate physically and biologically relevant conformational ensembles. In addition, there are several factors controlling the conformational flexibility of rings, including the presence of endocyclic double bonds (Flores-Ortega et al., 2007), the nature of substituents and the presence intramolecular interactions such as hydrogen bonds (Lyu et al., 2019). Lyu et al. (2019) recently showed that intramolecular hydrogen bonds restricted the pseudo-rotation path in deoxyribonucleosides, and the path characteristics depended on the strength of intramolecular interactions. In macrocycles, small structural modification to macrocycles, *e.g.* modification of exocyclic functionality, may lead to significant changes in conformation through hydrogen bonds and other intramolecular interactions (Appavoo et al., 2019). Such conformational changes are difficult to predict, as the correlation between ring bond rotations are not well studied.

A variety of coordinate systems have been developed to characterise ring puckers quantitatively. These techniques can be categorized into three general approaches. The first approach measures the perpendicular displacement of the ring atoms from a mean plane of the ring (Kilpatrick et al., 1947; Cremer and Pople, 1975); while the second approach makes use of triangular tessellation of the ring, and measures the associated angles between the reference plane and the triangular planes (Hill and

Figure 4.1: Two distinct pseudo-rotated conformations (green, orange) of (a) azepane and (b) methylcyclohexane. The best RMSD between conformations are 0.60 Å and 0.67 Å respectively. The RDKit implementation of RMSD calculation was used. Pseudo-rotation and the concomitant change in substituent orientation, *e.g.* axial and equatorial methyl groups in panel (b), can lead to diverse geometry.

Reilly, 2007). The last approach simply measures the ring torsion angles (Haasnoot, 1992); but this representation does not lend itself well to identifying pseudo-rotation. Methods used to analyse ring conformations based on perpendicular displacements of ring atoms such as Cremer Pople puckering coordinates (Cremer and Pople, 1975) are widely used in the community (Mayes et al., 2014; Paoloni et al., 2019). This representation has the advantage of using a reduced number of parameters, $N - 3$, to describe the geometry of an N-membered *monocyclic* ring. Hence, only two parameters are required to describe the conformational space of five-membered rings, and just three for six-membered rings. This representation has been also used as collective variables for the enhanced sampling of six-membered ring conformations in molecular dynamics studies (Sega et al., 2011).

To better understand ring conformational preferences, I extended the analysis to more complex ring systems, including larger sizes, bicyclic and polycyclic rings. I not only study their puckering preferences using Cremer-Pople puckering coordinates, but also identify the underlying constraints on their geometry and the change in substituent orientations upon puckering. More importantly, I have built quantitative models to convert from Cremer-Pople puckering coordinates to ring torsion angles, which thus allows us to understand the torsional changes upon pseudo-rotation. A novel knowledge-based conformational sampling scheme based on puckering parameters is also proposed. This approach allows efficient exploration of the conformational space, including the dominant canonical conformations and their associated pseudo-rotation.

I show that my sampling method can generate low energy ring conformations effectively.

4.2 Related Works

Although the general conformational preference of rings are not well understood, many efforts have been devoted to develop conformer sampling tools for ring conformers, especially macrocycles. Knowledge-based methods include CSD conformer generator (Cole et al., 2018), Conformator (Friedrich et al., 2019), and ETKDG (Wang et al., 2020), where endocyclic torsion potentials and ring templates are used in these methods. Molecular dynamic simulations and low mode dynamic simulations (Labute, 2010; Watts et al., 2014) can also be used.

Alternative methods such as inverse kinematics and ring breaking-and-stitching were proposed to sample ring conformation. Inverse kinematics method, BRIKARD (Coutsias et al., 2016), leverages techniques from kinematics and computational geometry for the sampling. It uses a recursive breadth-first construction to generate all sterically feasible ring combinations that are consistent with a given set of values of the sampled torsions. On the other hand, the ring breaking-and-stitching method breaks the rings into linear chains based on hierarchical rules, and dummy atoms are introduced. A given set of values of the allowed torsion angles that satisfy the ring closure constraints are used in the sampling. Local energy minimisation is usually performed after ring closure. Methods utilising ring breaking-and-stitching include PrimeMCS in Schrödinger Tools (Sindhikara et al., 2017), Conformator (Friedrich et al., 2019), and the conformer sampling tool in AutoDock (Forli and Botta, 2007).

4.3 Method

I first introduce the basic concept of Cremer Pople puckering parameters and substituent orientation angles, followed by an introduction to unique ring families (URFs) (Kolodzik et al., 2012) and the ring reconstruction procedures. After that, I build models to understand the relationship between puckering parameters, substituent orientation and torsion angles. I then discuss the metrics for model performance, followed by the implementation detail and the dataset used in this analysis.

4.3.1 Cremer Pople Puckering Parameters

Cremer and Pople (1975) proposed a coordinate system to describe the geometry of an N-membered *monocyclic* ring. It leverages the discrete Fourier transform technique to characterise the ring geometry with amplitudes and phase angles. Using this representation, pseudo-rotation in molecular rings can be completely elucidated. The mathematical details are discussed below.

Let the position of each atom, j, in an N-membered puckered ring be specified by the Cartesian coordinates (X_j, Y_j, Z_j); and let $\mathbf{R_j}$ be the corresponding position vector of ring atom j, with the origin, (0, 0, 0) as the geometrical center of the ring, such that it satisfies Equation 4.1:

$$\sum_{j=1}^{N} \mathbf{R_j} = \mathbf{0} \tag{4.1}$$

This constraint effectively suppresses translation of the planar reference. Two additional constraints, Equations 4.2 and 4.3, are imposed to suppress overall rotation of the planar reference about the x- and y-axes:

$$\sum_{j=1}^{N} z_j \cos\left(\frac{2\pi(j-1)}{N}\right) = 0 \tag{4.2}$$

$$\sum_{j=1}^{N} z_j \sin\left(\frac{2\pi(j-1)}{N}\right) = 0 \tag{4.3}$$

The orientation of the mean plane can now be determined by two vectors, $\mathbf{R}'$ and $\mathbf{R}''$, as in Equations 4.4 and 4.5. I denote the unit normal vector to the mean plane defined by $\mathbf{R}'$ and $\mathbf{R}''$ as $\mathbf{n}$, as defined in Equation 4.6, and the positive direction of $\mathbf{n}$ defines the "top" side of the ring.

$$\mathbf{R}' = \sum_{j=1}^{N} R_j \sin\left(\frac{2\pi(j-1)}{N}\right) \tag{4.4}$$

$$\mathbf{R}'' = \sum_{j=1}^{N} R_j \cos\left(\frac{2\pi(j-1)}{N}\right) \tag{4.5}$$

$$\mathbf{n} = \frac{\mathbf{R}' \times \mathbf{R}''}{|\mathbf{R}' \times \mathbf{R}''|} \tag{4.6}$$

Using the ring atom position vector and the unit normal vector $\mathbf{n}$, it is possible to compute the full set of displacements, z_j (for $j = 1, \ ..., \ N$), from the mean plane using the scalar products in Equation 4.7; this will also satisfy Equation 4.2 and 4.3 automatically:

$$z_j = \mathbf{R_j} \cdot \mathbf{n} \tag{4.7}$$

The general ring-puckering coordinates for an N-membered ring are calculated as follows:

If N is odd and $N >3$, I define q_m and ϕ_m using:

$$q_m \cos \phi_m = \left| \sqrt{\frac{2}{N}} \right| \sum_{j=1}^{N} z_j \cos \left(\frac{2\pi m(j-1)}{N} \right) \tag{4.8}$$

$$q_m \sin \phi_m = \left| \sqrt{\frac{2}{N}} \right| \sum_{j=1}^{N} z_j \sin \left(\frac{2\pi m(j-1)}{N} \right) \tag{4.9}$$

These formulae apply for $m = 2, \ 3, \ ..., \ \frac{(N-1)}{2}$. They represent a set of puckering coordinates with non-negative amplitudes, q_m, ($q_m \geq 0$), and phase angles, ϕ_m ($-\pi \leq \phi_m < \pi$).

If N is even, the coordinates in Equation 4.8 and 4.9 apply up to $m = \frac{N}{2} - 1$, but there is an additional puckering coordinate, as shown in Equation 4.10; note that the amplitude $q_{\frac{N}{2}}$ can take either sign:

$$q_{\frac{N}{2}} = \left| \sqrt{\frac{1}{N}} \right| \sum_{j=1}^{N} z_j (-1)^{j-1} \tag{4.10}$$

4.3.2 Ring Ordering

The Cremer Pople puckering parameter is atom-order dependent. The choice of the first atom and the atom numbering order, *i.e.* clockwise or anticlockwise, will affect the outcome. To overcome this ambiguity, I introduced a new atom numbering scheme

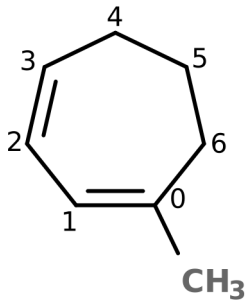


Figure 4.2: Atom ordering for 1-Methylcyclohepta-1,3-diene. There are two possible orderings, including: (0,1,2,3,4,5,6) and (3,2,1,0,6,5,4), based on the bond order criterion. Since atom 0 has one non-hydrogen (methyl) substituents, it has a higher connectivity than atom 3. Thus, the final atom ordering is (0,1,2,3,4,5,6).

that is based on the ring bond order, connectivity (number of substituents attached to the ring atoms), and the element types in the ring. The first atom is chosen with a bond that has the highest bond order, *i.e.* triple > double > aromatic > single. If the bond order of all the ring bonds is the same, such as in cyclohexane, then atomic connectivity is used to determined the first atom. Atoms with exocyclic double bonds have the highest priority, followed by atoms with two non-hydrogen substituents, followed by one non-hydrogen substituent, and finally no substituents. If the first atom cannot be determined by the rules above, then the atom with the minimum atomic number is assigned as the first atom. Otherwise, the first atom is assigned randomly when the ring is symmetric, such as in cyclohexane.

The direction of numbering can be determined by the sum of the ring bonds' bond orders, ring atoms' connectivities, and ring atoms atomic number in a ring path. The ring path takes $\frac{N+2}{2}$ atoms when N is even, and $\frac{N+1}{2}$ atoms when N is odd. For any N-membered ring, the path with the largest sum of bond orders is chosen. Similarly, the path with the largest connectivity sum is selected when there are multiple paths with same bond orders (an example is shown in Figure 4.2). If multiple possibilities exist, the path with the minimum sum of the atomic numbers from the first atom to atom $\frac{N+2}{2}$ when N is even (or atom $\frac{N+1}{2}$ when N is odd) is chosen.

For cyclic peptides, the volume of the amino acids side chain is taken into account for atom numbering, giving highest priority to the bulkiest side chain, tryptophan, and so on down to glycine. The amino acid ranks are listed in Table 4.1 (Zamyatnin, 1972).

Table 4.1: Amino Acid Volume Ranking.

Amino Acid	Volume (Å^3)	Rank
Tryptophan	227.8	1
Tyrosine	193.6	2
Phenylalanine	189.9	3
Arginine	173.4	4
Lysine	168.6	5
Isoleucine	166.7	6
Leucine	166.7	6
Methionine	162.9	8
Histidine	153.2	9
Glutamine	143.8	10
Valine	140.0	11
Glutamic acid	138.4	12
Threonine	116.1	13
Asparagine	114.1	14
Proline	112.7	15
Aspartic acid	111.1	16
Cysteine	108.5	17
Serine	89.0	18
Alanine	88.6	19
Glycine	60.1	20

4.3.3 Ring Substituent Orientation

To describe each substituent's orientation, I used a coordinate system defined by Cremer (1980), which is complementary to Cremer Pople puckering parameters. The mathematical details are given below.

Let $\mathbf{n}$ be the unit vector perpendicular to the mean plane (as defined in Cremer-Pople puckering parameters); let $\mathbf{s}_j$ be the unit vector pointing from a ring atom, j, to the corresponding substituent, S; let $\mathbf{u}$ be the vector that points from the geometrical center to the projection of the position of the ring atom onto the mean plane; and let $\mathbf{v}$ be the vector perpendicular to both $\mathbf{n}$ and $\mathbf{u}$. The substituent orientation can be described by two orientation angles, α and β, which describe the substituent position relative to the mean plane and to the ring center respectively, as shown in Figure 4.3.

The α angle is defined by Equation 4.11:

$$\cos\alpha = \mathbf{s}_j \cdot \mathbf{n} \tag{4.11}$$

α angle ranges from 0 to π radians, where 0 (or π) implies the vector, $\mathbf{s}$, is parallel to the z-axis, *i.e.* axial above (or below respectively) the mean plane; and $\pi/2$ means the vector $\mathbf{s}$ is perpendicular to the z-axis, *i.e.* equatorial orientation.

The β substituent angle is defined by Equations 4.12 and 4.13:

$$\mathbf{s}_j \cdot \mathbf{u}_j = \sin\alpha\cos\beta \tag{4.12}$$

$$\mathbf{s}_j \cdot \mathbf{v}_j = -\sin\alpha\cos\beta \tag{4.13}$$

The β substituent angle ranges from $-\pi$ to π radians, where 0 (or $\pm\pi$) indicates the substituent is outwardly (or inwardly) directed.

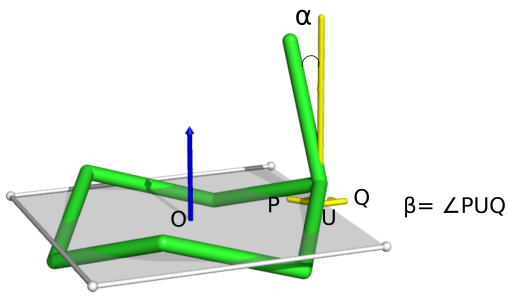


Figure 4.3: Definition of the substituent orientation angle α and β. Methylcyclohexane is used as an example with a mean plane (grey) cutting through the six-membered ring. The points P and U are projections of the methyl carbon and the ring atom that is attached to the methyl carbon onto the mean plane. The ring bonds are colored in green. O denotes the origin, which is also the geometrical center of the ring. The points O, U and Q are collinear. The β is defined by the angle between P, U and Q.

4.3.4 Unique Ring Families (URFs)

To study the puckering preferences of complex rings such as fused rings and spiro rings, I applied the concept of unique ring families (URFs) (Kolodzik et al., 2012) to decompose complex ring systems into multiple meaningful sub-groups. The puckering

parameters of these sub-groups and the orientation angles of the substituents attached to the sub-groups can be computed directly.

The calculation of URFs comprises two parts: (i) calculation of Relevant Cycles (RCs), where the RCs are defined as the union of all minimum cycle bases; and (ii) pairing of RCs if they are URF-pair-related (see Definition 1).

Definition 1 *Let C_1 and C_2 be two RCs in a molecular graph, G; then C_1 and C_2 are URF-pair related if and only if all of the following conditions hold:*

1. *$|C_1| = |C_2|$, i.e. the number atoms in each ring is the same;*
2. *$E(C_1) \bigcap E(C_2)$, i.e. the two rings share one or more bonds; and*
3. *There exists a set, S, of strictly smaller rings, c, in G such that $C_1 \bigoplus (\bigoplus_{c \in S} c) = C_2$.*

(a) (b)

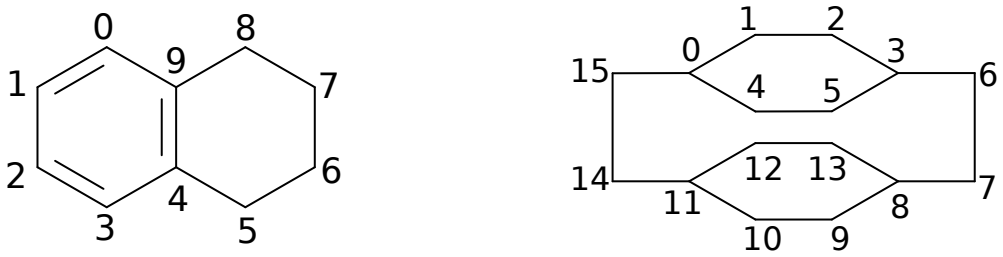


Figure 4.4: Example of unique ring families (URFs) calculation. (a) 1,2,3,4-tetrahydronaphthalene. There are two RCs in (a) $(0, 1, 2, 3, 4, 9)$ and $(4, 5, 6, 7, 8, 9)$. Since it does not exist any smaller rings satisfying definition 3, the two RCs are the two URFs in (a). (b) Tricyclo(8.2.2.24, 7)-hexadecane. There are six RCs: $(0, 1, 2, 3, 4, 5)$, $(8, 9, 10, 11, 12, 13)$,$(0, 1, 2, 3, 6, 7, 8, 9, 10, 11, 14, 15)$,$(0, 1, 2, 3, 6, 7, 8, 13, 12, 11, 14, 15)$, $(0, 4, 5, 3, 6, 7, 8, 9, 10, 11, 14, 15)$ and $(0, 4, 5, 3, 6, 7, 8, 13, 12, 11, 14, 15)$. There are three URFs in total. The two six atoms RCs form two URFs while the four RCs with 12 atoms are paired to form another URFs.

4.3.5 Reconstructing Cartesian Coordinates from Cremer-Pople Puckering Parameters

Cremer-Pople puckering parameters are uniquely defined and can be easily calculated from the Cartesian coordinates of the atoms in an N-membered ring. On the other hand, the Cartesian coordinates of any N-membered rings can be derived from its

$N-3$ puckering parameters and the additional $2N-3$ internal coordinates that describe the planar reference ring. The calculation comprises the following steps:

1. Calculation of z evaluations;
2. Projection of bond lengths and bond angles onto the mean plane;
3. Ring partition; and
4. Calculation of ring partition coordinates.

Step 1: Calculation of z elevations

Cremer Pople puckering parameters are required for the calculation, and the inversion formulae for an N-membered ring are given below:

$$z_j = \left|\sqrt{\frac{2}{N}}\right| \sum_{m=2}^{\frac{N-1}{2}} q_m \cos\left(\phi_m + \frac{2\pi m(j-1)}{N}\right), \qquad \text{if } N \text{ is odd} \tag{4.14}$$

$$z_j = \left|\sqrt{\frac{2}{N}}\right| \sum_{m=2}^{\frac{N}{2}-1} q_m \cos\left(\phi_m + \frac{2\pi m(j-1)}{N}\right) + \frac{1}{N} q_{\frac{N}{2}} (-1)^{j-1}, \qquad \text{if } N \text{ is even} \tag{4.15}$$

for ring atoms $j = 1, ..., N$.

Step 2: Projection of bond lengths and bond angles

Once the z_j coordinates have been determined, the bond lengths, r', and bond angles, θ', of the planar reference ring can be computed by projecting the N bond lengths, r, and $N-3$ bond angles, θ, of the puckered ring onto the mean plane. The N projected bond lengths, r', and $N-3$ projected bond angles, θ', suffice to determine the (x, y) coordinates of the projected atoms. The initial bond lengths and bond angles are listed in Appendix B, Tables B.1 and B.2, respectively. Note that the reference bond lengths and bond angles will vary with ring size. For ring atoms i, j, and k:

$$r'_{ij} = \left|\sqrt{r_{ij}^2 - (z_i - z_j)^2}\right| \tag{4.16}$$

$$\cos\theta'_{ijk} = \frac{(z_k - z_i)^2 - (z_k - z_i)^2 - (z_k - z_i)^2 + 2r_{ij}r_{jk}\cos(\theta_{ijk})}{2r'_{ij}r'_{jk}} \tag{4.17}$$

Step 3: Ring partition

To calculate the (x, y) coordinates of the planar reference ring, the ring is partitioned into three segments, S_1, S_2 and S_3, by inscribing a triangle inside the ring. The number of atoms, bond length, r, and number of bond angles, θ, in each fragment are listed in Appendix B, Table B.3. This partitioning keeps the error propagation due to coupling of the ring parameters to a minimum.

Step 4: Ring partition coordinate calculations

The (x, y) coordinates of the atoms of each segment can be calculated from the bond lengths r' and bond angles θ' values. In the next step, the coordinates of the vertices of the inscribed triangle are determined by lengths of the sides of the triangle, namely R_1, R_2, and R_3, which are equivalent to the distance between the terminal atoms of each segment. One point of this triangle is placed at the origin, and the second point on the positive x-axis. The coordinates of the remaining vertex of the triangle can be calculated easily using R_1, R_2, and R_3. In the remaining steps, the three segments are rotated such that the terminal atoms coincide with appropriate segments. Finally, the planar ring is translated such that the geometric center coincides with the origin, thus yielding the final set of Cartesian coordinates.

The procedure outlined can be applied to any N-membered monocyclic ring with or without symmetry. Note that this procedure is sensitive to the puckering parameters, bond lengths and bond angles in the planar reference ring. Poor parameter choice will lead to numerical error.

Here, I focused on generating ring conformations of monocyclic rings only. This framework can be extended in conjunction with URF decomposition to generate conformers for more complex rings, such as bicyclic fused rings and spiro rings.

4.3.6 Conformational Sampling of Rings

Kernel Density Estimation (KDE) was used to learn the ring puckering preferences. To facilitate the learning process, the Cremer-Pople parameters were mapped to Cartesian coordinates $(q_m \cos\phi_m,\ q_m \sin\phi_m)$ for the KDE calculation. A Gaussian kernel was used in the calculation. The samples drawn from the model were then converted back to puckering parameters, and returned distinct ring conformations

using the procedures outlined in Section 4.3.5. Once the ring backbone conformation is fixed, the ring substituent position can be updated accordingly using the relationship between endocyclic torsion angles and exocyclic torsion angles, Equation 4.19 in Section 4.3.7, with appropriate parameters (see Appendix B, Table B.7). Note that the exocyclic bond angles were kept fixed in the sampling. Note that this approach does not require force field minimization, although as discussed below, energy minimization can also improve cases where actual bond lengths or angles differ slightly from ideal geometry.

4.3.7 Connection between Ring Puckering, Substituent Orientation and Torsion Angles

To understand the change of substituent orientation upon pseudo-rotation, I proposed two models, Models 1 and 2, to predict the orientation angles, α and β. Both models share the same functional form, Equation 4.18, but the model parameters, A_i, $B_{i,m}$, $C_{i,m}$, D_i, $E_{i,m}, F_{i,m}$, and G_i are different. α_i and β_i denote the orientation angles of the substituents attached to ring atom, i; q_m and ϕ_m are the puckering amplitudes and phase angles respectively; $M = \frac{N-1}{2}$ when N is odd, otherwise $M = \frac{N}{2} - 1$; $\mathbb{1}_{N,2(M+1)}$ is an indicator function, and is equal to 1 when $N = 2(M+1)$; and N is the number of atoms in the ring. Note that a single model is not sufficient to describe substituent orientation, since it also depends on other factors such as the nature of substituents and the relative stereochemistry of the stereo-centers. Multiple sub-models are thus required.

Models 1 and 2:

$$\alpha_i, \beta_i = A_i + \sum_{m=2}^{M} B_{i,m} q_m \cos(\phi_m + \frac{2\pi m(i-1)}{N}) + \sum_{m=2}^{M} C_{i,m} q_m \sin(\phi_m + \frac{2\pi m(i-1)}{N}) + D_i \mathbb{1}_{N,2(M+1)} (-1)^{i-1} q_{\frac{N}{2}} + \sum_{m}^{M} E_{i,m} q_m^2 \cos(2(\phi_m + \frac{2\pi m(i-1)}{N})) + \sum_{m}^{M} F_{i,m} q_m^2 \sin(2(\phi_m + \frac{2\pi m(i-1)}{N})) + G_i \mathbb{1}_{N,2(M+1)} (-1)^{i-1} q_{\frac{N}{2}}^2 \quad (4.18)$$

Alternatively, the substituent orientation can be described by exocyclic torsion angles and bond angles. A linear model, Model 3, Equation 4.19, was proposed to describe the rotational relationship between the exocyclic torsion angle and the neighbouring

endocyclic torsion angle. The endocyclic torsion angles are defined by the four ring atoms, $(i-1,\ i,\ i+1,\ i+2)$; while θ_i^{exo} is the exocyclic torsion angle defined by the substituent atom, s_i attached to ring atom i, and the three ring atoms $(i,\ i{+}1,\ i{+}2)$, so the exocyclic torsion is defined by $(s_i,\ i,\ i{+}1,\ i{+}2)$. H_i and J_i are model parameters. Since the endocyclic and exocyclic share the same rotatable bond defined by ring atom $(i,\ i+1)$, the equation can be simplified to Equation 4.20, *i.e.* setting $J_i = 1$. It describes the rotation of the exocyclic torsion angle around the endocyclic torsion angle. Similarly, the parameter H_i depends on the nature of the substituents, and the relative stereochemistry of the stereo-center. Multiple sub-models are required.

Model 3:

$$\theta_i^{exo} = H_i + J_i\theta_i^{endo} \tag{4.19}$$

$$\theta_i^{exo} = H_i + \theta_i^{endo} \tag{4.20}$$

Torsion angles are an alternative way to measure or define ring geometries. They are widely used in conformational analysis of small rings. Inspired by de Leeuw et al. (1984), I proposed a linear model, Model 4, Equation 4.21, to convert the Cremer-Pople puckering parameters to endocyclic torsion angles for general N-membered ring. This model helps understand the torsional changes upon pseudo-rotation. Variables $q_m,\ \phi_m,\ \theta_i^{endo}$, and $\mathbb{1}_{N,2(M+1)}$ are defined as above; $L_i, P_{i,m}, Q_{i,m}, R_i$ are the model parameters.

Model 4:

$$\theta_i^{endo} = L_i + \sum_{m=2}^{M} P_{i,m}q_m\cos(\phi_m + \frac{2\pi m(i-1)}{N}) + \sum_{m=2}^{M} Q_{i,m}q_m\sin(\phi_m + \frac{2\pi m(i-1)}{N}) + R_i\mathbb{1}_{N,2(M+1)}(-1)^{i-1}q_{\frac{N}{2}} \tag{4.21}$$

For all models, a random sample of 50% of the data was used to estimate the model parameters, and the rest was used to assess the performance of the models. All model parameters were estimated by the least-squares method. The model parameters are listed in Appendix B.4 to B.7. The performance metrics are discussed in Section 4.3.8.

4.3.8 Metrics for Sampling and Model Performance

Two metrics were used to assess the quality of the generated conformations, namely heavy atom root-mean-square deviation (RMSD) and Torsion Fingerprint Deviation (TFD), as discussed in Chapter 2.5. The lowest energy conformation obtained from CREST was used as the reference conformation in each case. Symmetry was taken into account for the RMSD calculation.

Three metrics: squared circular correlation coefficient R^2_{circ}, the mean angular error (MAE), and standard deviation of the angular distance, were used to assess the performance of our proposed models. The angular distance (angular error) is defined by Equation 4.22, which was introduced in Chapter 2.9 Equation 2.37

$$d(\theta_{\text{predicted}}, \theta_{\text{actual}}) = \min(\theta_{\text{predicted}} - \theta_{\text{actual}}, 2\pi - (\theta_{\text{predicted}} - \theta_{\text{actual}})) \tag{4.22}$$

where $\theta_{\text{predicted}}$ and θ_{actual} are the predicted angles and actual angles respectively.

4.3.9 Ramachandran Plot and Eccentricity for Cyclic Peptides

To provide a better understanding of the ring geometry of cyclic peptides, I computed the (ϕ,ψ) torsion angles (Ramachandran et al., 1963). I also calculated the eccentricity, which is a measure of "roundness" of a ring (Wang et al., 2020). Eccentricity, e, is a non-negative real value that characterizes the shape of a conic section. A value 0 indicates a circle and 1 indicates an ellipse.

4.3.10 Implementation

RDKit (Landrum, 2018) was used to read in molecules, generate initial conformers for the simulation and write conformers. The implementation of RMSD and TFD in RDKit were used. RingDecomposerLib (Flachsenberg et al., 2017) was used to identify the relevant cycles and the unique ring families in a molecule. NumPy (van der Walt et al., 2011) was used to calculate the Cremer-Pople puckering parameters, orientation angles, model parameters of the proposed models, model performance metrics, and eccentricity of cyclic peptides. The implementation of kernel density estimation (KDE) in Scikit-Learn (Pedregosa et al., 2011) was used.

4.3.11 Data

Over 130,000 small molecules were selected from Crystallography Open Database (COD (Gražulis et al., 2009, 2012)) (63814 molecules) and ZINC database (Sterling and Irwin, 2015) (67009 molecules), including natural products and macrocycles. In addition, I generated a set of cyclic peptides (CP), including 8661 cyclic tetrapeptides (CTP) and 2249 cyclic pentapeptides (CPPs). The molecules from COD and ZINC contain carbon (C), nitrogen (N), oxygen (O), and sulphur (S) atoms in the ring, with rings having up to 20 atoms. The peptide datasets contain head-to-tail cyclic tetrapeptides and cyclic pentapeptides, *i.e.* cyclization from the N-terminus to the C-terminus, yielding a set of 12-membered and 15-membered rings. Their sequences are composed of fourteen of the twenty naturally occurring L-amino acids, see Table 4.2.

Table 4.2: Fourteen of the naturally occurring amino acids were used to generate the cyclic peptides dataset.

Property	Amino Acids
Special	Cysteine, Glycine
Charged	Histidine, Lysine, Aspartic Acid, Glutamic Acid
Polar Neutral	Serine, Threonine
Hydrophobic	Alanine, Valine, Leucine, Phenylalanine, Tyrosine, Tryptophan

For molecules from ZINC and cyclic peptides, ETKDG (Riniker and Landrum, 2015) in RDKit was used to generate initial conformations, while the molecules from COD used the X-ray crystal structures as initial conformations. I computed the lowest energy conformation for each molecule including the cyclic peptides using CREST (Grimme, 2019; Pracht et al., 2020), using the density functional tight binding quantum method GFN2 (Grimme et al., 2017; Bannwarth et al., 2019) for energy evaluation. The iMTD-GC workflow, as introduced in Chapter 3, was used in the search. Note that CREST may break the molecule into fragments in the outputs; when this happened, they were discarded in my analysis.

To demonstrate the effectiveness of using puckering preferences in sampling ring conformations, I selected twenty simple molecules, including monocyclic rings with substituents and endocyclic double bonds, see Appendix B, Table B.8.

4.4 Results and Discussion

4.4.1 Small and Medium-sized Ring Puckering Preferences

A relatively small number of conformational clusters were observed for 5- to 8-membered rings, reflecting their canonical conformations. For example, Figures 4.5a and 4.5b show the puckering preferences for flexible 6-membered rings with no endocyclic double bond. The peaks at $q_3 \approx \pm 0.6$ and $q_3 \approx 0$ correspond to the celebrated chair and boat conformations respectively. As expected, the chair conformation is more frequently observed than the boat conformation. The phase angle, ϕ_2, is uniformly distributed, suggesting free pseudo-rotation in both forms.

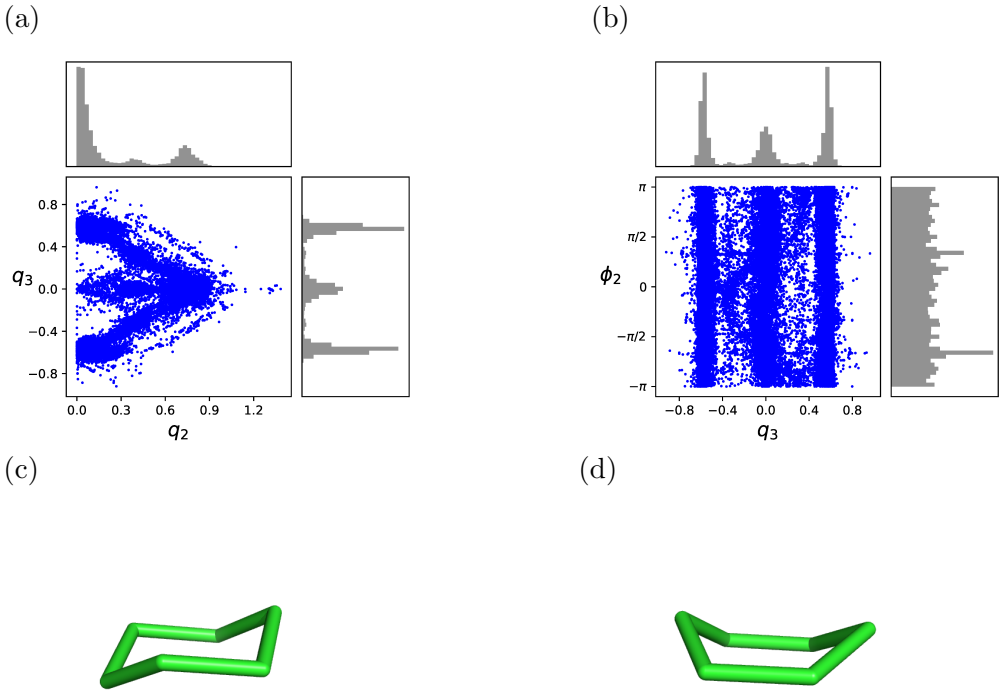


Figure 4.5: Puckering preferences of 6-membered rings with no endocyclic double bonds: (a) Coupled amplitudes q_2 and q_3. The mode at $q_3 \approx \pm 0.6$ and $q_3 \approx 0$ correspond to the chair and boat conformation respectively. (b) Coupling between amplitude q_3 and the phase angle ϕ_2. The pseudo-rotation is free in chair and boat conformations. (c) chair conformation ($q_2 = 0.03, q_3 = 0.57, \phi_2 = 0.00$); (d) boat conformation ($q_2 = 0.62, q_3 = 0.01, \phi_2 = 0.00$).

In contrast, the presence of endocyclic double bonds or shared aromatic bonds restrict both puckering and pseudo-rotation, as shown in Figures 4.6a and 4.6b. The puckering

amplitude, q_3, and phase angle, ϕ_2, exhibit a sinusoidal relationship in Figure 4.6b. These relationships hold for both simple monocyclic rings, complex bi- and polycyclic rings.

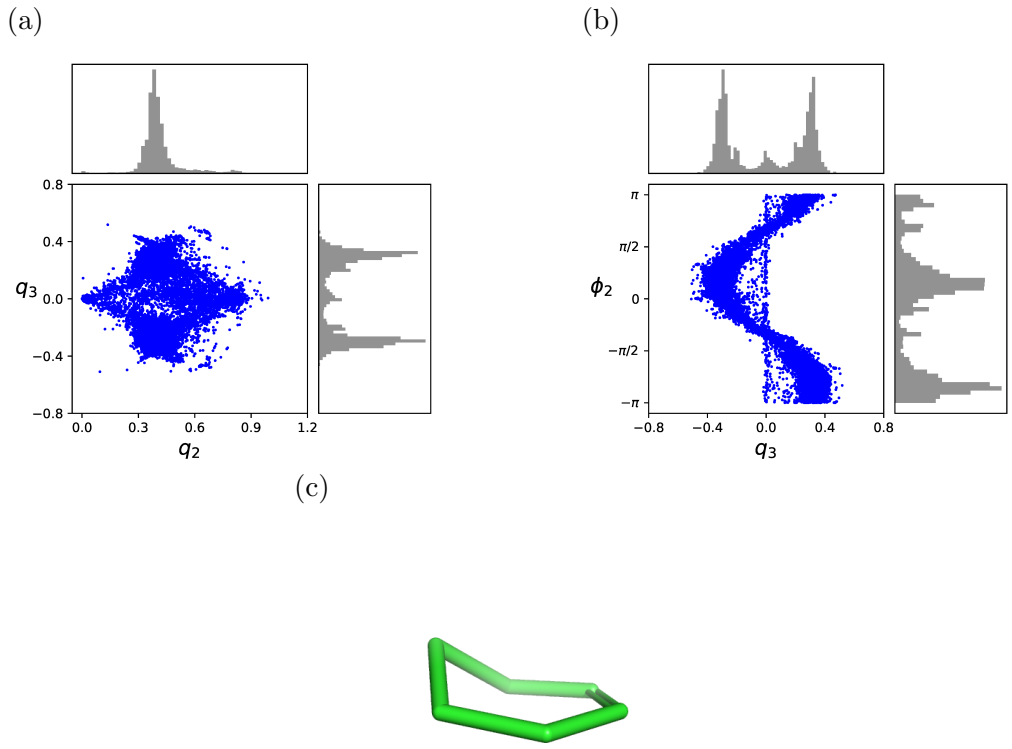


Figure 4.6: Puckering preferences of 6-membered rings with one endocyclic double bond, including shared aromatic bond: (a) Coupling between amplitudes q_2 and q_3; (b) Coupling between amplitude q_3 and phase angle ϕ_2. The pseudo-rotation is restricted by the presence of double bond, and exhibit a sinusoidal relationship with amplitude q_3. (c) half chair conformation ($q_2 = 0.45, q_3 = 0.28, \phi_2 = -1.92$).

For 7- and 8-membered rings, an additional phase angle, ϕ_3, is required. Phase-phase couplings are evident in some conformational clusters. For example, Figure 4.7a shows that there are three conformational clusters in 7-membered rings with no endocyclic double bonds. The predominant conformations are the twist-chair and chair conformations, as illustrated in Figures 4.7c and 4.7d. The puckering amplitudes (q_2, q_3) fall into a narrow range and the pseudo-rotations are restricted in this region, as shown in Figure 4.7b. The phase angles ϕ_2 and ϕ_3 are strongly coupled, and they are marginally uniformly distributed. This coupling suggests the minimum energy pathway of the chair-twist-chair pseudo-rotation. As suggested by Bocian et al. (1975),

the pseudo-rotation map can be approximated by Equation 4.23, with varying fixed values (ϕ_2^*, ϕ_3^*), slopes (K_2, K_3), and an angle ϕ $(-\pi \leq \phi < \pi)$. In this case, $K_2 = 3$ and $K_3 = 1$. The pseudo-rotation pathways are valid for all monocyclic, bicyclic and polycyclic rings with heteroatoms.

$$\begin{aligned} \phi_2 &= \phi_2^* + K_2\phi \\ \phi_3 &= \phi_3^* + K_3\phi \end{aligned} \tag{4.23}$$

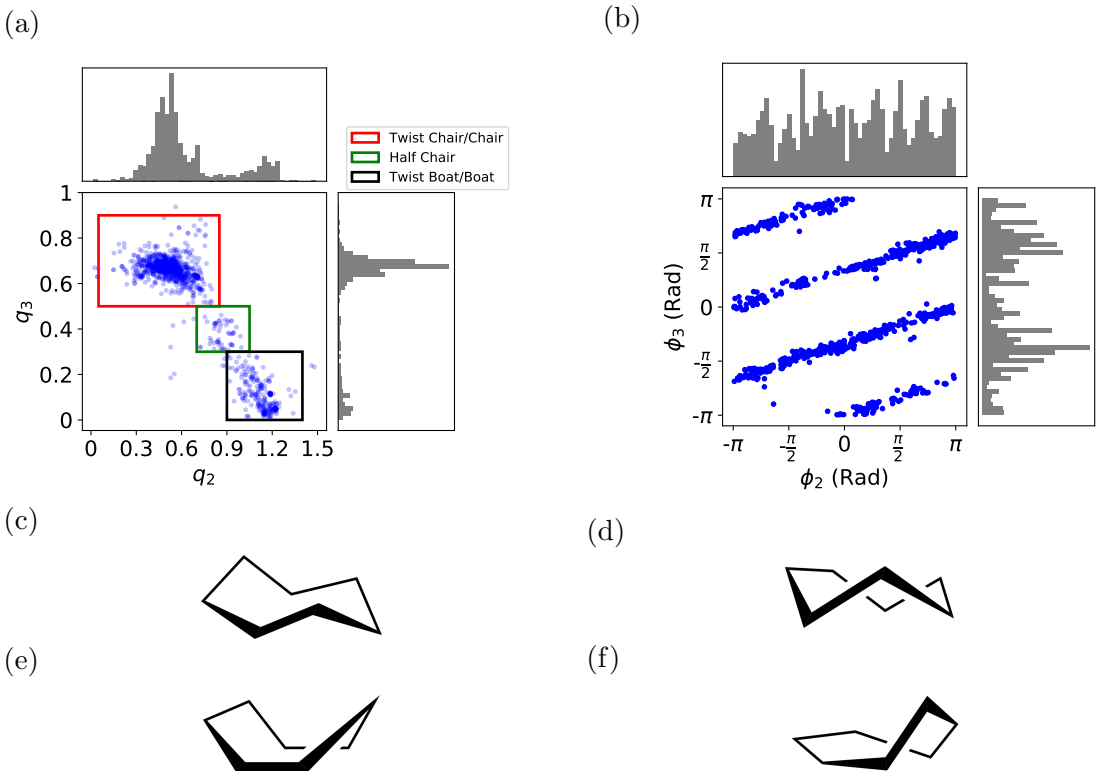


Figure 4.7: Puckering preferences of 7-membered rings with no endocyclic double bonds. (a) Coupling between puckering amplitudes q_2 and q_3. Twist-chair and chair conformations (indicated by red box) are frequently observed in the lowest energy conformation, followed by boat and twist boat conformation (indicated by black box). The half chair (indicated by green box) is the transition structure from chair to boat, and it is occasionally observed. Note that the color boxes only show the coarse boundary of the conformational clusters. (b) Coupled phase angles of the chair and twist-chair conformations, as indicated by the red box in (a). Example conformation of cycloheptane. (c) chair; (d) twist chair; (e) boat; (f) half chair.

Bulky substituents and adjacent rings often induce significant steric clashes, and result in concomitant changes in conformational preferences. The increase in amplitude q_2

and decrease in amplitude q_3 indicate a conformational change from chair to half chair ($0.7 < \phi_2 < 1$) and boat conformations ($\phi_2 > 1$). The pseudo-rotations are free in these clusters, *i.e.* the phase angles are randomly distributed.

For 8-membered rings, the couplings of amplitudes and phase angles vary between clusters. In particular, the boat-chair conformation shows strong phase-phase couplings, as illustrated in Figure 4.8a. Similarly, the pseudo-rotation map in Equation 4.23 can be applied, with different model parameters.

(a) (b)

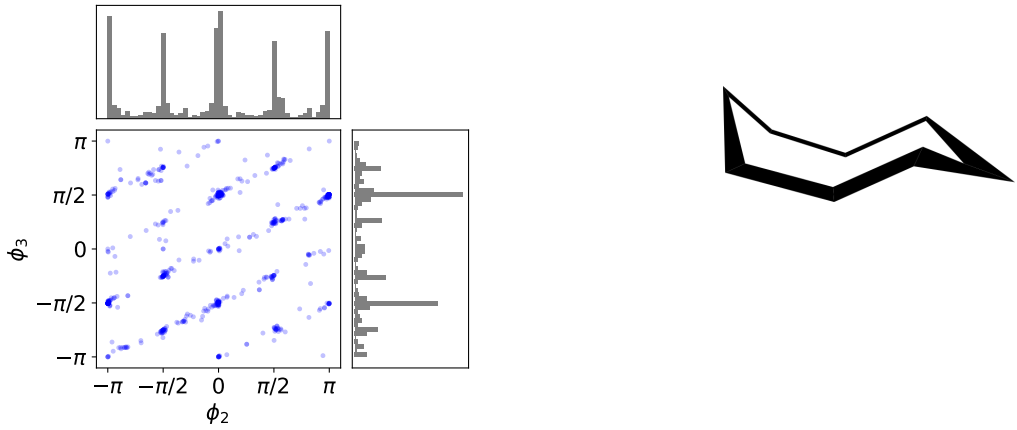


Figure 4.8: (a) Pseudo-rotation map of the boat-chair conformation in 8-membered rings with no endocyclic double bonds. (b) Example of the boat-chair conformation of cyclooctane.

4.4.2 Effect of Endocylic Double Bonds

To assess the effect of endocyclic double bonds on the conformational preferences of rings, I selected 7-membered rings with one and two endocyclic double bonds. I further separated the observations by the location of endocyclic double bonds. Figure 4.9a shows three conformational clusters in seven membered-rings with single endocyclic double bond, corresponding to the chair, half-chair and boat conformations, which are the same as the case without double bonds. However, the population of chair conformations decreases, while the population of half chair and boat conformations increase. The pseudo-rotations in all three clusters are restricted, as illustrated in Figures 4.9b to 4.9d. In the chair and twist-chair region, the phase angle, ϕ_3, is relatively fixed with small variations in phase angle, ϕ_2, while in the boat and twist-boat region, the phase angle ϕ_2 is fixed while the phase angle ϕ_3 varies. The half

chair conformation exhibits strong coupling between phase angles.

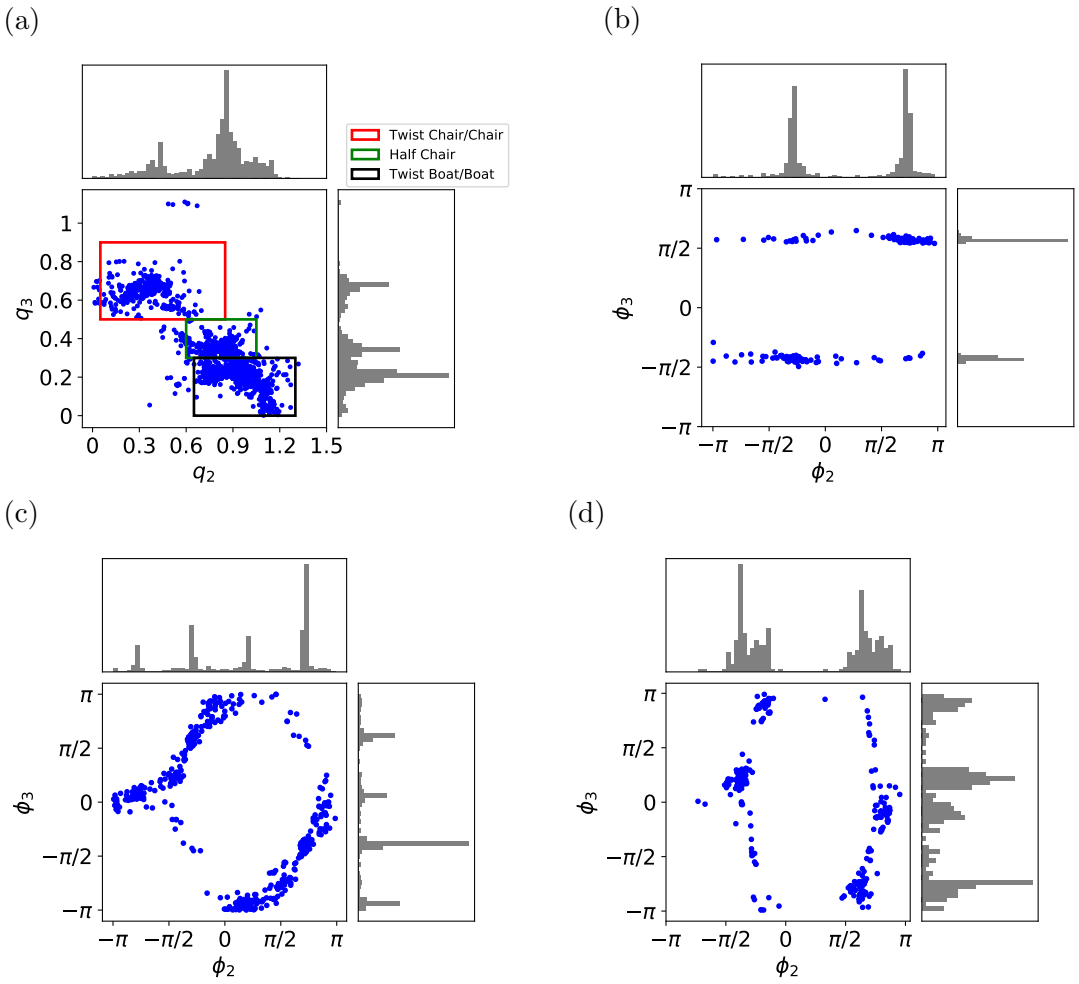


Figure 4.9: Puckering preferences of 7-membered rings with one endocyclic double bond. (a) Puckering amplitudes (q_2, q_3). Similarly, the red box indicates the chair/twist chair conformations; green box indicates half chair conformations, and black box indicates the boat conformations. Coupled phase angles in (b) chair/twist chair conformation; (c) half chair; and (d) boat/twist boat conformation. The pseudorotation are restricted in all three clusters.

As the number of endocyclic double bonds increases, the number of degrees of freedom of the ring system decreases. The location of the double bonds strongly influences the puckering preferences, as shown in Figure 4.10. The double bonds in 1,3-cycloheptadiene and 1,4-cycloheptadiene-like structures (Figure 4.10a and 4.10c respectively) impose different steric constraints, and lead to contrasting phase-phase coupling.

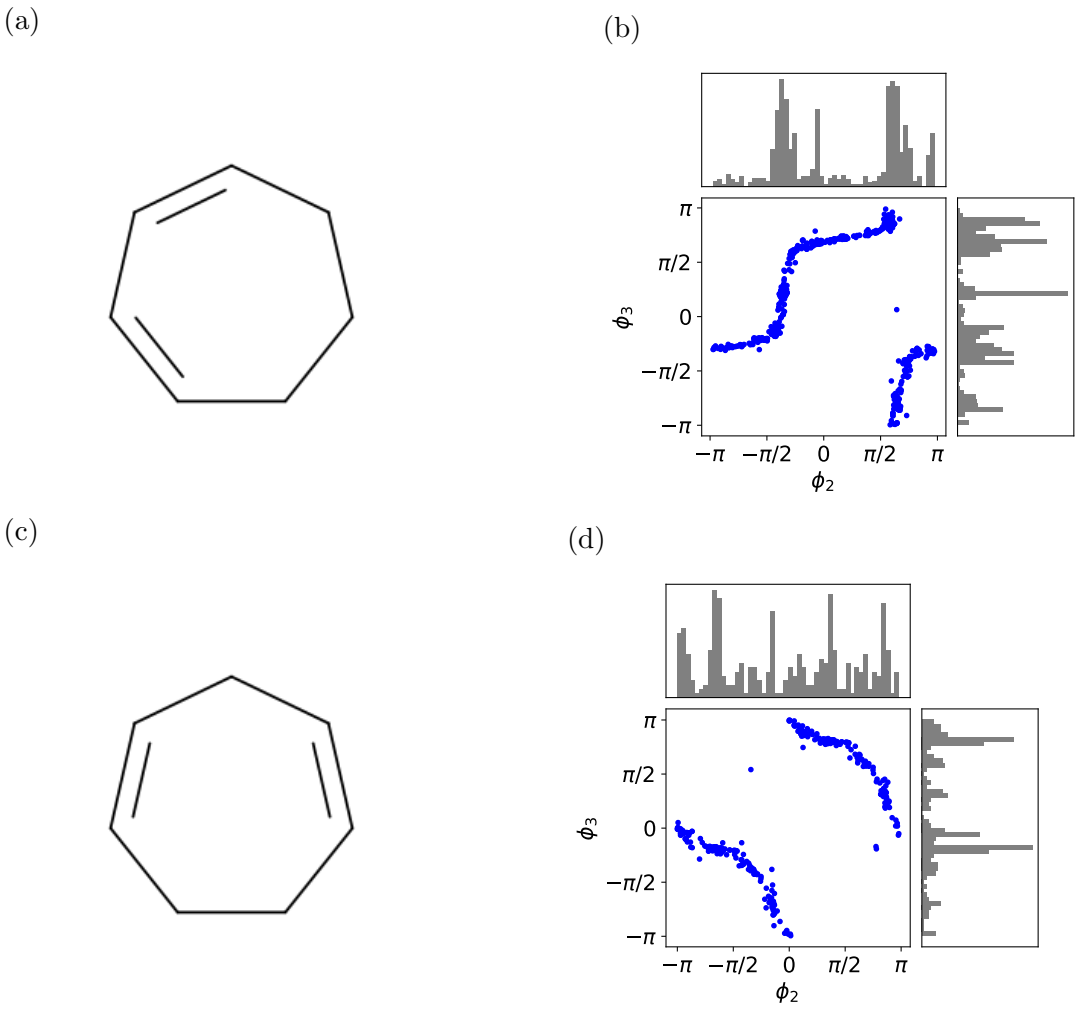


Figure 4.10: 7-membered rings with two endocyclic double bonds and their associated phase angles coupling. (a) 1,3-cycloheptadiene, and (b) the highly-coupled phase angles of the low energy conformations observed in 7-membered rings with double bonds at the 1 and 3 positions. (c) 1,4-cycloheptadiene, and (d) again, the highly-coupled phase angles of the low energy conformations observed in 7-membered rings with double bonds at the 1 and 4 positions. Monocyclic, bicyclic and polycyclic rings are all included in our analysis, and it should be noted that the double bond could also be a shared aromatic bond. The relative location of the endocyclic double bonds imposes different constraints on the system, and results in visibly different phase-phase couplings.

For larger rings, the number of conformational clusters increases, while the coupling between puckering amplitudes and phase angles becomes more complex. It should be noticed that small local structural changes may result in significant changes in conformation through transannular repulsion and intramolecular interactions. To gain further insight into long range coupled ring bond torsion angles, I performed

cluster analysis on a set of cyclic peptides.

4.4.3 Cyclic Peptides

Peptide cyclization imposes additional constraints on the system, and thus reduces the thermally-accessible conformational space of the resultant cyclic peptides relative to their linear counterparts (Edman, 1959). There are several factors governing the backbone conformation of cyclic peptides, including the size and properties of the amino acid side chains, presence of N-methylation, and formation of γ- and β-turns (Loiseau et al., 2003). Analysing the puckering preferences helps us to understand the relative influence of these factors.

The configuration of the amide bonds provides important information to determine the dominant backbone conformation adopted by the cyclic peptides. The partial double bond character of the carbon-nitrogen bond in amide bonds renders them planar, resulting in either *cis* (C) or *trans* (T) amides. I can thus classify the conformations based on the sequence of *cis*- or *trans*-amide bonds, as described in (Loiseau et al., 2003), for example, for cyclic tetrapeptides, all-*cis* ("CCCC") or all-*trans* ("TTTT") amides. Typically, the *trans*-amide bond is preferred in acyclic peptides, large cyclic peptides and proteins. Figures 4.11a and 4.11b, however, show that the *cis*-amide bond is preferred in both cyclic tetrapeptides and cyclic pentapeptides. In small cyclic peptides, high ring strain reduces the energy barrier between *cis* and *trans* isomers. All-*trans* and single-*cis* (CTTT and CTTTT) configurations are less favored in both tetra- and pentapeptides due to high transannular strain, and they exist only with explicit stabilization from one or more intramolecular hydrogen bonds. Such stabilization leads to γ- turns in cyclic tetrapeptides, and γ- and β-turns in cyclic pentapeptides, as reflected by their Ramachandran (ϕ,ψ) dihedral angles in Figure 4.12. The puckering amplitudes and phase angles are thus highly restricted in such conformational clusters. It should be noted that these turns are favored by the *in vacuo* calculation, and may not reflect the conformations observed in solution. The positional preferences of amide carbonyl groups are key to understanding the formation of such intramolecular hydrogen bonds, which will be discussed in next section.

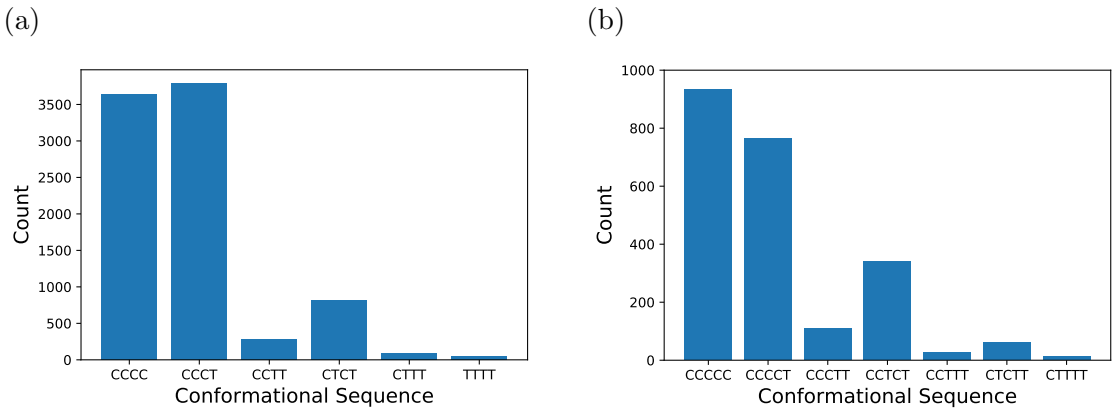


Figure 4.11: Conformational sequence preferences in (a) cyclic tetrapeptide and (b) cyclic pentapeptides. *cis* amide bond is favored in cyclic tetra- and pentapeptides.

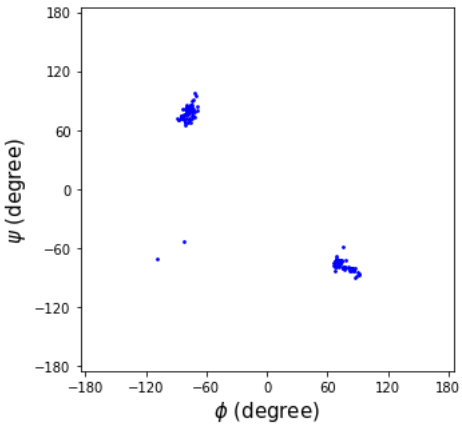


Figure 4.12: Ramachandran plot showing the backbone (ϕ,ψ) torsion angles at the α-carbon atoms in $i+1$-th amino acid of cyclic tetrapeptides with all-*trans*, TTTT, amide configuration.

Main chain-main chain intramolecular interactions were not observed in cyclic tetrapeptides with two or more *cis*-amide bonds; nor were they seen in cyclic pentapeptides with three or more *cis*-amide bonds. Transannular repulsion, main chain-side chain, and side chain-side chain intramolecular interactions appear to be the major driving forces behind the conformational preferences seen in these cases. Small structural modifications, such as the change of amide bonds and/or side chain orientations, may induce significant steric clashes, and lead to conformational switching. Here, I followed the nomenclature used by Loiseau et al. (2003), where the orientation of amide carbonyl is denoted by "U" when it is oriented above the mean plane; while it is denoted by "D" when it is oriented below the mean plane. Figures 4.13a and 4.13b

show the puckering amplitude preferences of two canonical conformations in all-*cis*-amide cyclic tetrapeptides, and they differ by the orientation of one amide bond. The two canonical forms (CCCC-DDDD and CCCC-UDDD) exhibit distinct phase-phase couplings, as illustrated in Figures 4.13c to.4.13f. Similar phenomena are observed in cyclic pentapeptides. Furthermore, the formation of main chain-side chain interactions and/or side chain-side chain interactions give rise to two sub-clusters within same configuration (CCCC-DDDD) with diverse geometries, as illustrated in Figures 4.14a and 4.14b. The eccentricty values of the two sub-clusters are shown in Figure 4.14c The orientation of the side chain C_β atoms play important roles in the formation of these interactions.

(a)

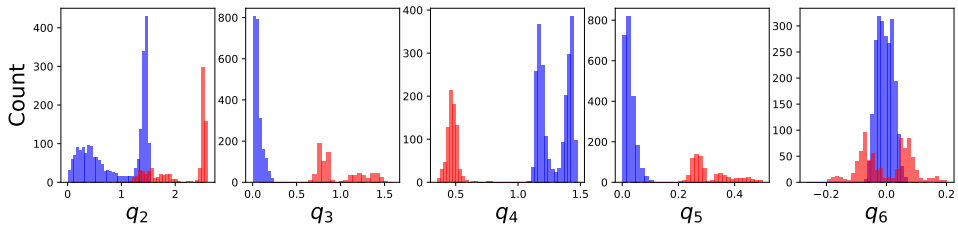
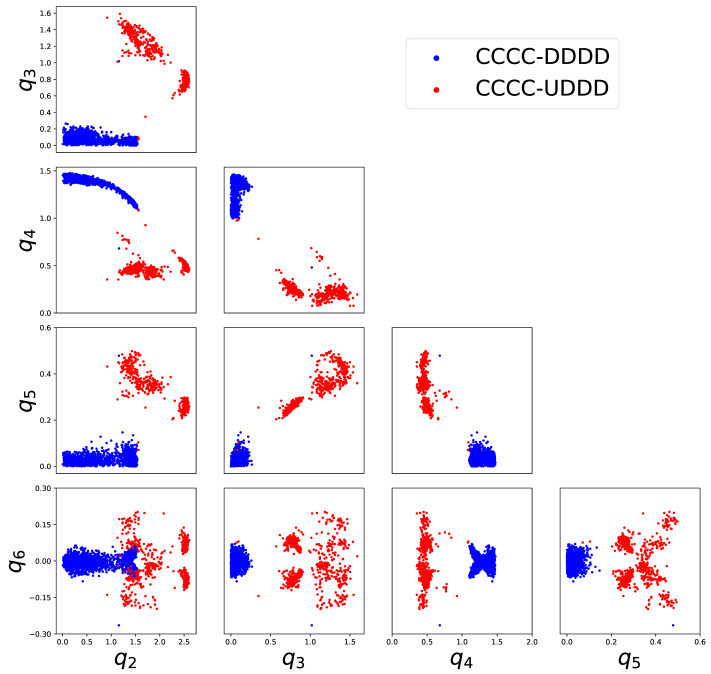


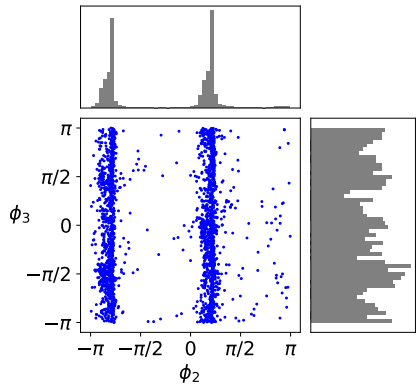
Figure 4.13: (a) Marginal distribution of the ring puckering amplitudes preferences $(q_2, q_3, q_4, q_5, q_6)$ for two types of all-*cis* conformation in cyclic tetrapeptides (CCCC-DDDD is colored blue and CCCC-UDDD is colored red). The two clusters are defined by the α orientation angle of the amide carbonyl oxygen, where U indicates $\alpha < \frac{\pi}{2}$, and D indicates $\alpha > \frac{\pi}{2}$. In panel (a), two modes are observed in puckering amplitudes q_2 and q_4, indicating presence of multiple sub-clusters. (b) Pairwise joint distribution of the ring puckering amplitudes preferences $(q_2, q_3, q_4, q_5, q_6)$ for two conformational clusters of all-*cis* conformation in cyclic tetrapeptides. The puckering preferences of CCCC-DDDD conformations are more concentrated than the one in CCCC-UDDD conformations. ϕ_2 and ϕ_3 phase-phase couplings in (c) CCCC-DDDD conformation and (d) CCCC-UDDD conformation. ϕ_2 and ϕ_5 phase-phase couplings in (e) CCCC-DDDD conformation and (f) CCCC-UDDD conformation.

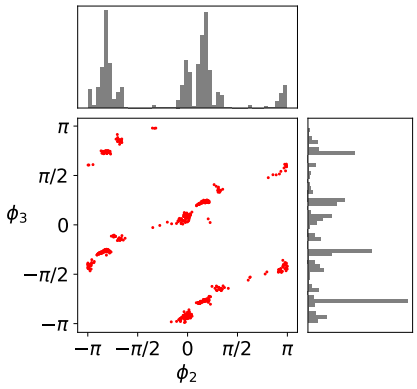
(b)



(c)

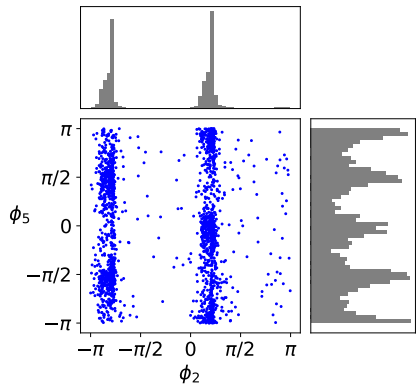
(d)





(e)

(f)



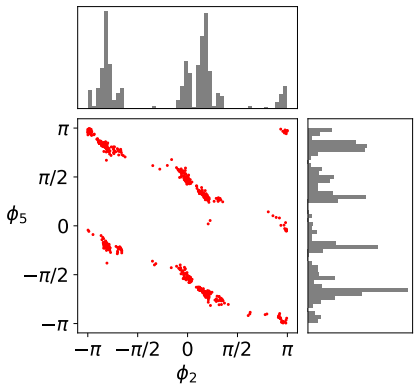


Figure 4.13: (Continued)

(a) (b)

(c)

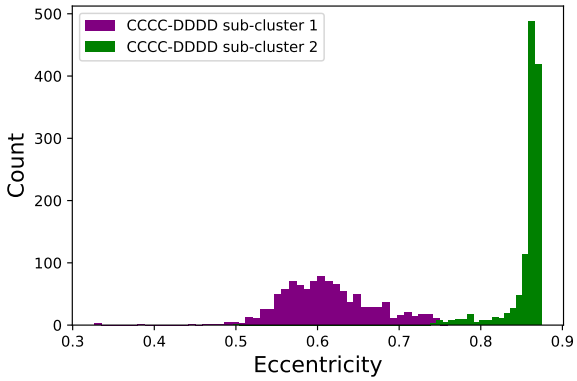


Figure 4.14: Example conformation from CCCC-DDDD (a) sub-cluster 1; (b) sub-cluster 2. Hydrogen atoms and side chains are not shown in panel (a) and (b). (c) Ring eccentricity values for two sub-clusters of CCCC-DDDD conformations, as colored by purple (sub-cluster 1) and green (sub-cluster 2). The main chain-side chain and side chain-side chain intramolecular interactions give rise to the diverse geometries.

To further understand the cyclic backbone conformation, I calculated the Ramachandran (ϕ,ψ) dihedral angles and the eccentricity of the backbone. The Ramachandran plots in Figures 4.15a and 4.15b show that the (ϕ, ψ) angle preferences of cyclic tetrapeptides and pentapeptides are similar to the standard secondary structures observed in proteins. Figure 4.15c and 4.15d show contrasting eccentricity values between clusters, for example all-*trans* cyclic tetrapeptides give a mode at 0.3, while alternating CTCT cyclic tetrapeptides give a mode at 0.8, indicating diverse geometries between clusters.

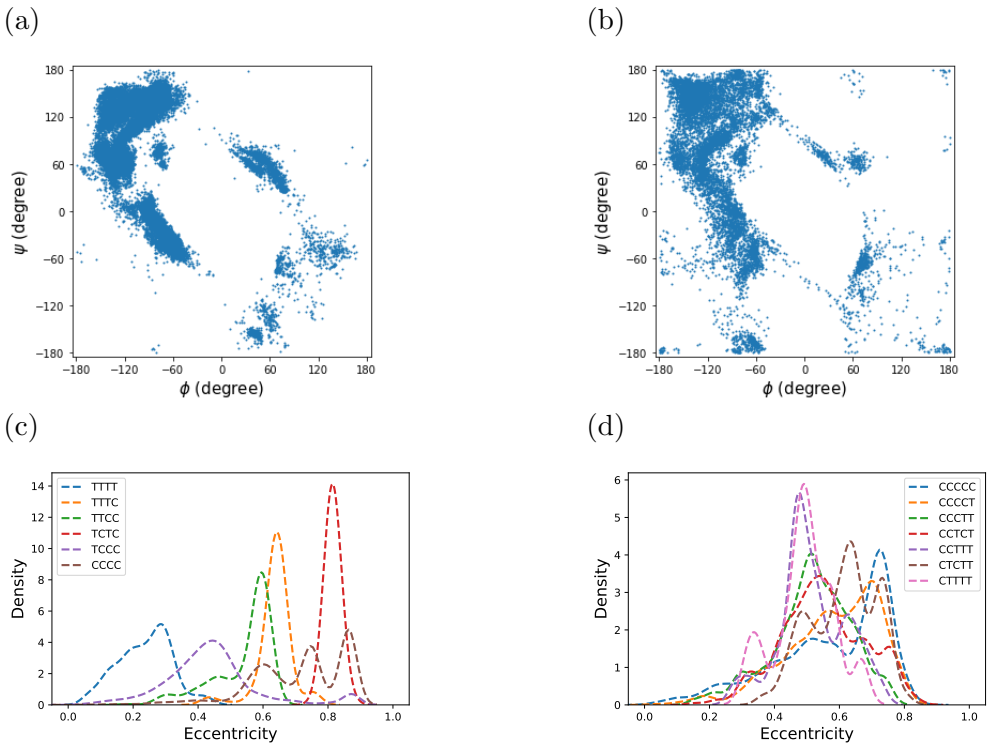


Figure 4.15: Ramachandran plots for (a) cyclic tetrapeptides and (b) cyclic pentapeptides. The (ϕ, ψ) angles preferences are quite similar to the standard secondary structures observed in proteins. Eccentricity values in (c) cyclic tetrapeptides and (d) cyclic pentapeptides. The change in amide configurations lead to diverse geometries.

I thus have shown that Cremer-Pople puckering parameters are a useful representation to understand ring puckering for both small rings and macrocycles including cyclic peptides, and analyzed the associated effects of endocyclic double bonds on ring puckering. I have also revealed the influence of configuration and orientation of amide on ring geometries. To gain further insights, I will examine the substituent orientations and its relationship to puckering preferences.

4.4.4 Effects of Substituent Orientation and Functionality

The size and functionality of substituents are two of the key factors determining the ring geometries, and their effects vary with ring size. I thus separated the lowest-energy conformations according to ring sizes: small (5- and 6-membered) rings, medium (7- to 11-membered) rings, and macrocycles (12-membered or larger rings). In particular, I assessed the following substituent types: hydroxyl group, alkoxy group, methyl group, carbonyl group, halogens and bulky substituents. The bulky

substituents are defined by the following SMARTS pattern: [CX4;R]!@;-[CX4H0, CX4H1,CX4H2], where the substituent carbon consists of at most two hydrogen atoms, and is attached to a carbon ring atom with a single non-ring bond.

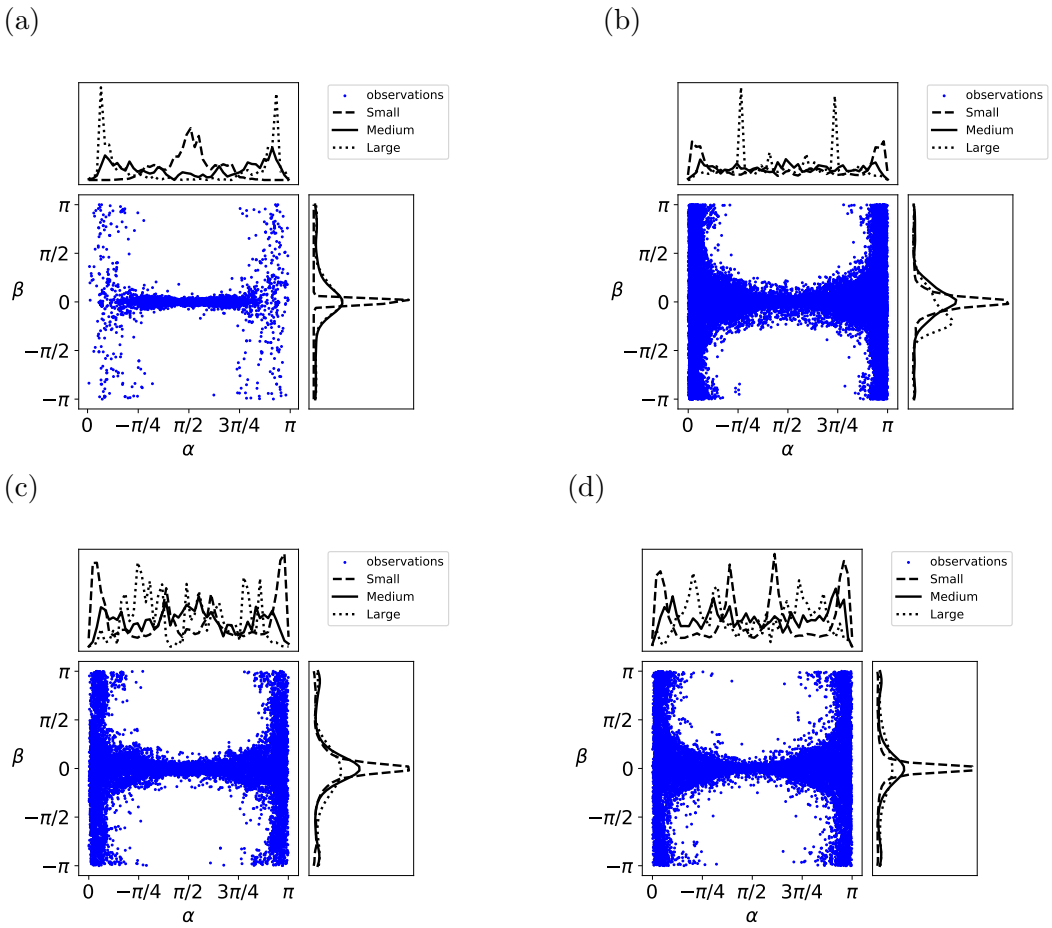


Figure 4.16: Substituent orientation preferences. The marginal distribution of orientation angles of substituents in small, medium-sized, and macrocycles are represented by dashed, solid, and dotted lines respectively. (a) carbonyl, (b) methyl (CH_3), (c) alkoxy (d) hydroxyl (e)halogens (fluorine), (f) halogens (chlorine), (g) halogen (bromine), (h) bulky substituents. For halogens, due to small number of observations in medium and large rings, the marginal distribution and the observations are not shown in the figure. All substituents tend to be directed outwardly, as indicated by the β angles. The α orientation angles vary between substituents and ring size.

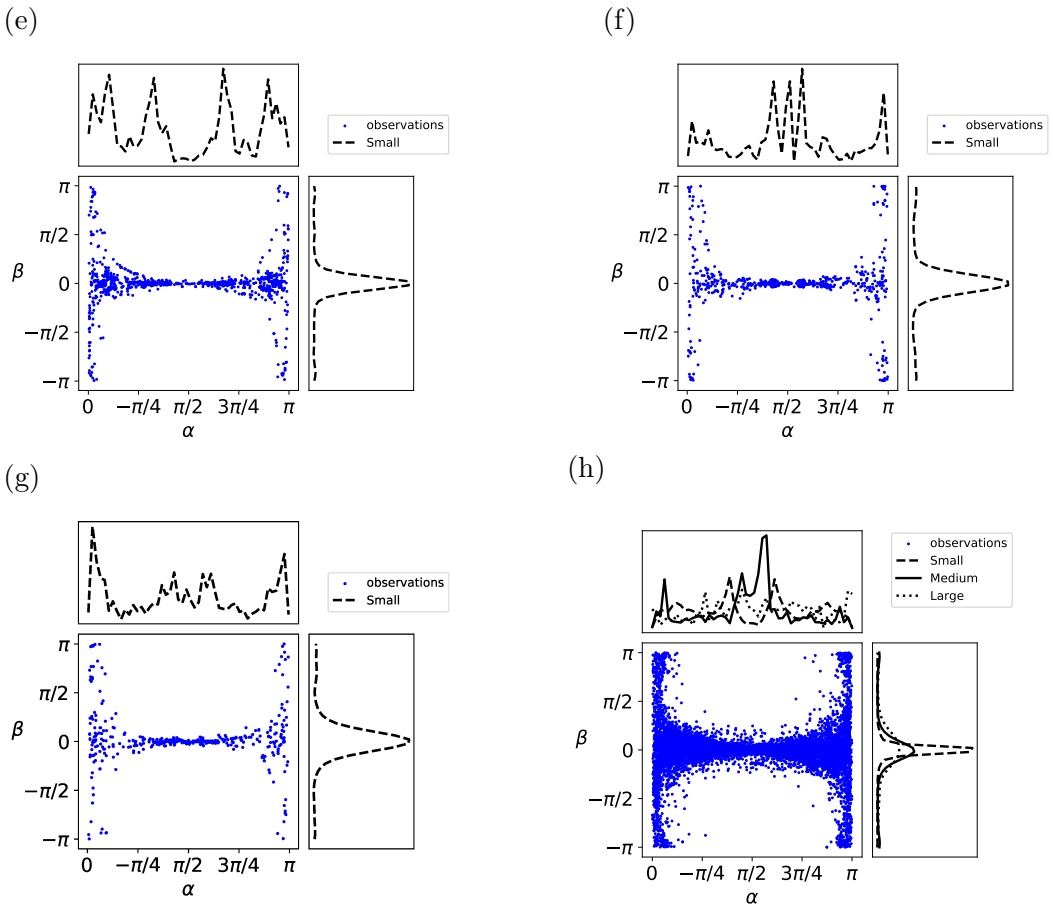


Figure 4.16: (Continued)

Figure 4.16 shows the orientation preferences of different substituent types. As might be expected, ring substituents tend to be outwardly directed (relative to the ring center) in small and medium-sized rings, *i.e.* β is close to zero, regardless of the nature of the substituents. Their α angle preferences, however, depend on both ring size and the nature of substituents. For example, Figure 4.16a shows the substituent orientation preferences of the carbonyl functional group. Due to the exocyclic double bond, its movement is restricted compared to other single bonded small substituents such as hydroxyl and methyl. The carbonyl oxygen thus tends to be equatorial to the mean plane, *i.e.* $\alpha \approx \frac{\pi}{2}$ in small rings, and preferences change as the ring size increases. Besides exocyclic double bonds, endocyclic double bonds also restrict the exocyclic motion. Figure 4.17 shows the substituent orientation preferences of methyl groups in small rings, and the α angle is bounded when the methyl is attached to a ring atom that is linked to a neighbouring ring atom with a shared endocyclic double bond. The influence of endocyclic bonds is weakened in medium-sized rings and macrocycles,

and the α angle can therefore adopt a wider range of values. Furthermore, substituents including carbonyl and hydroxyl are allowed to be quasi-axial to the mean plane and inwardly-directed in macrocycles, which are sterically unfavorable in small and medium-sized rings.

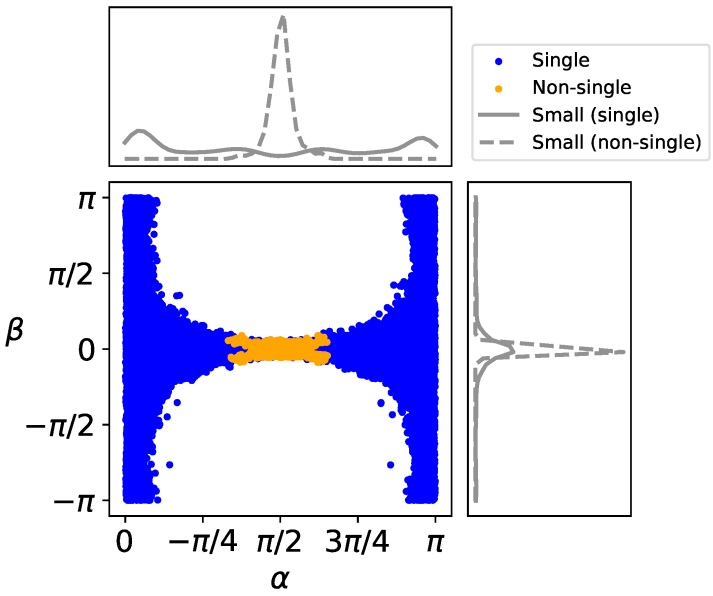


Figure 4.17: Effect of endocyclic double bond on substituent orientation angle. The orientation angles of methyl group attached to small rings with endocyclic double bonds are bounded (colored in orange).

In addition to endocyclic double bonds, the formation of long range intramolecular interactions in cyclic peptides, for example main chain-side chain interactions and side chain-side chain interaction, typically alter the conformational preferences and lead to multiple sub-clusters. The positional preferences of amide carbonyls and the side chain C_β atoms help us to understand the effect of these interactions.

Figure 4.18 shows the side chain C_β atoms and amide carbonyl orientation angles preferences in CCCC-DDDD conformational cluster. Both C_β and amide carbonyl groups are generally located accordingly to avoid steric clashes. The sub-clusters orientation angle preferences (colored by purple and green) provide a clear understanding of the effect of long range intramolecular interactions on substituent orientation. Figure 4.19 illustrates how the side chains interact in different conformational sub-clusters. On the other hand, the amide carbonyl groups are aligned to form main chain-main chain hydrogen bonds in all-*trans* (TTTT) conformations (example in Figure 4.20), lead to

rigidification of amide carbonyls, and form γ-turns, as shown in Figure 4.21. The C_β atoms also move accordingly to avoid steric clashes.

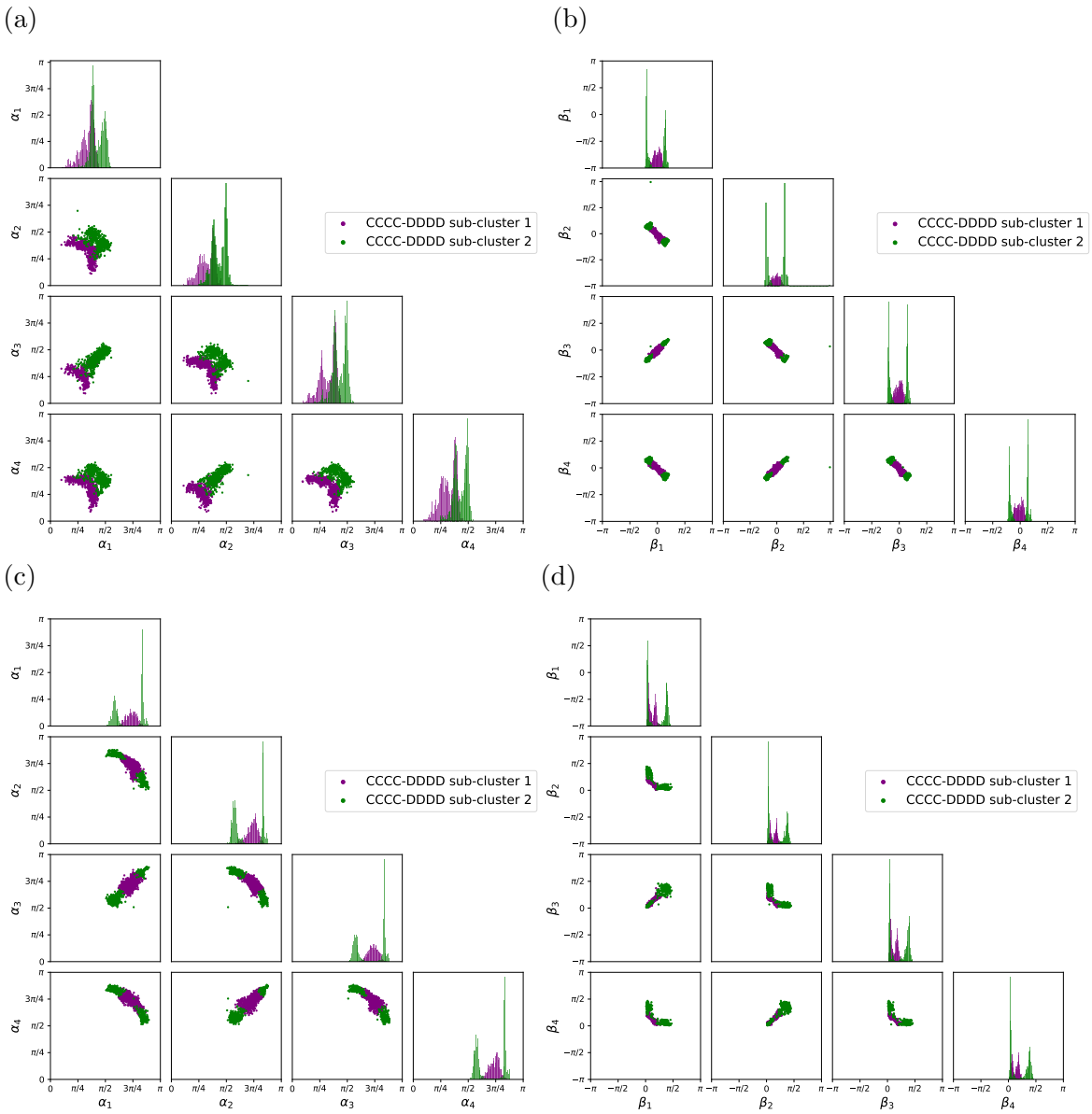


Figure 4.18: C_β and amide carbonyl orientation preference in cyclic tetrapeptdies with CCCC-DDDD conformations. (a)-(b) α, β angles of C_β. (c)-(d) α, β angles of amide carbonyl. Substituents orientation angles are correlated, and the angles preferences change (colored by purple and green) in order to align the long range intramolecular interactions such as CH-π interactions and hydrogen bonds.

(a) (b)

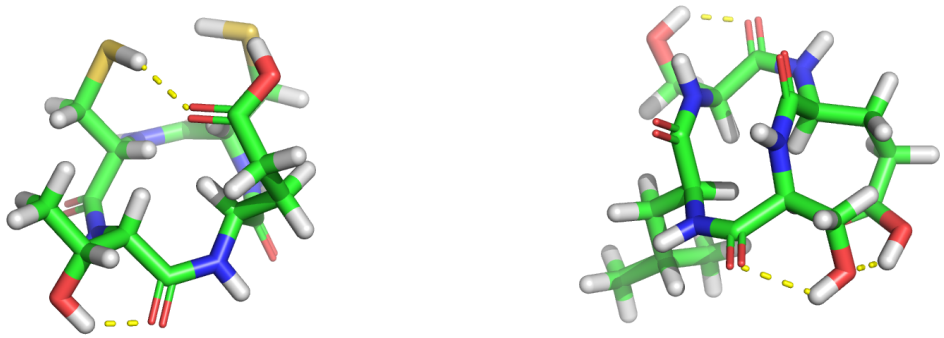


Figure 4.19: Example of CCCC-DDDD conformations: (a) sub-cluster 1; (b) sub-cluster 2. The yellow dots indicate the intramolecular hydrogen bonds.

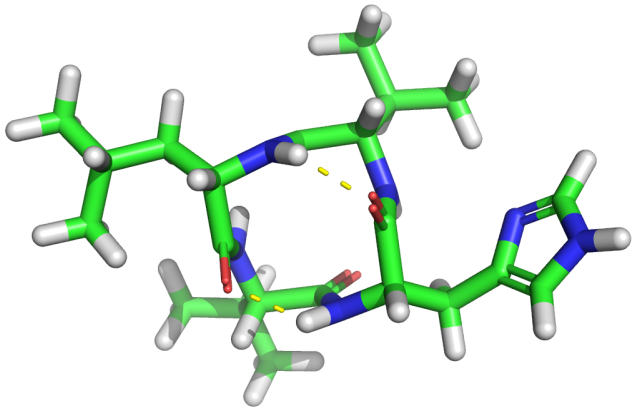


Figure 4.20: Example of all-*trans* TTTT conformation. There are four intramolecular hydrogen bonds, with two sets above the ring mean plane (indicated by yellow dots) and two sets below the mean plane (not shown).

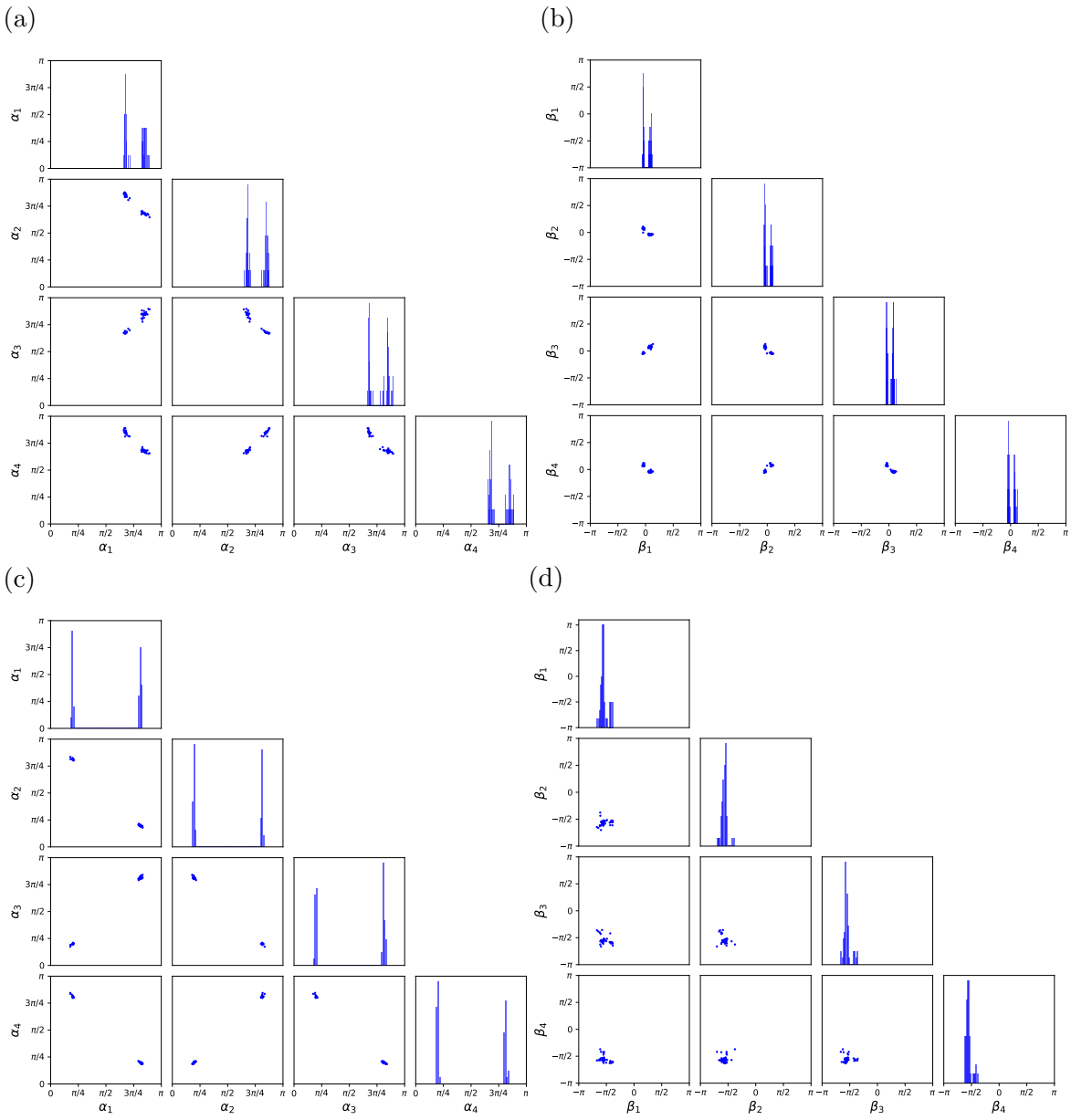


Figure 4.21: C_β and amide carbonyl orientation preference in cyclic tetrapeptdies with all-*trans* (TTTT) conformations. (a)-(b) α, β angles of C_β. (c)-(d) α, β angles preferences of amide carbonyl. Substituent orientations are restricted in order to align the main chain-main chain intramolecular hydrogen bonds.

In addition to C_β orientation, I calculated the side chain torsion angles, χ_1. Figure B.1 in Appendix B shows multimodality in χ_1 angles, which is consistent with side chain torsion angles observed in protein secondary structures. This suggests the side chain conformations can be easily sampled using standard side chain rotamer libraries (Dunbrack, 2002).

Our extended Cremer-Pople representation provides a means to understand correlated positional preferences in ring substituents; however, it is not clear what the relationship between the puckering preference and substituent orientation is, especially in macrocycles. I therefore developed simple model, Equation 4.18 to predict α and β orientation angles. Tables 4.3 and 4.4 show the predictive performance of the α and β orientation angles of carbonyl groups in 5-, 6-membered rings, and amide carbonyl in cyclic tetrapeptide CCCC-DDDD conformation at given positions. The low mean angular error and high square circular correlation coefficient suggest an excellent fit of the proposed model. Note that since the β angle falls into narrow range in small rings, and small variation will lead to low squared circular correlation.

Table 4.3: Predictive performance of carbonyl α substituent orientation angle at a given position. Each model's performance is given in terms of squared circular correlation coefficient, R^2_{circ}, mean angular error, MAE, and standard deviation of angular error.

Ring Size	Cluster	Performance (Position, R^2_{circ}, MAE, S.D.)
5	Cluster 1 (Envelope)	(1, 0.990, 0.026, 0.028) (2, 0.986, 0.023, 0.028) (3, 0.994, 0.022, 0.016)
6	Cluster 1 (Chair)	(1, 0.974, 0.028, 0.026), (2, 0.949, 0.034, 0.037)
6	Cluster 2 (Chair)	(1, 0.974, 0.028, 0.026), (2, 0.949, 0.034, 0.037)
6	Cluster 3 (Boat)	(1, 0.997, 0.035, 0.031), (2, 0.997, 0.032, 0.035),
12	CCCC-DDDD (sub-cluster 1)	(1, 0.995, 0.012, 0.009) (4, 0.994, 0.013, 0.010) (7, 0.994, 0.013, 0.010) (10, 0.994, 0.013, 0.010)
12	CCCC-DDDD (sub-cluster 2)	(1, 0.999, 0.010, 0.010) (4, 0.999, 0.011, 0.010) (7, 0.999, 0.011, 0.010) (10, 1.000, 0.010, 0.009)

Table 4.4: Predictive performance of carbonyl β substituent orientation angle at a given position. Each model's performance is given in terms of squared circular correlation coefficient, R^2_{circ}, mean angular error, MAE, and standard deviation of angular error.

Ring Size	Cluster	Performance (Position, R^2_{circ}, MAE, S.D.)
5	Cluster 1 (Envelope)	(1, 0.410, 0.017, 0.015) (2, 0.620, 0.014, 0.012) (3, 0.337, 0.016, 0.012)
6	Cluster 1 (Chair)	(1, 0.514, 0.021, 0.022), (2, 0.570, 0.015, 0.018)
6	Cluster 2 (Chair)	(1, 0.514, 0.021, 0.022), (2, 0.570, 0.015, 0.018),
6	Cluster 3 (Boat)	(1, 0.929, 0.028, 0.025), (2, 0.934, 0.027, 0.025),
12	CCCC-DDDD (sub-cluster 1)	(1, 0.908, 0.067, 0.051), (4, 0.914, 0.066, 0.048), (7, 0.912, 0.065, 0.049), (10, 0.909, 0.068, 0.048)
12	CCCC-DDDD (sub-cluster 2)g	(1, 0.994, 0.041, 0.044), (4, 0.994, 0.043, 0.042), (7, 0.993, 0.046, 0.046), (10, 0.995, 0.040, 0.044)

4.4.5 Connection between Puckering Parameters, Endocyclic and Exocyclic Torsion Angles

As mentioned earlier, measuring torsion angles is an alternative way to quantify ring puckering, and is often used in conformational analysis of small rings. de Leeuw et al. (1984) discussed the connection between ring puckering coordinates and torsion angles for small rings. Here, Equation 4.21 was proposed to convert Cremer-Pople puckering parameters to torsion angles for general N-membered rings. Table 4.5 shows the predictive performance in 5-, 6-membered rings and 12-membered cyclic tetrapeptide. All sub-models gave high squared circular correlation coefficient values, $R^2_{\text{circ}} > 0.9$, and low mean angular error, MAE< 0.14 radians ($\approx 8.0^\circ$).

Table 4.5: Predictive performance of endocyclic torsion angles. Each model's performance is given in terms of squared circular correlation coefficient, R^2_{circ}, mean angular error, MAE, and standard deviation of angular error. The endocyclic torsions (3,6,9,12) in cyclic tetrapeptide are the torsion angles of the amide bonds, which are not shown here.

Ring Size	Cluster	Performance (Torsion, R^2_{circ}, MAE, S.D.)
5	Cluster 1 (Envelope)	(1, 0.999, 0.012, 0.018), (2, 0.999, 0.016, 0.023), (3, 0.999, 0.016, 0.023), (4, 0.999, 0.027, 0.033), (5, 0.999, 0.024, 0.029)
6	Cluster 1 (Chair)	(1, 0.971, 0.019, 0.018), (2, 0.964, 0.021, 0.018), (3, 0.953, 0.021, 0.019), (4, 0.938, 0.021, 0.024), (5, 0.889, 0.036, 0.027), (6, 0.917, 0.033, 0.027)
6	Cluster 2 (Chair)	(1, 0.974, 0.018, 0.016), (2, 0.968, 0.021, 0.017), (3, 0.958, 0.022, 0.019), (4, 0.933, 0.022, 0.025), (5, 0.902, 0.034, 0.028), (6, 0.926, 0.031, 0.026)
6	Cluster 3 (Boat)	(1, 0.999, 0.025, 0.028), (2, 0.999, 0.030, 0.022), (3, 0.999, 0.030, 0.027), (4, 0.999, 0.028, 0.034), (5, 0.999, 0.041, 0.033), (6, 0.998, 0.038, 0.037),
12	CCCC-DDDD (sub-cluster 1)	(1, 0.884, 0.124, 0.094), (2, 0.824, 0.126, 0.089), (4, 0.876, 0.137, 0.096), (5, 0.796, 0.132, 0.090), (7, 0.875, 0.135, 0.099), (8,, 0, 779, 0.129, 0.095), (10, 0.883, 0.133, 0.093), (11, 0.774, 0.126, 0.090),
12	CCCC-DDDD (sub-cluster 2)	(1, 0.997, 0.037, 0.042), (2, 0.993, 0.045, 0.050), (4, 0.997, 0.047, 0.043), (5, 0.989, 0.060, 0.057), (7, 0.997, 0.042, 0.039), (8, 0.992, 0.051, 0.051), (10, 0.997, 0.040, 0.041) (11, 0.993, 0.048,0.048)

Equation 4.19 describes the relationship between the change in substituent exocyclic torsion angles with respect to the neighbouring endocyclic torsion angles, and the performances are summarised in Table 4.6. All sub-models gave high squared circular correlation coefficient, $R^2_{\text{circ}} > 0.97$, and small mean angular error, < 0.1 radian ($\approx 5.7°$), highlighting the power of the models. These models allowed us to update substituents positions efficiently in my proposed sampling scheme. Note that exocyclic bond angles will also change upon puckering, but their relationship with ring puckering parameters is not discussed here.

Table 4.6: Predictive performance of the substituent exocyclic torsion angles. Each model's performance is given in terms of squared circular correlation coefficient, R^2_{circ}, mean angular error, MAE, and standard deviation of angular error.

Ring Substituent	Performance (Sub-Model, R^2_{circ}, MAE, S.D.)
Carbonyl (C=O)	(1, 0.998,0.040,0.039)
Methyl (CH3)	(1, 0.997,0.052,0.039),(2, 0.994,0.061,0.041), (3, 0.993,0.029,0.028), (4, 0.998,0.056,0.039), (5, 0.998,0.047,0.038)
Hydroxyl (OH)	(1, 0.998,0.054,0.038),(2, 0.998,0.057,0.040)
Alkoxy (-O-)	(1, 0.998,0.044,0.039), (2, 0.980,0.038,0.037 (3, 0.998,0.048,0.041)
Bulky carbon (-CH0, CH1, CH2)	(1, 0.996, 0.064, 0.052),(2, 0.972, 0.063, 0.036), (3, 0.975, 0.082, 0.049),(4, 0.996, 0.068, 0.054) ,
Halogen (-F)	(1, 0.998,0.046,0.033), (2, 0.998,0.048,0.036)
Halogen (Cl)	(1, 0.996,0.060,0.048), (2, 0.997,0.054,0.043)
Halogen (Br)	(1, 0.998,0.057,0.045), (2, 0.998,0.052,0.052)

4.4.6 Ring Reconstruction

To assess the performance of the proposed sampling method, I selected 20 simple ring molecules including small and medium sized monocyclic substituted and unsubstituted rings. None of the molecules contained any acyclic rotatable bonds. The GFN2-computed lowest energy conformation obtained from the CREST were used as the reference conformations. Note that the conformation with the lowest RMSD does not necessary give the lowest TFD values. Here, I reported the TFD values of the conformation with the lowest RMSD values. Figure 4.22 shows two examples, cycloheptane and 4,4-dimethylhexanone. Both have generated conformations (without energy minimisation) that are very similar to their corresponding reference conformations, with low RMSD values (0.12 Å and 0.16 Å) and TFD values (0.06 and 0.05).

In general, my proposed method gives low average TFD value (0.05), and an average RMSD value 0.09 Å on the selected cyclic molecules, see Appendix B.9. It demonstrates the effectiveness of our proposed method. Note that the large RMSD values are ascribed to the deviation in bond lengths and bond angles. Local geometry optimisation with bond length and bond angles will help generate a better conformation with lower RMSD values.

(a) (b)

Figure 4.22: Alignment of conformations generated by my method (in orange) and the lowest energy conformation sampled by CREST (in green), for (a) cycloheptane, and (b) 4,4-dimethylcyclohexanone. The sampled conformations are very similar to the lowest energy conformation, with RMSD values of 0.12 Å and 0.16 Å, and TFD values of 0.06 and 0.05, respectively. The torsion deviations are small in both cases, and the deviation in bond lengths and bond angles lead to larger RMSD values. RDKit was used to compute the RMSD and TFD values.

4.5 Summary

In this chapter, I explored the use of Cremer Pople puckering parameters to study the conformational preferences of a diverse set of rings, including macrocycles and cyclic peptides. By standardizing the atom ordering of ring atoms, I was able to elucidate the puckering preferences for general N-membered ring molecules from GFN2-computed low energy structures. The extended representation provided a means to characterise the geometries of ring substituents, thus enabling us to study the coupled motion of ring substituents upon puckering.

I showed that the ring conformations are generally clustered, based on their canonical conformations. The presence of endocyclic double bonds and shared aromatic ring bonds reduce the flexibility of the rings and alter the puckering preferences. The pseudo-rotation is usually restricted, except in some canonical conformations such as chair and boat conformation in flexible 6-membered rings. In addition, the substituent orientation preferences have great influence on ring geometries. The substituent α orientation angle preferences normally depends on the nature of substituent and the ring size. On the other hand, the β orientation angle is rigid in small and medium-sized rings, regardless of the nature of substituents, while it shows variation in macrocycles. The formation of intramolecular interaction between substituents leads to rigidification of substituent position and unique puckering preferences. Different models were proposed to understand the local change of geometries, including change of substituent orientation, endocyclic and exocyclic torsion angles upon puckering.

A novel knowledge-based conformer sampling tool based on the puckering preferences was proposed. Kernel density estimation (KDE) was used to learn the puckering preferences and sample new conformations. Using the physical understanding of the rotational dependence between endocyclic and exocyclic torsion angles, the substituent orientation can be updated accordingly. To progress to larger ring systems, more data is necessary for the density estimation. Future work should focus on increasing sampling with additional accurate quantum mechanics (QM) energy calculation and developing better density estimation technique to learn the coupled puckering preferences in large rings effectively. The resulting puckering preferences derived from conformations with QM energies can then be utilized to sample low energy macrocycle conformations efficiently.

My proposed models and sampling framework are general and readily extensible to larger and more complex ring systems. An improved understanding of the conformational preference of cyclic molecules will accelerate the sampling of low energy conformers for a wide range of computational modelling applications.

Chapter 5

Understanding Conformational Entropy in Small Molecules

Most of the work in this chapter has been reproduced from the following works:
L. Chan, G. M. Morris, G. R. Hutchison. Understanding Conformational Entropy in Small Molecules. *ChemRxiv* 2020 10.26434/chemrxiv.12671027

5.1 Background

While entropy is a major driving force in many physical and chemical processes and is a key component of the free energy of a molecule, it can be challenging to calculate with standard quantum mechanics methods. Proper consideration in flexible molecules, even within a rigid rotor approximation, requires not just the calculation of the translational, rotational and vibrational partition functions, but sampling all thermally accessible conformational degrees of freedom. Several efforts have focused on both exhaustive quantum mechanical evaluations of multiple conformers (Speybroeck et al., 2005; Ellingson et al., 2006; Zheng et al., 2011; Simón-Carballido et al., 2017; Wu et al., 2019), and empirical estimates of the entropy from multiple thermally accessible conformers (Ghahremanpour et al., 2016). Other efforts have used molecular dynamics with various force fields, which may not yield the same accuracy as modern quantum mechanics methods (Head et al., 1997; Peter et al., 2004; Chang et al., 2005, 2007; Suárez et al., 2011).

In principle, the number of possible conformers increases exponentially with the number of degrees of freedom of a molecule. In the solution and gas phase, many bonds have low torsional energy barriers (*e.g.*, sp^3−sp^3 single bonds), while in the solid

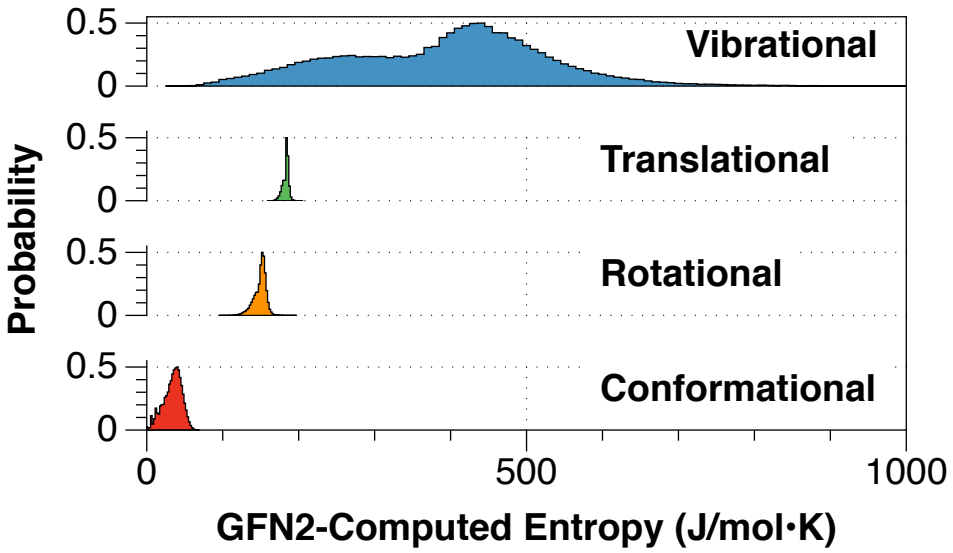


Figure 5.1: Distributions of GFN2-calculated vibrational, translational, rotational, and conformational entropies across 93021 molecules in the datasets, see Section 5.2.1.

state, steric effects may restrict free torsional motion. Thus, it is common practice in conformer generation to focus on sampling hundreds or thousands of geometrically diverse conformers, (O'Boyle et al., 2011b; Hawkins, 2017) and using fast molecular mechanics force fields for energy evaluations – even if they do not always correlate well with more accurate electronic structure methods (Kanal et al., 2018; Rai et al., 2019; Folmsbee and Hutchison, 2020).

Recent improvements in density functional tight-binding approximations (Grimme et al., 2017; Bannwarth et al., 2019; Grimme, 2019; Pracht et al., 2020) and in availability of computational resources have enabled the work I present here: an evaluation of conformer ensembles and the corresponding entropies of over 120,000 small molecules with up to twenty rotatable bonds, and comprising over 12 million conformers. Figure 5.1 highlights one of the key results that the vibrational entropy usually has the largest contribution to the total entropy, followed by translational and rotational entropies. The conformational entropy makes the smallest contribution. The median conformational entropy comprises 36.3 J/mol·K, and while relatively small, should not be neglected.

Although the relative contribution of conformational entropies to total entropies is small, the calculation time is 200-300 times longer than for the vibrational calculations, with a median time of 1.01 hours per compound, and an average of 2.08 hours per compound for a dual core job using the GFN2 method, as illustrated in Figure 5.2.

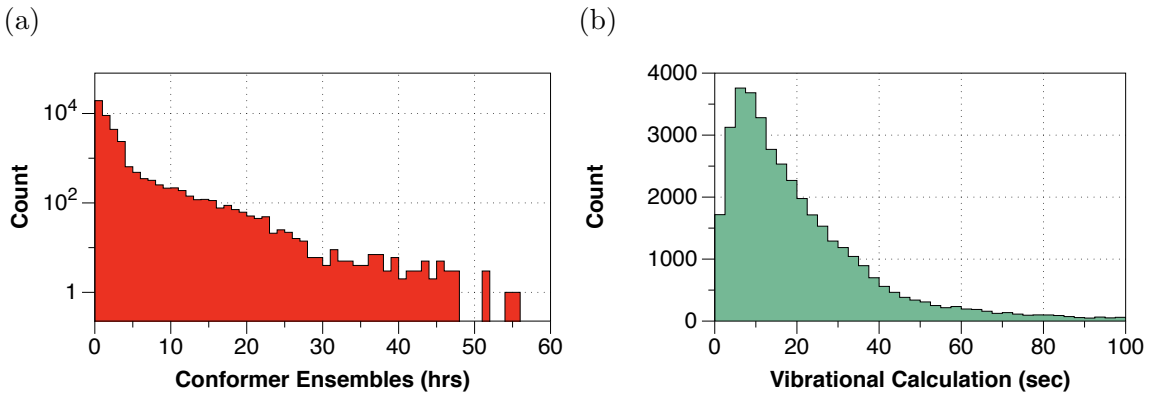


Figure 5.2: Histograms showing the counts of (a) CREST dual-core run times (in hours) on a logarithmic scale; and (b) GFN2 dual-core vibrational run times (in seconds) across a subset of 39005 molecules in the dataset, see Section 5.2.1.

To overcome this computational bottleneck, I elucidate a physical interpretation of the components of conformational entropy by developing a statistical model across the datasets I curated, as described in Section 5.2.1.

5.2 Data and Methods

Here, I first summarise the set of molecules used for the study, followed by a description of the methods.

5.2.1 Data

Over 120,000 compounds were drawn from the Crystallographic Open Database (COD) (Gražulis et al., 2009, 2012) as well as more complex organic macrocycles from the ZINC database (Sterling and Irwin, 2015), and consisted of any of the following elements: hydrogen, boron, carbon, nitrogen, oxygen, fluorine, silicon, phosphorus, sulfur, chlorine, bromine and iodine. The set includes a wide range of molecular sizes, with up to 128 heavy atoms, up to 181 bonds, and up to twenty rotatable bonds.

To assess the model performance, I split the data into two sets: (i) a training set (93021 molecules) and (ii) a testing test (15547 molecules) selected from the ZINC database, namely ZINC-I set. Most of the analysis and models focused on the training set, and the testing set was used to assess the predictive performance. The InChI key (Heller et al., 2013) of each molecule in both the training set and testing set was

computed to ensure there was no overlap between datasets. To examine the models' predictive performance, an additional cyclic tetrapeptide (CTP) testing set was assembled. This set contains 8661 head-to-tail cyclic tetrapeptides, *i.e.* cyclization from the *N*-terminus to the *C*-terminus, and thus gives a set of 12-membered rings. They are composed of fourteen out of twenty naturally occurring *L*-amino acids (see Table 5.1). Both ZINC and peptide test sets contain dissimilar molecules from the training sets, with a median of 0.57 and 0.53 of Tanimoto similarity (Ralaivola et al., 2005) between the extended connectivity fingerprints with diameter 6 (Rogers and Hahn, 2010) respectively.

Table 5.1: Fourteen of the twenty naturally occurring amino acids that were used to generate the cyclic tetrapeptides (CTPs) test set.

Type	**Amino Acids**
Special	Cysteine, Glycine
Charged	Histidine, Lysine, Aspartic Acid, Glutamic Acid
Polar Uncharged	Serine, Threonine
Hydrophobic	Alanine, Valine, Leucine, Phenylalanine, Tyrosine, Tryptophan

5.2.2 Calculation of Entropies

For all molecules from ZINC, RDKit (Landrum, 2018) was used to generate initial conformations. For molecules from COD, the X-ray crystal structures were used as initial conformations. The molecular geometries of all molecules from ZINC and COD were then optimized using the GFN2 method (Grimme et al., 2017; Bannwarth et al., 2019), followed by conformer sampling using the iterative metadynamic sampling and genetic crossover (iMTD-GC) method implemented in the CREST program (Grimme, 2019; Pracht et al., 2020), including additional geometry optimization of the final conformational ensemble. The lowest energy conformer was selected for calculating the vibrational modes to evaluate standard rigid rotor harmonic oscillator vibrational, translational, and rotational entropies (Grimme, 2012). Note that the CREST calculation may break molecules into fragments; molecules that were fragmented in its final output were excluded from our analysis.

Table 5.2: Feature definitions as SMARTS expressions used in the calculation of descriptors. The definition of donors and acceptors are adapted from (Gobbi and Poppinger, 1998).

Name	SMARTS
Rotatable bond	[!$(*#*)&!D1]-!@[!$(*#*)&!D1]
Methyl group	[CX4H3]
Acyclic amide	[#7]!@;-[CX3]=[O]
Acyclic ester	[#8]!@;-[CX3]=[O]
Acyclic thioamide	[#7]!@;-[CX3]=[SX1]
Cyclic amide	[NX3]@[CX3]=[O]
Cyclic ester	[OX2H0]@[CX3]=[O]
Cyclic thioamide	[NX3]@[CX3]=[SX1]
Donor	[$([N;!H0;v3,v4&+1]),$([O,S;H1;+0]),n&H1&+0]
Acceptor	[$([O,S;H1;v2;!$(*-*=[O,N,P,S])]),$([O,S;H0;v2]), $([O,S;-]),$([N;v3;!$(N-*=[O,N,P,S])]), n&H0&+0,$([o,s;+0;!$([o,s]:n);!$([o,s]:c:n))]

5.2.3 Methods

5.2.3.1 Degrees of Freedom

The calculation of the conformational entropy of a molecule is associated with the distribution of its conformers. It is thus necessary to understand the factors controlling the conformer population. The number of conformers, in principle, grows with the number of degrees of freedom in a molecule, for example number of rotatable bonds, number of methyl groups and number of degrees of freedom in flexible rings. To understand the effect of increasing number of degrees of freedom, three descriptors were introduced, namely the number of rotatable bonds , N_{rotor}, the number of methyl groups, N_{CH3}, and total ring flexibility, R_f^{Total}.

The number of rotatable bonds and number of methyl groups are the counts of substructures defined by the SMARTS patterns in Table 5.2. The rotatable bond is defined as an atom which is not triply bonded and not one-connected, *i.e.* terminally connected by a single non-ring bond to an equivalent atom. An additional descriptor, total ring flexibility, R_f^{Total}, was introduced to represent the number of degrees of freedom in flexible rings.

5.2.3.2 Ring Flexibility

Inspired from results published by Cremer and Pople (1975) that showed the conformational space of an N-membered *monocyclic* ring can be described by $N-3$ parameters, one can derive a descriptor based on the number of ring atoms. To explain the actual number of degrees of freedom in more complex ring structures, the concept of unique ring families (URFs) (Kolodzik et al., 2012) is required. The calculation of unique ring families (URFs) was introduced in Chapter 4.3.4.

Once the URFs are identified, a subgroup ring flexibility score, R_f^S, can be computed using Equation 5.1. Building on the results from Cremer and Pople (1975) and Chapter 4, the number of degrees of freedom in a relevant cycle is given by N_r-3-p, where N_r is the number of atoms in the relevant cycle, and p is the penalty to account for the effect of presence of endocyclic double bonds and shared aromatic bonds. Counts of non-single bonds in a ring were used as the penalty. The maximum function in Equation 5.1 avoids negative values in an over-constrained ring system such as aromatic rings. R_s indicates the number of relevant cycles in a URF. The subgroup ring flexibility is thus the average ring flexibility of a URF:

$$R_f^{\mathrm{S}} = \frac{1}{R_S}\sum_{r=1}^{R_S}\max(N_r - 3 - p, 0) \tag{5.1}$$

Beyond the effect of non-single bonds, ring junction types such as spiro (at least two rings sharing a common ring atoms), fused (at least two rings sharing two adjacent atom) and bridged rings (at least two rings sharing more than two atoms), have great influence on the associated conformer population. The number of degrees of freedom decreases as the number of ring junctions increases. I therefore introduced a penalty based on the counts of various types of ring junctions, as described in Table 5.3. Furthermore, delocalisation of electrons in amide, ester, and thioamide group (see definitions in Table 5.2) partially restrict the rotations of a ring bond. A penalty was also included to account for their associated effects.

The total ring flexibility, R_f^{Total}, is simply the sum of subgroup ring flexibility, as shown in Equation 5.2. Three examples are shown in Figure 5.3.

$$R_f^{\mathrm{Total}} = \sum_{t=1}^{T} R_f^t \tag{5.2}$$

where t is the index of the subfamilies, and T is the total number of subfamilies in the molecule.

Table 5.3: Ring Penalties. A fused ring penalty is only applied to fused rings with shared atoms linked by a single bond.

Condition	Penalty, p
Non-single bond	Count of non-single bonds
Spiro	Count of spiro atoms
Fused	Count of fused single bonds
Bridge	Count of bridgehead atoms
Polycyclic	Count of polycyclic atoms
Cyclic amide	Count of cyclic amides
Cyclic ester	Count of cyclic esters
Cyclic thioamide	Count of cyclic thioamides

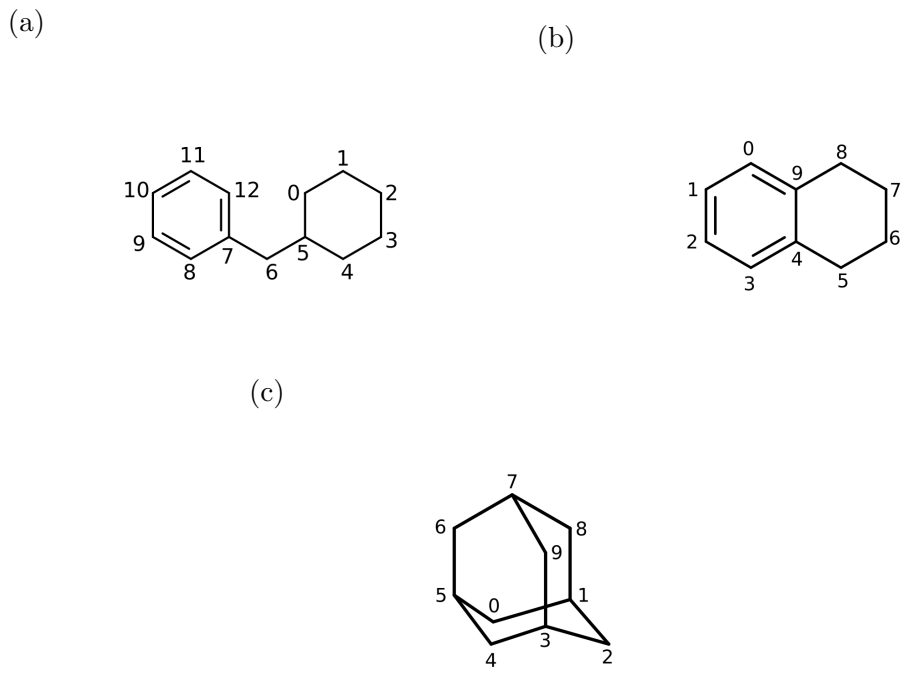


Figure 5.3: Example of URFs calculations: (a) cyclohexylmethylbenzene; (b) 1,2,3,4-tetrahydronaphthalene; (c) adamantane. In (a), there are two URFs: URF0, (0,1,2,3,4,5); and URF1, (7,8,9,10,11,12). The subgroup ring flexibility of URF0 and URF1 are 3 and 0 respectively. The total ring flexibility is 3. In (b) there are two URFs: URF0, (0,1,2,3,4,9); and URF1, (4,5,6,7,8,9). The subgroup ring flexibility of UFR0 and URF1 are 0 and 2 respectively, and the total ring flexibility is 2. In (c), there are four URFs: URF0 (0, 5, 6, 7, 8, 1), URF1 (0, 1, 2, 3, 4, 5), URF2 (3, 4, 5, 6, 7, 9), and URF3 (1, 2, 3, 9, 7, 8). Since each URF contains 3 bridgehead atoms, the corresponding subgroup ring flexibility is 0. Hence, the total ring flexibility is 0.

Note that *cis-trans* isomers in fused ring systems impose different steric constraints to the system and may give different conformational entropy values. I did not penalise *cis-trans* isomerism in this analysis.

5.2.3.3 Chemical Functionality

Besides the number of degrees of freedom, the chemical functionality and the molecular shape have great influence on the conformer population. For instance, delocalisation of electrons in acyclic amide, ester and thioamides groups (see definitions in Table 5.2) reduce the conformational flexibility and render them planar. A descriptor based on the count of these functional groups in a molecule was introduced. I denote the count of these functional groups as N_{SG}. Note that the definitions of amide and ester also match other functional groups, such as urea and carbamate, and result in multiple matched groups.

5.2.3.4 Foldability

Additionally, intramolecular interactions such as hydrogen bonds and π-π stacking can have strong influence on the molecular shape. Typically, formation of such interactions leads to a so-called "folded" structure, and reduces the conformational flexibility. Accurate prediction of the formation of intramolecular interactions will improve the prediction of conformational entropies. Multiple efforts have been made to characterise the topologies that are likely to form intramolecular hydrogen bonds (Bilton et al., 2000; Kuhn et al., 2010), by analysing the X-ray crystal structures from the Cambridge Structural Database (CSD) (Groom et al., 2016) and the Protein Data Bank (PDB) (Berman et al., 2000). Bond angle and torsion angle preferences were studied and a set of motifs with a high probability of forming intramolecular hydrogen bonds were suggested. To predict potential formation of intramolecular interactions, I applied junction analysis to study the characteristics of the shortest path of these potential interactions. In particular, intramolecular hydrogen bonds and π-π stacking (see definitions in Table 5.4) were considered in this analysis. Junction analysis (Monod et al., 2004) was originally used to identify structurally conserved positions of amino acids in antibodies. I applied the same concept to study the substructure positional preferences in molecules involving intramolecular hydrogen bonds and π-π stacking. The calculation is described next.

Table 5.4: Distance and angle constraints used to determine the intramolecular interactions. The distance and angle constraints are adapted from Schrödinger (Schrödinger, LLC, 2020).

Interactions	Distance and Angle Constraints
π-π stacking	i. Maximum distance between ring centroids of two rings is 4.4 Å.
	ii. Angles between the ring planes is less than 30°.
Hydrogen bonds	i. Maximum distance is 2.5 Å.
	ii. Minimum donor angle is 120°.
	iii. Minimum acceptor angle is 90°.

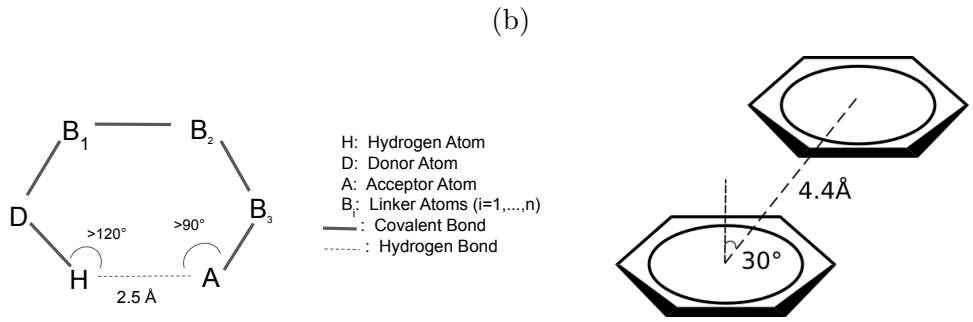


Figure 5.4: Graphical illustration of the distance and angle constraints in (a) hydrogen bond and (b) π-π stacking.

Given a path defined by bonds between atoms in a molecular graph, the path is aligned such that the terminal nodes are the atoms with desired intramolecular interactions. For instance, one can treat the first atom as the hydrogen bond donor and the last atom as the hydrogen bond acceptor. Gaps are inserted in between when the path is shorter than the maximum path length. This is not unlike protein squence aligment and loops of variable length between common secondary structural elements in proteins, in particular complementarity-determining regions (CDRs) in antibodies (Monod et al., 2004).

Inspired by the motif analysis in Bilton et al. (2000); Kuhn et al. (2010), I developed a method that can characterise the path involved in hydrogen bonds using the element type, atomic hybridization and whether the atom is in a ring. The hybridization of an atom is inherited from the definition in RDKit (Landrum, 2018). Note that the oxygen in amides, and the ether oxygen in ester groups, hydroxyl oxygen in carboxyl

groups are classified as sp^2 hybridized. The path direction was unified such that the path began with the donor atom and ended with the acceptor atom. I further separated the motifs based on the acceptor types for junction analysis. Five types of acceptors were considered, including carbonyls, hydroxyls, alkoxy groups, acyclic nitrogens and nitrogens in heterocycles.

Intramolecular aromatic interactions such as π-π stacking also influence molecular shape. Instead of using the atom and bond properties, I studied the positional preferences of a set of functional groups, including carbamates, ureas, ketones, ethers, esters, amides. The position of a specified atom in the corresponding substructure and the orientation of an atom in the substructure bonded to the specified atom were taken into account, as described in Table 5.5. To standardise the atom ordering, the position of the aromatic ring atom that is closer to the substructure of interest was used as beginning atom. If multiple substructures exist in the same path, the substructures with the highest rank was used. The orientation is denoted as "F" if the bonded atom is *followed* by the specified atom, and is denoted as "B" if the bonded atom comes *before* the specified atom.

Table 5.5: SMARTS expressions used to identify functional groups in π-π stacking structural motifs. Urea type 1, 2 and 3 indicate cyclic urea, urea with one nitrogen in a ring, and acyclic urea respectively. Similarly, amide type 1, 2, and 3 indicate cyclic amide, amide with nitrogen in ring and acyclic amide respectively. The general amide SMARTS pattern will match all types of amide. The position 1 indicates location of the atom of interest in the corresponding functional group. The position 2 indicates the location of atom that bonded to atom of interest, and is used to determine the orientation.

Functional Group	SMARTS	Pos. 1	Pos. 2
Carbamate	[#8][C](=O)[#7]	1	0
Urea Type 1	[#7&R][C&R](=O)[#7&R]	1	N/A
Urea Type 2	[#7&R][C&!R](=O)[#7&!R]	1	0
Urea Type 3	[#7&!R][C&!R](=O)[#7&!R]	1	N/A
Ketone	[#6][C](=O)[#6]	1	N/A
Ether	[#6][O][#6]	1	N/A
Ester	[#8][C](=O)[#6]	1	0
Amide	[#7][C](=O)[#6]	1	0
Amide Type 1	[#7&R][C&R](=O)[#6&R]	1	0
Amide Type 2	[#7&R][C&!R](=O)[#6&!R]	1	0
Amide Type 3	[#7&!R][C&!R](=O)[#6&!R]	1	0

Table 5.6: Position and Count Thresholds

Intramolecular Interactions	**Thresholds**
Hydrogen Bond	(i) >= 10% of the total count at a given position. (ii) minimum 10 observations at a given position.
π-π Stacking	(i) >= 10% of the total count at a given position (ii) minimum 5 observations at a given position.

The over-represented atom types or functional groups at given positions (see thresholds in Table 5.6) were then used to identify motifs with desired properties. Once the motifs in a molecule were identified, the foldability score can be computed directly as follows:

Foldability score for a motif with intramolecular hydrogen bonds (F_{HBond}):

$$W_r = \sum_{i=1}^{K} \sum_{j=1}^{L_i} w_r^{ij} \mathbb{1}_{r \in ij} = \sum_{i=1}^{K} \sum_{j=1}^{L_i} \frac{1}{L_i} \mathbb{1}_{r \in ij} \tag{5.3}$$

$$F_{HBond} = \sum_{r=1}^{R} \min(W_r, 1) \tag{5.4}$$

where K is the number of donors in the molecule; L_i is the number of possible acceptors interact with a fixed donor i, and R is the number of rotatable bonds respectively; W_r is the score for rotatable bonds, r; and w_r^{ij} is the weight of rotatable bond r in the path from donor i to acceptor j; $\mathbb{1}_{r \in ij}$ is a indicator function, and is 1 if the rotatable bond r is in the path from donor, i, to acceptor, j. The foldability score, F_{HBond}, is simply the expected number of rotatable bonds found in the molecular subgraph containing intramolecular hydrogen bonds. Note that the donor atoms can potentially interact with multiple acceptors, and the final geometry depends on the strength of each hydrogen bond pair and the overall steric interaction. To avoid overestimation, I calculated the expected number of rotatable bonds restricted by the formation of hydrogen bonds, rather than the count of rotatable bonds. Equal weight was used for all possible pairs in the analysis.

Foldability score for a motif with π-π stacking:

$$F_{\pi-\pi} = \sum_{p=1}^{p} N_{\text{rotor}}^{p} \tag{5.5}$$

where N^{p}_{rotor} is the number of rotatable bonds found in the shortest path between two aromatic rings that form π-π stacking; and P is the total number of possible π-π interactions in the molecule. The foldability score, $F_{\pi-\pi}$, is thus the number of rotatable bonds found in molecular subgraph containing π-π stacking interaction.

5.2.4 Models

Using all the descriptors mentioned above, a linear model, LR-Best, was proposed to predict the conformational entropies of small molecules. There were six descriptors in total: (i) the number of rotatable bonds (N_{rotor}), (ii) the number of methyl groups (N_{methyl}), (iii) the total ring flexibility (R_f^{Total}), (iv) the number of functional groups, *i.e.* amides, ester and thioamides (N_{SG}), (v) the hydrogen bond foldability (F_{Hbond}), and (vi) the π-π stacking foldability ($F_{\pi-\pi}$). $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$ are the model parameters.

LR-Best:

$$\begin{aligned} S_{\text{conf}} = \beta_0 &+ \beta_1 \log(N_{\text{rotor}} + 1) + \beta_2 \log(N_{\text{methyl}} + 1) + \beta_3 \log(R_f^{\text{Total}} + 1) \\ &+ \beta_4 \log(N_{\text{SG}} + 1) + \beta_5 \log(F_{\text{HBond}} + 1) + \beta_6 \log(F_{\pi-\pi} + 1) \end{aligned} \tag{5.6}$$

LR-1:

$$S_{\text{conf}} = \beta_7 + \beta_8 \log(N_{\text{rotor}} + 1) \tag{5.7}$$

where β_7 and β_8 are model parameters.

A baseline linear model (LR-1) with number of rotatable bonds (N_{rotor}) as a sole input feature (Ghahremanpour et al., 2016) and multiple machine learning models were included for comparison. The machine learning models included (i) least absolute shrinkage and selection operator (LASSO), (ii) ridge regression, (iii) kernel ridge regression, and (iv) a neural network (NN). Extended connectivity fingerprints with diameter 6 (ECFP6) (Rogers and Hahn, 2010) with 4096 bits was used as the inputs for all machine learning models. The implementation of the models is discussed in Section 5.2.5.

5.2.5 Implementation

RDKit (Landrum, 2018) was used to read molecules, generate initial conformers for molecular dynamic simulation, calculate molecular descriptors and generate the

ECFP6 fingerprints. RingDecomposerLib (Flachsenberg et al., 2017) was used to identify the URFs of a molecule. A Python package, statsmodels (Seabold and Perktold, 2010), was used to estimate model parameters of the proposed linear model and the baseline model. The implementations of LASSO, ridge regression, kernel ridge regression, and cross-validation in Scikit-learn (Pedregosa et al., 2011) were used. Keras (Chollet, 2015) was used to construct the neural network.

5.2.6 LASSO and Ridge Regression

ECFP6 with 4096 bits was used as model inputs. The hyperparameter (penalty), α, was optimised by 3-fold cross-validation with a grid-search over a parameter grid, $\alpha \in \{0.01, 0.05, 0.1, 0.5, 1.0, 5.0, 10.0\}$.

5.2.7 Kernel Ridge Regression (KRR)

ECFP6 with 4096 bits was used as model's inputs. The Tanimoto (Jaccard) kernel (Ralaivola et al., 2005) was used in KRR. The Nyström approximation (Williams and Seeger, 2001) was used to approximate the feature map. The hyperparameter α was optimized by 3-fold cross-validation with a grid search over a grid $\alpha \in \{0.001, 0.01, 0.1, 1, 10\}$.

5.2.8 Neural Network

ECFP6 with 4096 bits was again used as the model's inputs. Rectified Linear Unit (ReLu) activation functions (Nair and Hinton, 2010) were used in the hidden layers of the neural network model. A linear activation function was used for the final layer. The ADAM optimizer (Kingma and Ba, 2014) was used for the optimisation. I also added elastic net regularization in the hidden layers, with $L_1 = 10^{-4}$ and $L_2 = 10^{-4}$. The hyperparameters, *i.e.* number of hidden nodes, number of layers and dropout rate were optimised by cross-validation with grid search over a parameter grid: $N_{\text{Hidden}} \in \{32, 64\}$, $N_{\text{Layer}} \in \{3, 4, 5\}$) and dropout (Srivastava et al., 2014) rate $\in \{0.5, 0.6\}$.

5.3 Results and Discussions

A set of saturated hydrocarbons, specifically alkanes, were used to investigate the effect of increasing the number of degrees of freedom on conformational entropy. In an unbranched n-alkane, C_nH_{2n+2}, the low torsional energy barrier of carbon-carbon single bonds enables all bonds to rotate freely and result in different conformations. In principle, with low torsional barriers and all bonds being equal, the number of conformers should increase exponentially with the count of rotatable bonds ($\approx 3^{n-3}$), assuming three possible local minima per rotatable bond. However, symmetry, correlated dihedral angles, and excluded volume often reduce the number of thermally accessible conformers (Vansteenkiste et al., 2003; Speybroeck et al., 2005; Ellingson et al., 2006; Wu et al., 2019).

Rather than an exponential growth in the number of possible conformers of linear alkanes, the number of *low-energy conformers* increases sub-linearly on a logarithmic scale, when evaluated either with exhaustive systematic conformer enumeration (Confab) (O'Boyle et al., 2011b) using a standard molecular force field (MMFF94) (Halgren, 1996; Halgren and Nachbar, 1996), or when using CREST conformer generation with the GFN2 method, as illustrated in Figures 5.5a and 5.5b respectively. The curves fit roughly to a power-law function, with exponents $\approx 1.5 - 2.6$, depending on the method and the energy window.

Since the number of low-energy conformers increases relatively slowly (*i.e.* sub-exponentially) with the number of rotatable bonds, the conformational entropy will therefore increase logarithmically, as found by the computed CREST / GFN2 entropies. For short alkane chains ($n < 4$ carbon atoms), the increase in conformational entropy is approximately linear, and approximately logarithmic or perhaps close to constant for long chains (see Figure 5.5c). One can understand that in long chains, dihedral motion in the center of the molecule will inherently restrict otherwise free rotations to avoid steric clashes—a concept known as *excluded volume* in polymer theory (Hill, 1986). These results match previous detailed quantum chemical calculations of conformational entropy in linear alkanes (Vansteenkiste et al., 2003; Speybroeck et al., 2005; Ellingson et al., 2006; Wu et al., 2019).

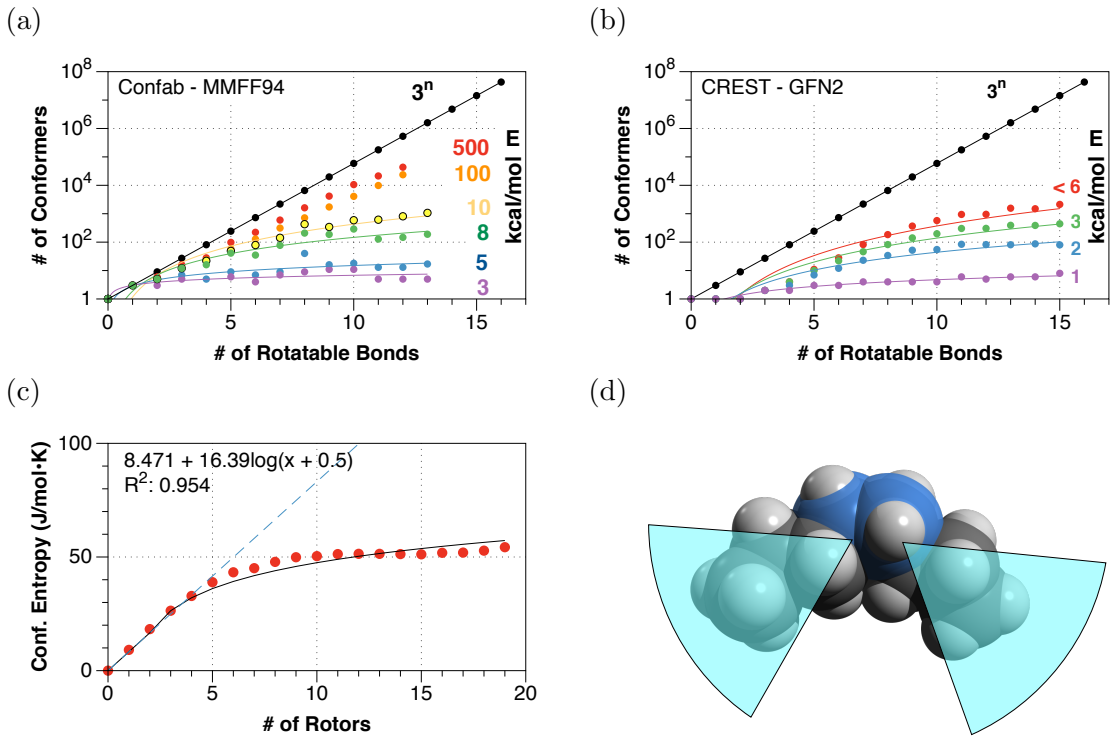


Figure 5.5: Conformational entropies for increasing lengths of n-unbranched alkanes, C_nH_{2n+2}. The counts of conformations in (a) and (b) are shown on a logarithmic scale. (a) Number of alkane conformers within a given energy range (in kcal/mol) of the global minimum (*i.e.*, within 3, 5, 8, 10, 100, and 500 kcal/mol) using Confab exhaustive sampling with the MMFF94 force field. (b) Number of conformers within a given energy window in kcal/mol of the global minimum using CREST sampling and the GFN2 method. (c) Conformational entropies for calculated for n-alkanes using CREST / GFN2. Note that for smaller hydrocarbons ($n < 4$ carbons) the scaling is approximately linear, and beyond $n = 8 - 10$ carbons, the conformational entropies are roughly constant. (d) Schematic of central torsion in octane C_8H_{18} indicating potential steric bumping (clashing carbons shown in blue) between the two molecular ends.

Figure 5.6a shows the conformer populations across the set of ~93,000 molecules at different GFN2-computed energy cutoffs (shown in different colors up to 6 kcal/mol), and the number of conformers within 6 kcal/mol of the global minimum grows at a logarithmic rate, reaching $\sim 10^3$ conformers for molecules with twenty rotatable bonds. Across the set, this still suggests the number of rotatable bonds is a useful predictor of the number of thermally-accessible conformers, and thus the conformational entropy — even if in larger molecules, the degrees of freedom are inherently correlated.

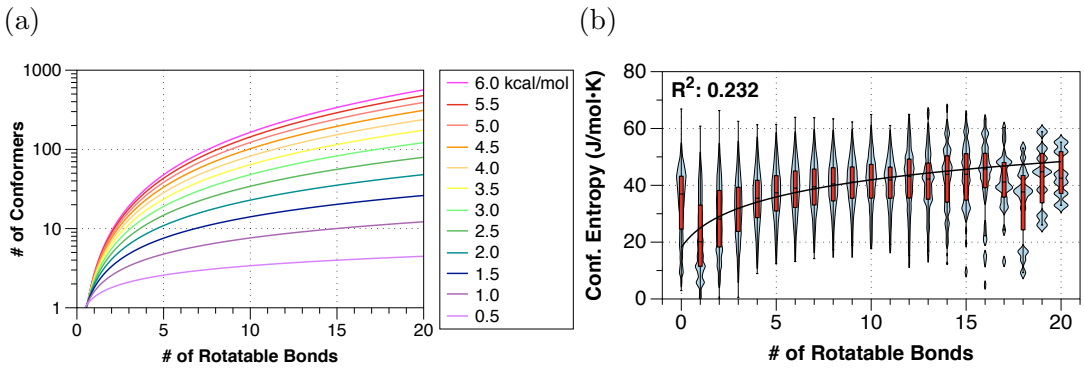


Figure 5.6: (a) Scaling of the number of conformers across the ∼93,000 molecules in the training set, on a logarithmic scale, within a given energy threshold, as a function of the number of rotatable bonds; and (b) correlation between the number of rotatable bonds, N_{rotor}, and GFN2-calculated conformational entropies, shown as violin plots for each rotatable bond bin. The line indicates a logarithmic best fit, *i.e.* $a + b\log(N_{\text{rotor}} + 1)$, with a coefficient of determination of 0.232; this highlights the need for better predictors than simply the number of rotatable bonds. Note how the linear model underestimated the conformational entropies of molecules with no rotatable bonds.

Beyond simple linear alkanes, branched alkanes and cycloalkanes can be used as models to understand other components of the conformational entropy. Both polypropylene chains and highly branched alkanes exhibit logarithmic increases in CREST-computed conformational entropy, based on the number of terminal CH_3 groups (see Figures 5.7a and 5.7b). Note that methyl groups are known to increase entropy as hindered rotors (Irikura, 1998, 2020). The magnitude of the methyl rotor entropies are higher from the CREST/GFN2 ensembles than previous quantum chemical estimates (*i.e.* 9.1 J/mol·K from CREST/GFN2 vs. 6.8 J/mol·K from HF/6-31G(d) using a hindered rotor model) (Irikura, 1998, 2020), but reflect that beyond iso-pentane, correlations between multiple CH_3 groups slow the increase in conformational entropy to logarithmic. Similarly, while cycloalkanes have fewer torsional degrees of freedom ($N - 3$ for an N-membered ring), the CREST-computed conformational entropy increases logarithmically with the ring size (see Figure 5.7c).

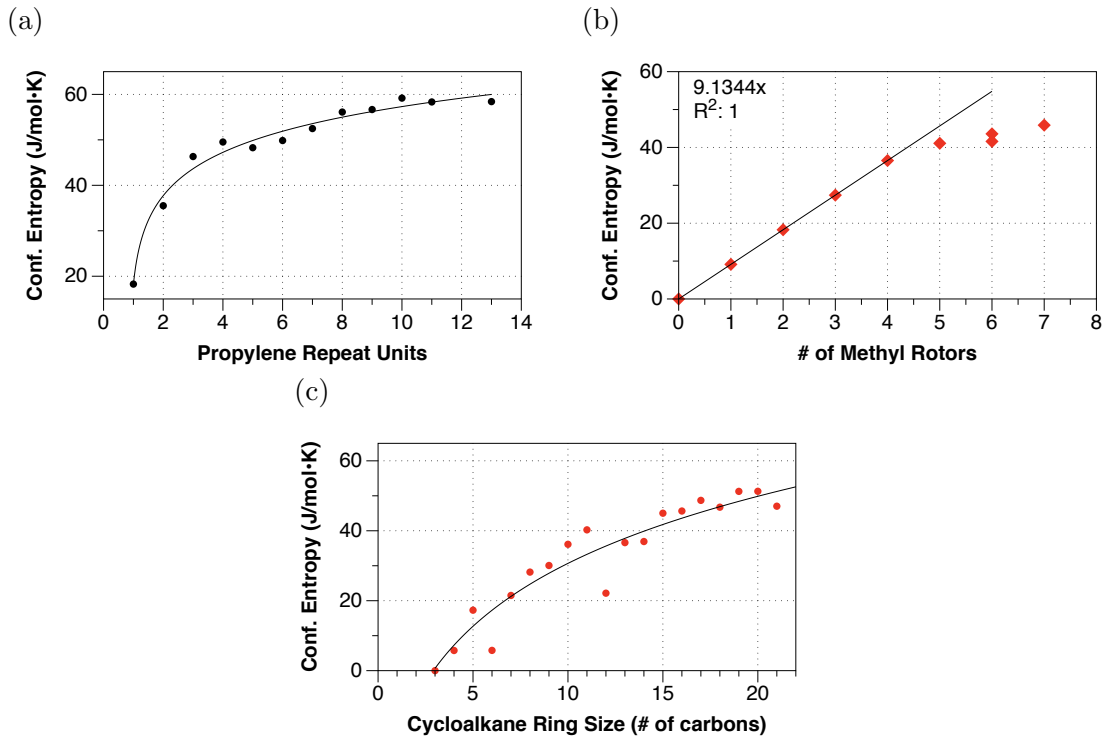


Figure 5.7: Conformational entropies for polypropylenes, branched alkanes and cycloalkanes. (a) Conformational entropies calculated for increasing lengths of polypropylene chains, as a function of the number of repeat units, illustrating approximately logarithmic increase; and (b) branched alkane chains as a function of the number of terminal methyl rotors. (c) Conformational entropies calculated for increasing ring size of cycloalkanes, n-C_nH_{2n}. The conformational entropies tends to grow logarithmically with ring size.

Building from the simple alkanes, it can be seen that the conformational entropy has multiple components based on the torsional degrees of freedom, including rotatable bonds, terminal methyl groups, and correlated motions in flexible rings, such as the cycloalkanes.

Rings can also be joined together, forming more complex spiro, fused and bridged rings. Figure 5.8 shows good correlation between the total ring flexibility, R_f^{Total}, and the GFN2-computed conformational entropies of a set of complex rings, with a Pearson correlation coefficient, $R^2 = 0.7$. Note that the complex rings set contains 70 molecules (see Appendix C.1), including monocycles, fused rings, bridged rings and spiro rings. They do not contain any acyclic rotatable bonds.

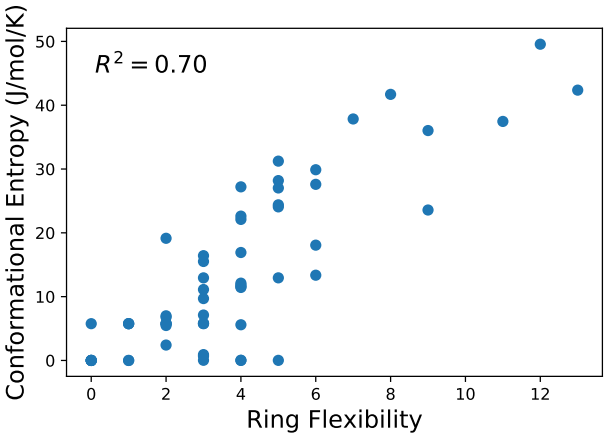


Figure 5.8: Relationship between GFN2-computed conformational entropy and total ring flexibility, R_f^{Total}. Conformational entropies of 70 simple molecules, including unsubstituted monocyclic heterocycles, fused rings, bridged rings, spiros, and rings with simple substituents (*e.g.*, hydroxyl, carbonyl) were calculated. Note that these molecules do not contain any rotatable bonds and methyl groups. The total ring flexibility shows good correlation with conformational entropy, with $R^2 = 0.70$.

5.3.1 Intramolecular Interactions

Beyond the number of degrees of freedom in molecules, formation of intramolecular interactions tend to reduce the conformational flexibility of molecules. Intramolecular hydrogen bonds were frequently ($\approx 36\%$) observed in the GFN2-computed low energy structures, followed by π-π interactions ($\approx 6.8\%$). Face-to-Face and parallel π-π stacking were the dominant forms, see Figure 5.9. That said, T-shaped π-stacking was less frequently observed in the database, and therefore was not included in our analysis. Furthermore, molecules may share the same structural motifs in forming intramolecular interactions, and repeating motifs, *i.e.* subgraph sharing same size and same atom types, were removed in the following analysis. The atom types are defined by the atom's element, hybridization, and whether the atom is in ring. Figure 5.10 shows the distribution of the path length of the motifs containing hydrogen bonds and π-π stacking. Note that the path length is the number of bonds between the terminal atoms, and the bond between the donor atom and hydrogen atom was not counted. The path length of the motifs containing intramolecular hydrogen bonds was widely varying, and ranged from 3 to 35. Multiple intramolecular interactions were found in large motifs, *i.e.* subgraph with more than 10 bonds. On the other hand, the distribution of the path length of π-π stacking motifs was skewed. Similarly, long range π-π stacking was rare and normally required several intramolecular interactions

to facilitate their formation. To gain structural insights into these path characteristics, I performed junction analysis and summarised the results below.

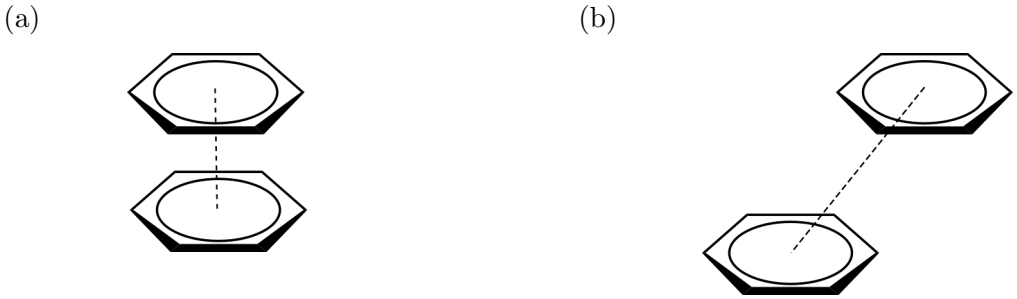


Figure 5.9: π-π stacking interactions: (a) face-to-face; and (b) parallel.

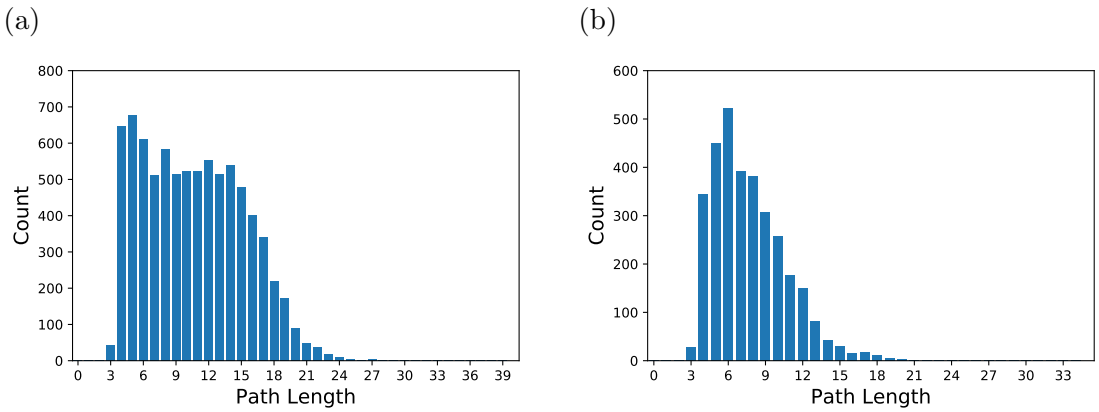


Figure 5.10: Distribution of the path length in (a) hydrogen bonds motifs and (b) π-π stacking motifs. The path length is the number of bonds between terminal atoms. Note that the bond between the donor atom and hydrogen atom in hydrogen bond motifs was not counted.

5.3.1.1 Hydrogen Bonds

As mentioned in Section 5.2.3.4, the motifs were grouped according to the acceptor types. Motifs containing up to nine bonds (ten atoms) were considered. The acceptor groups chosen—carbonyl oxygen, hydroxyl oxygen, alkxoy oxygen, acyclic and cyclic nitrogens, covered more than 90% of the observed intramolecular hydrogen bonds. Figure 5.12 shows that carbonyl and hydroxyl groups are frequently observed in the motifs.

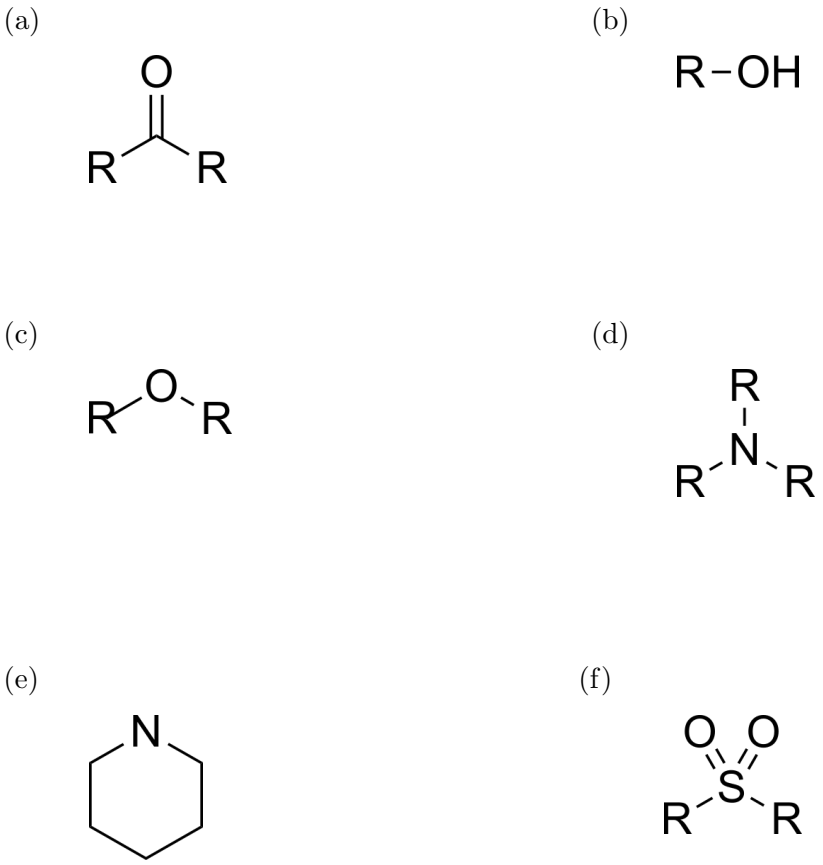


Figure 5.11: Example of hydrogen bond acceptors: (a) carbonyl group; (b) hydroxyl group ; (c) alkoxy group; (d) acyclic nitrogen; (e) heterocycle nitrogen; (f) sulfonyl group.

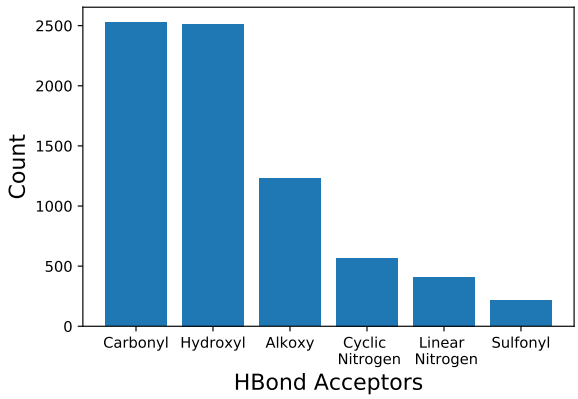


Figure 5.12: Count of unique intramolecular hydrogen bond structural motifs containing six different hydrogen bond acceptors in the dataset. Motifs with carbonyl or hydroxyl groups as hydrogen bond acceptor were frequently observed.

The junction analysis revealed that carbonyl acceptors usually form intramolecular hydrogen bonds with hydroxyl (sp^3 hybridized oxygen) groups, ester oxygen (sp^2 hybridized oxygen), and amide nitrogen (sp^2 nitrogen), as shown in Figure 5.13a and Figure 5.14. A large variety of atom types were observed along the path. Similarly, hydroxyl and alkoxy oxygen acceptors have a high propensity to form hydrogen bonds with hydroxyl group, as shown in Figures 5.13b and 5.13c.

There are approximately 5.4% and 7.6% motifs containing acyclic and cyclic nitrogen acceptors respectively. For motifs with cyclic nitrogen acceptors, I divided into two classes: (i) sp^3 hybridized nitrogen acceptor and (ii) sp^2 hybridized nitrogen acceptor. The hybridization of the heterocycle nitrogen imposed different geometrical constraints to the system, as reflected in the path variation shown in Figures 5.13e and 5.13f.

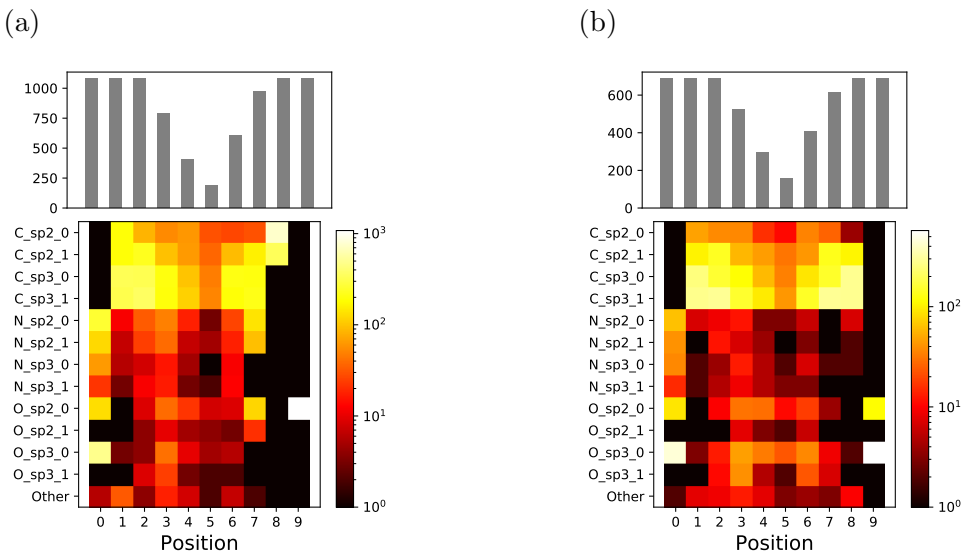


Figure 5.13: Positional analysis for a given intramolecular hydrogen bond acceptor type, showing counts of molecules with each type of atom. Counts of motifs containing each kind of intramolecular hydrogen bond are shown as histograms above each heat map, where black represents zero, and a color spectrum from dark red, red, orange, and yellow to white maps to the number of motifs in each category. Position 0 and 9 are the donor atom and acceptor atom respectively. Plots are shown for each type of H-bond acceptor: (a) carbonyl; (b) hydroxyl; (c) alkoxy; (d) acyclic nitrogen (e) heterocyclic sp^3 nitrogen; and (f) heterocyclic sp^2 nitrogen. The atom types are defined by the atom's element, hybridization, and whether the atom is in ring, *e.g.,* "C_sp3_0" is an sp^3 carbon not in a ring. In general, a large variety of atom types were observed in the motifs with carbonyl, hydroxyl or alkoxy acceptors. The hybridization state of heterocycle nitrogen acceptor induced different geometrical constraints, and gave rise to the variation in path.

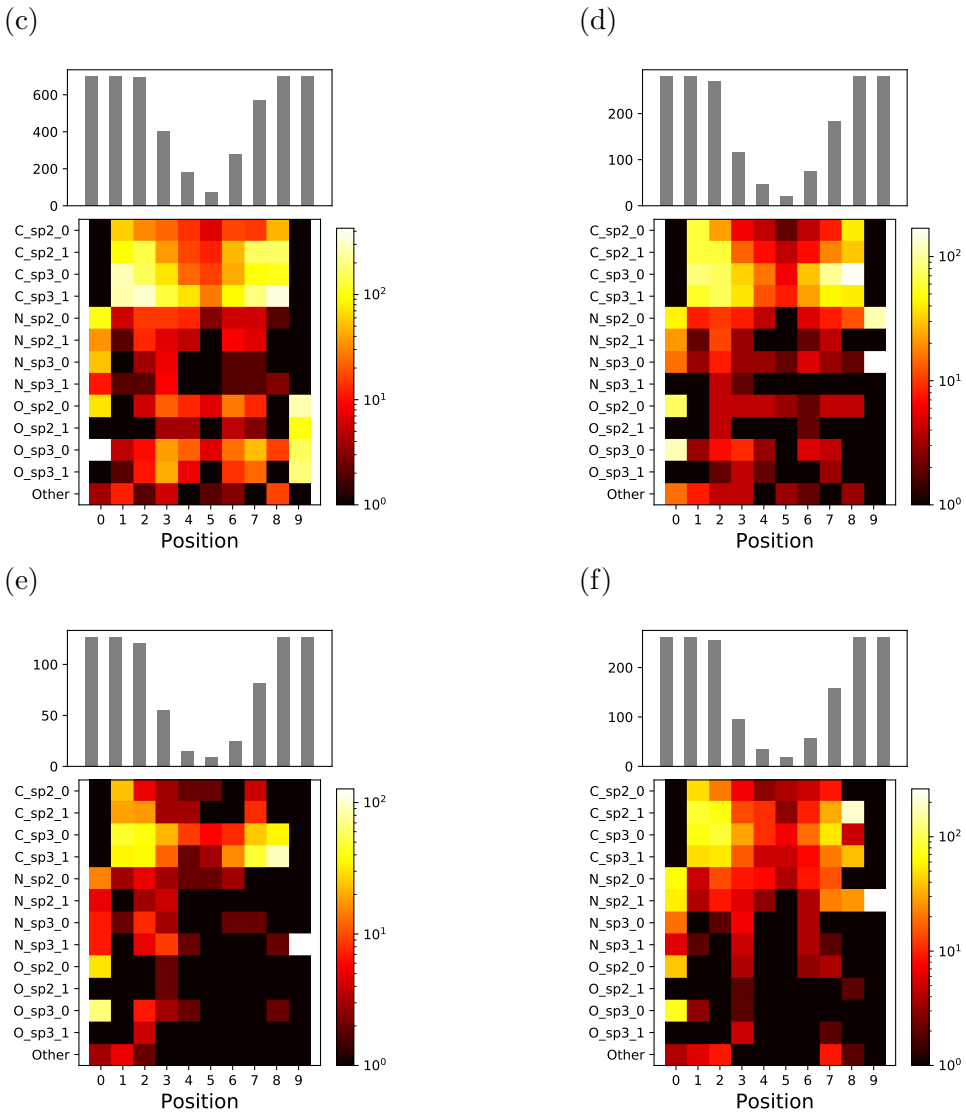


Figure 5.13: (Continued)

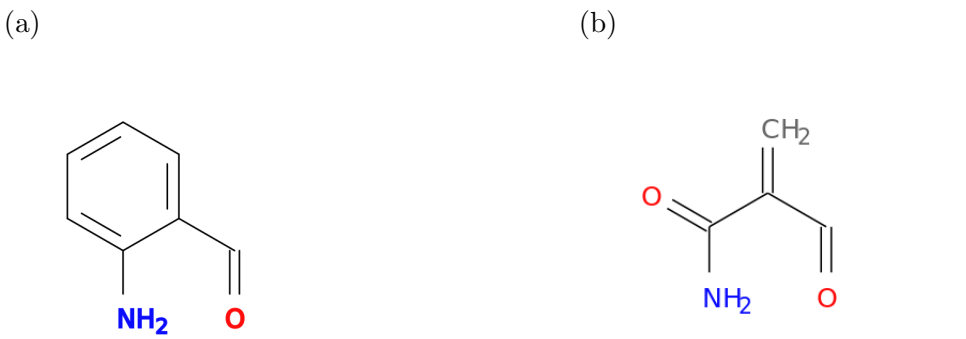
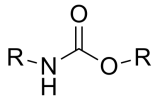


Figure 5.14: Intramolecular hydrogen bonds motif examples

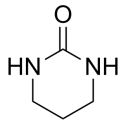
5.3.1.2 Face-to-Face and Parallel π-π stacking

Motifs with up to 16 atoms were considered in the analysis. Position 0 and 15 are the aromatic ring atoms. Six functional groups were considered in π-π stacking, including carbamate, urea, ketone, ester, ether and amide, which contributed 80% of the observed π-π stacking. Figure 5.16 shows that amide group was widely observed, followed by ether group.

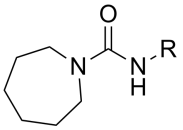
(a)



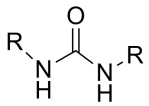
(b)



(c)

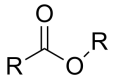


(d)

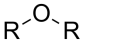


(e)

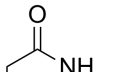
(f)



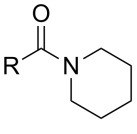
(g)



(h)



(i)



(j)

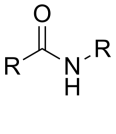


Figure 5.15: Example of functional groups involving in π-π stacking interactions: (a) carbamate group; (b) cyclic urea; (c) urea with one nitrogen in a ring; (d) acyclic urea; (e) ketone; (f) ester; (g) ether; (h) cyclic amide; (i) amide with nitrogen in a ring; (j) acyclic amide.

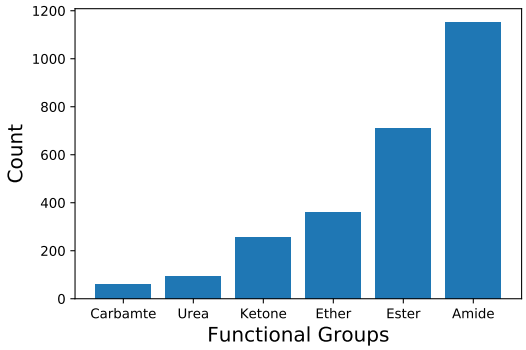


Figure 5.16: Counts of unique π-π stacking structural motifs containing six different functional groups. Amide was widely observed in the motifs. Note that the motif may contain more than one functional groups.

Figure 5.17a shows the position of carbamate carbons and the orientation of the C−O bond was highly conserved. Its paired functional group, ester, was commonly observed near the other end and the direction of the C−O bond was also conserved. Three forms of urea were considered, including cyclic ureas (type 1), ureas with one nitrogen in a ring (type 2), and acyclic ureas (type 3), see Figure 5.15. The positional preferences varied between them. Figure 5.17b shows that the position of the cyclic urea was conserved, while Figure 5.17d shows widely varying positional preferences of acyclic amide. Note that the acyclic amides often came in pairs. Amide was the common paired functional group in the motifs containing cyclic ureas or ureas with one nitrogen in a ring, see Figures 5.17b and 5.17c respectively.

Figure 5.17e shows that ketone group was frequently found in a short path, and typically came after the aromatic rings. For longer paths, it was found in the middle of the path and often formed intramolecular hydrogen bonds with hydroxyl group, in order to support longer range π-π stacking and other intramolecular interactions. It paired functional groups included ether and ester. The ether oxygen was normally found next to the terminal aromatic rings.

Figure 5.17f shows multiple preferred positions of ester carbon, and the orientation of the C−O single bond was position dependent. Its paired functional group included sulfonamide group and ether. The position of sulfur in sulfonamide group and the orientation of S−N group were highly conversed. The positions of ether oxygen, however, showed a huge variation.

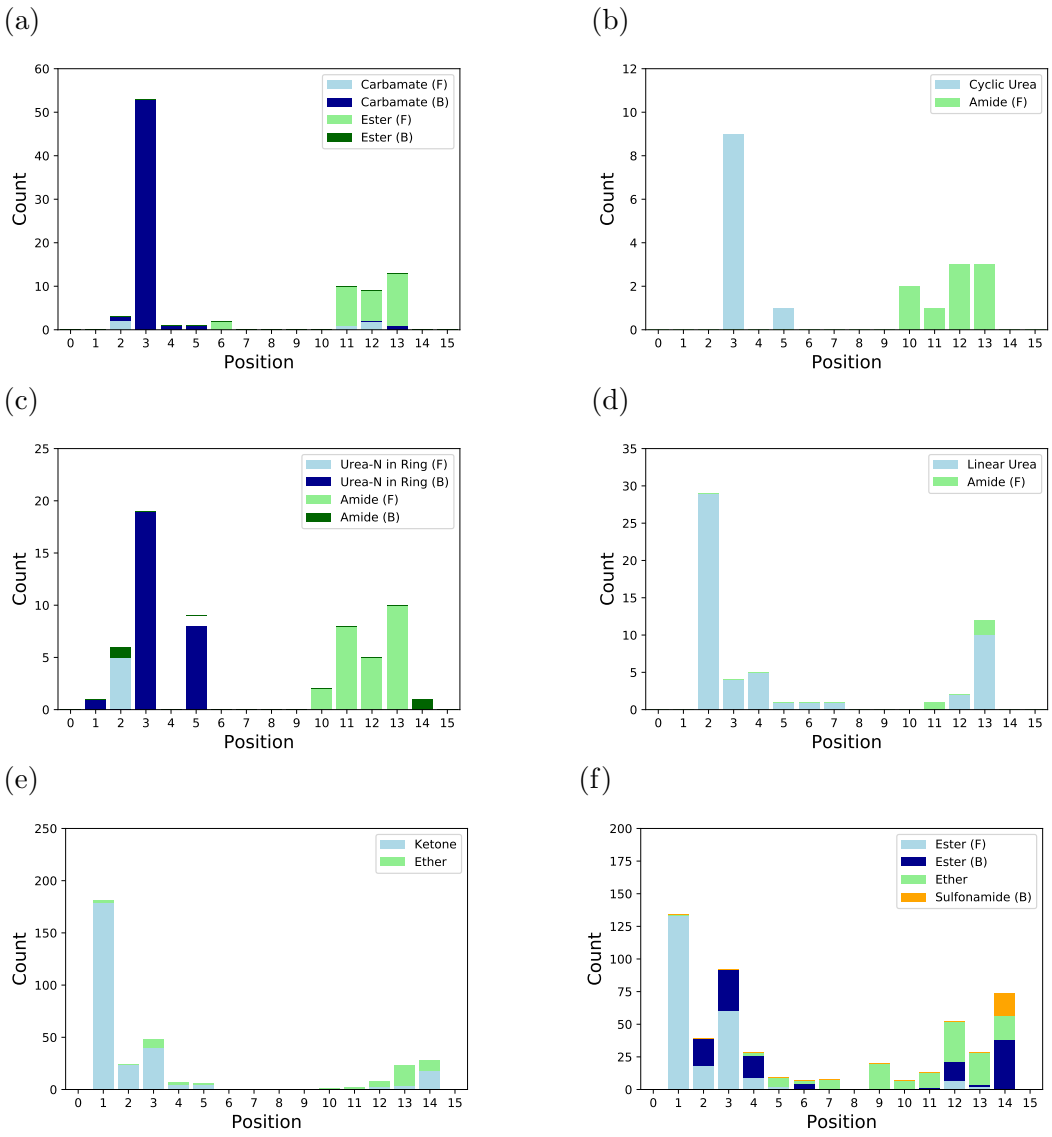


Figure 5.17: Counts of functional groups by position in the training set. Atoms 0 and 15 belong to the aromatic rings. The functional groups are: (a) carbamate; (b) cyclic urea; (c) urea with one nitrogen in a ring; (d) acyclic nitrogen; (e) ketone; (f) ester; (g) ether; and (h) acyclic amide (type 0), (i) cyclic amide (type 1); (j) amide with nitrogen in a ring (type 2); (k) acyclic amides with one or more ring bonds along the shortest connecting path. The orientations of the C−O bond in carbamates and esters, the S−N bond in sulfonamides, and the C−N bond in amides are specified thus: (F) indicates that the oxygen or nitrogen atoms are *followed* by the carbon or sulfur atom, while (B) indicates that the oxygen or nitrogen atoms comes *before* the carbon or sulfur atom. The positions of carbamate, urea, and ketone groups tend to be conserved, while amide positions show a huge variation.

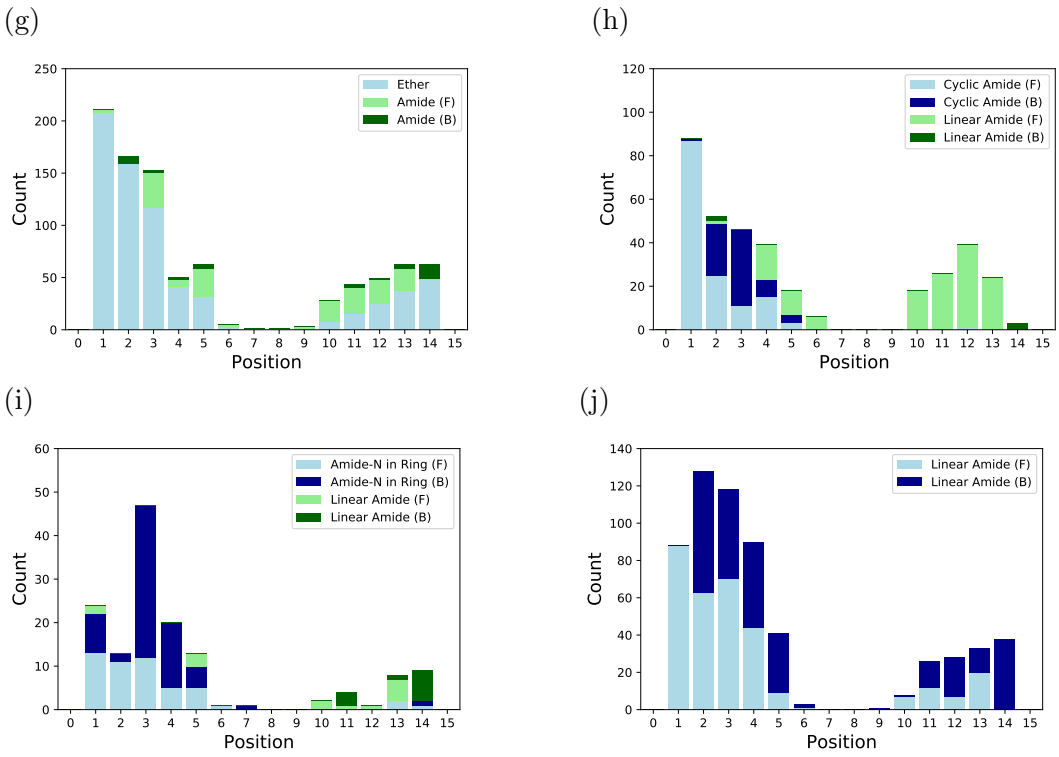


Figure 5.17: (Continued)

Similar to ureas, three different types of amide groups, including cyclic amides, amides with the amide nitrogen in a ring and acyclic amides, were considered. Figures 5.17h to 5.17j show that the amide carbon position was predominantly found at the first, second or third positions, however, the orientation of the C−N bond was position dependent. Cyclic amide and acyclic amide were typically came in pairs, as shown in Figure 5.17h.

Despite the high variation in positional preferences in some functional groups, some rules were generalised to identify potential intramolecular π-π stacking. To assess the generalizability of these positional preferences, I compared the positional preference of amide carbon in the training set and also in the peptides set. Figure 5.18 shows different positional preferences in the training set and peptide set, and this suggests that different models may be required for peptides.

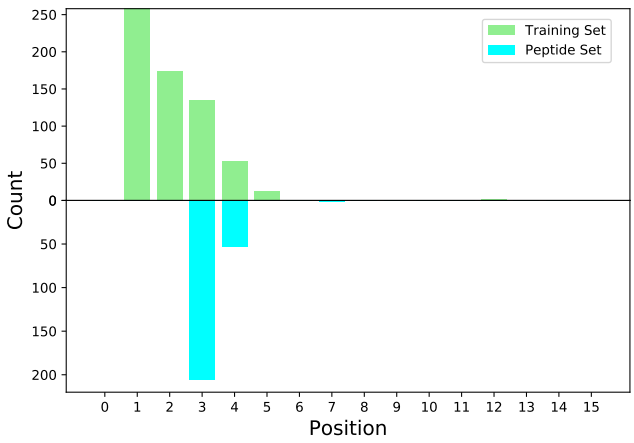


Figure 5.18: Amide positional preferences in the training set and peptide set.

The junction analysis provided insights into the positional preferences of common functional groups in molecules with intramolecular hydrogen bonds and π-π stacking interactions. The over-represented groups (see the thresholds in Table 5.6) were then used to identify potential interactions in unseen datasets and compute the foldability descriptors defined in Equations 5.4 and 5.5. These descriptors and path features were used to develop models to predict conformational entropy. The model performances are discussed next.

5.3.2 Models to Predict Conformational Entropy and their Performance

I compared the proposed model LR-Best, Equation 5.6, with a baseline model, LR-1, Equation 5.7, and multiple statistical and machine learning models, including linear regression, LASSO, ridge regression, kernel ridge regression (KRR), and neural networks (NN).

Before looking at the predictive performance, I assessed the model assumptions in my proposed and baseline linear models, as illustrated in Figure 5.19. For the baseline model, there was no evidence the normality assumption was violated, however, the residuals plots in Figure 5.19b shows that the model underestimated the conformational entropies of molecules with no rotatable bonds, and the variance of the residuals depended on the fitted values, suggesting the homoscedasticity and that the linear relationship did not hold. On the other hand, I did not have sufficient evidence that the model assumptions were violated for the proposed model (LR-Best).

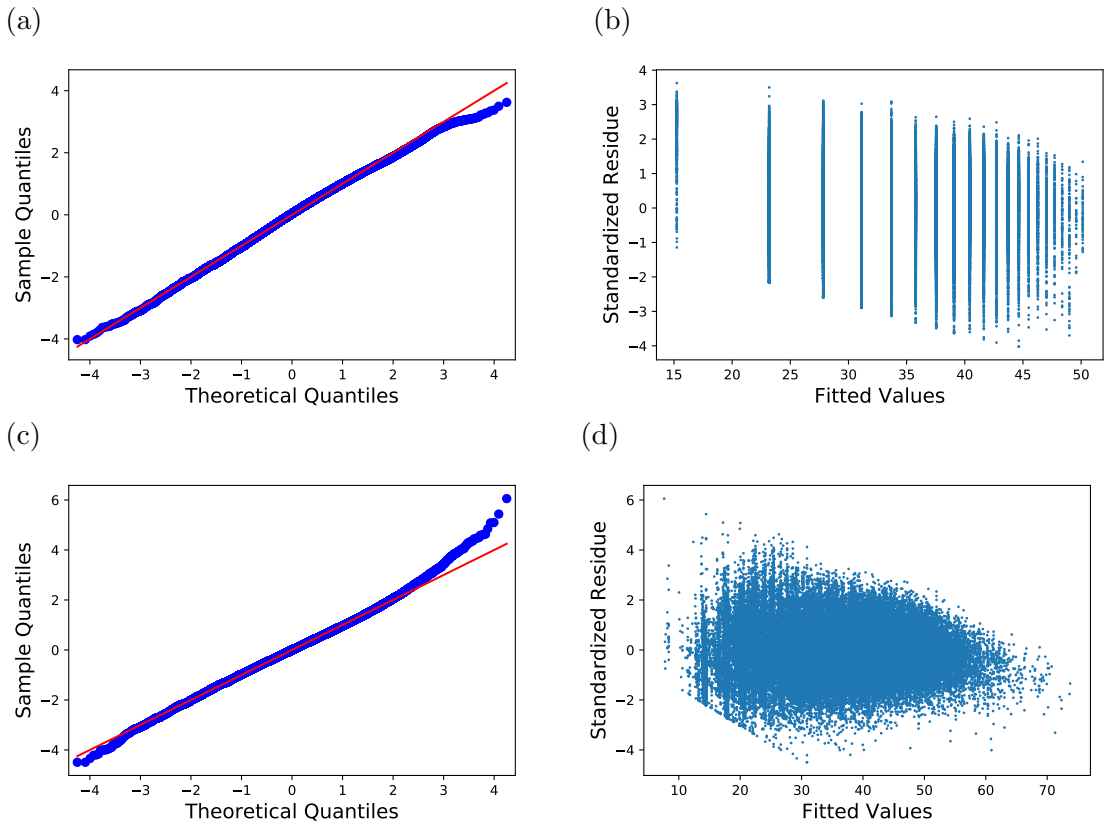


Figure 5.19: Models Diagnostics: (a) LR-1 Q-Q plot (b) LR-1 residual plot. The predicted conformational entropies of molecules with no rotatable bonds are underestimated and the variance of the residuals are not equal. (c) LR-Best Q-Q plot (d) LR-Best residual plot. There is no evidence that the model assumptions are violated for LR-Best model.

Table 5.7: LR-Best Models Summary. The parameters associated with foldability (F_{Hbond}, $F_{\pi-\pi}$), and counts of functional groups (N_{SG}) are negative, indicating the conformational entropy decreases as the these variables increase. All parameters differ significantly from zero. The negative value in parameter associated with total ring flexibility is inconsistent with my previous results.

Coefficient	Estimate	Std. Error	t	Pr(> | t |)
Intercept	8.15	0.08	99.44	$< 10^{-3}$
$\log(N_{\text{rotor}} + 1)$	9.02	0.05	180.93	$< 10^{-3}$
$\log(N_{\text{Methyl}} + 1)$	13.03	0.04	375.39	$< 10^{-3}$
$\log(N_{\text{SG}} + 1)$	-2.44	0.05	-50.77	$< 10^{-3}$
$\log(F_{\text{HBond}} + 1)$	-1.72	0.05	-32.47	$< 10^{-3}$
$\log(F_{\pi-\pi} + 1)$	-0.59	0.03	-17.34	$< 10^{-3}$
$\log(R_f^{\text{Total}} + 1)$	-0.23	0.03	-7.51	$< 10^{-3}$
$R^2 = 0.714$				

Table 5.7 shows negative values of the parameters associated with the number of specified functional groups, foldability with intramolecular hydrogen bonds, and π-π stacking, which suggest that the conformational entropy decreases as these variables increase. Small p-values from the t-tests indicate these parameters are significantly different from zero. Surprisingly, the parameter associated with the ring flexibility is slightly negative, which is not consistent with observations in cycloalkanes (Figure 5.7c) and the small cyclic molecules subset shown in Figure 5.8. This indicates our proposed descriptor, ring flexibility, may not fully capture the conformational entropy of complex rings in the training set. The LR-Best model was also in close agreement with the GFN2-calculated conformational entropy in the training set, with a coefficient of determination, $R^2 = 0.715$.

To assess the predictive power of all models, I calculated the mean absolute error between the model-predicted and GFN2-computed conformational entropies for two independent test sets, ZINC-I and the peptides set. The proposed linear model (LR-Best) outperformed the other machine learning models (LASSO, Ridge, KRR and DNN), giving a mean absolute error of 4.77 and 4.65 J/mol $\cdot$ K respectively (see Table 5.8). Figure C.1 in Appendix C shows the correlation between the predicted entropies and the GFN2-computed entropies in both ZINC-I and peptide test sets for all models. The LR-Best model gave the highest correlation, with $R^2 = 0.75$ and $R^2 = 0.61$ respectively. The ECFP6 fingerprints only consider local information about any given atom, and the global topological information including longer range intramolecular interactions therefore cannot be encapsulated in such representations. This limits the predictive power of models based on such short range features. The KRR approach failed to obtain good predictions in peptides, as the cyclic peptides are likely too dissimilar from molecules in the training data.

Table 5.8: Entropy prediction model performance. Comparison of the mean absolute error (MAE) between the model-predicted and GFN2-computed conformational entropies, in J/mol$\cdot$ K, for the training set and both test sets, namely the ZINC-I and CTP sets. LR-1 is a single-variable linear model, with the number of rotatable bonds as the sole explanatory variable. LR-Best gives the lowest MAE in both test sets.

Model	Training (MAE)	ZINC-I Set (MAE)	CTP Set (MAE)
LR-1	8.67	8.83	9.00
LR-Best	5.16	**4.77**	**4.65**
LASSO	5.55	5.47	6.76
Ridge	4.95	5.29	5.83
KRR	5.90	5.87	8.79
DNN	5.22	5.26	6.98

5.4 Summary

In summary, my analysis showed that the conformational entropy of small molecules increases logarithmically with the number of degrees of freedom in the small molecules. Despite the possible number of conformers increasing exponentially with the number of rotatable bonds, inherent correlation between multiple rotatable bonds and terminal CH_3 groups restricts the number of thermally-accessible conformations greatly. Intramolecular interactions such as π-π stacking and intramolecular hydrogen bonds further reduce the number of thermally-accessible conformers, and decrease the conformational entropy as a result. Such effects, here in small molecules, relate to Levinthal's paradox which states finding the native folded state of a protein by a random search among all possible configurations can take an enormously long time, and the energy landscapes found in protein folding (Levinthal, 1968; Zwanzig et al., 1992; Dill and Chan, 1997). The contribution of ring entropy from flexible rings has to be assessed carefully. A standardized atom numbering scheme was introduced to study the path characteristic of hydrogen bonds and π-π stacking and sets of rules were generalised to identify potential interactions in unseen molecules. I introduced a new descriptor called *foldability* which took the effect of intramolecular interactions on conformational entropy into account, and thus improved the prediction of the conformational entropy component of standard molecular entropy. The resulting linear model, based on a physical understanding of the various contributions to conformational entropy, outperformed current machine learning methods that used ECFP-based features, and gave a mean absolute error of 4.8 J/mol· K, or $\approx$ 0.34 kcal/mol at 300 K. My approach facilitates the calculation of thermodynamic properties and provides insights into the effect of intramolecular interactions on conformational preferences and the intrinsic correlation between rotors in a molecule. This work can also be extended to predict the change in solvation entropy as well as ligand conformational entropy upon protein-ligand binding, and thus provide better estimates of binding free energies for drug discovery (Head et al., 1997; Chang et al., 2007).

Chapter 6

Conclusions and Future Directions

6.1 Summary

In Chapters 1 and 2, I gave an overview of the computer-aided drug and materials design process, and highlighted one of its key challenges: sampling diverse low energy conformers and finding the lowest energy conformer of a flexible small molecule. I introduced the basic concepts of molecular conformation and their representation, and described various systematic and stochastic conformer sampling methods. Evaluating conformational energy is a complementary task to rank conformers for downstream applications. The most accurate method for energy evaluation requires solving the Schrödinger equation, and different levels of numerical approximations were discussed, including the standard quantum mechanics methods, semi-empirical quantum mechanics methods, molecular mechanics force fields and machine learning methods. Two of the most significant challenges arise in two ways: (i) the enormous conformational space of flexible molecules, and (ii) the high computational cost when standard quantum mechanics methods are used for energy evaluations.

The advancement in hardware and the rise of machine learning allow us to tackle this challenge. One of these machine learning techniques, namely Bayesian optimisation, was used (Brochu et al., 2010; Snoek et al., 2012). It is a sequential strategy for global optimisation. A surrogate model, typically Gaussian Process (GP) (Williams and Rasmussen, 2006), is used to approximate the function of interest, followed by an evaluation at a new query point. This process is repeated until the stopping criterion is reached. The search strategy only requires pointwise evaluation, while it does not make any assumptions about the functional forms of the objective being evaluated. This technique is widely used in the machine learning community for hyper-parameter

tuning (Snoek et al., 2012; Law et al., 2019), and I applied this technique to search for the lowest energy conformation of a molecule (Chan et al., 2019, 2020a).

In addition to sampling, I introduced a computational framework to analyse conformers and model their conformational preferences. In conformational analysis, internal coordinates, *i.e.* bond lengths, bond angles and torsion angles, are frequently used, as they describe the relative positions between atoms, so as to eliminate the effect of translation and rotation along x-, y-, z-axes. Conventional summary statistics and distance measures, however, are not readily applicable to these internal coordinates, due to the circular nature of the variables. The basics of circular data analysis was introduced, including circular distance, circular correlation, and the von Mises distribution, which is an analogue of normal distribution on a circle. I also discussed multiple metrics to examine the quality of the sampled conformations.

In Chapter 3, I first assessed the effectiveness of the Bayesian optimisation (BO) technique in searching the lowest energy conformation of a molecule with up to six rotatable bonds. I introduced and compared the Bayesian optimisation algorithm (BOA) (Chan et al., 2019) with a systematic search method, Confab (O'Boyle et al., 2011b), and a uniform random search, using a molecular mechanics force field energy, namely the Merck molecular force fields (MMFF94) (Halgren, 1996; Halgren and Nachbar, 1996). In BOA, Gaussian Process (GP) was used as the surrogate model, with a locally periodic kernel, which encapsulated the periodicity of the torsion potentials derived from experimental determined X-ray crystal structures. Standard acquisition functions of expected improvement (EI) (Mockus et al., 1978) and GP lower confidence bound (GP-LCB) (Srinivas et al., 2009) were used. The BOA algorithm outperformed both systematic enumeration and uniform search. It frequently found lower energy structures than the lowest energy conformations found by the systematic search. As expected, uniform random search performed worst, as it suffered from the combinatorial explosion of number of conformers with increasing number of rotatable bonds. More importantly, the BOA required orders of magnitude fewer energy evaluations to reach the optimal solution. Its performance improved when the number of energy evaluation increased, despite the cubic time complexity of BOA.

The flexible Bayesian optimisation framework allows us to further incorporate our prior knowledge through the acquisition function. Adjacent torsion angles are inherently correlated to avoid steric clashes and align intramolecular interactions, such as hydrogen bonds and $\pi - \pi$ stacking. Inspired their use in protein structure modelling,

I used bivariate von Mises mixture models to capture the correlation between adjacent torsion angles. A new knowledge-based acquisition function, knowledge-based expected improvement (KEI) was proposed (Chan et al., 2020a). It is a product of standard EI and a collection of bivariate von Mises mixture models of substructures containing correlated torsions. The proposed acquisition function biased the search towards to low energy regions, and accelerated the early stage of the search. BOA with KEI (BOKEI for short) was compared with BOA with standard EI (BOA-EI for short) using two energy functions: (i) geometry-optimised MMFF94, and (ii) a density functional tight binding method, GFN2 (Bannwarth et al., 2019). A genetic algorithm (GA) was also included in the benchmark assessment where geometry-optimised MMFF94 was used. BOKEI outperformed BOA-EI in both cases for molecules up to 18 rotatable bonds, and GA for molecules up to 11 rotatable bonds. BOKEI was particularly efficient in the early stage of the search, and the effect diminished as more energy evaluations were used, as both BOKEI and BOA-EI converged on the same conformation. The new approach (BOKEI) did not show substantial increase in computational time, compared with original approach (BOA-EI).

Additionally, the conformational analysis on adjacent torsion angles revealed the deficiencies of the MMFF94 force fields, which failed to capture the correlated torsions in some substructures. The discrepancy between the GFN2-computed low energy structures and the crystal structures were ascribed to the formation of intra- and intermolecular interactions.

Thus far, I mainly focused on sampling conformers that arise from rotation about acyclic rotatable bonds and the associated torsional preferences. The conformation of cyclic structures remained fixed throughout the sampling. I therefore explored the use and extensions of Cremer-Pople puckering parameters (Cremer and Pople, 1975; Cremer, 1980) to describe the geometries of a conformationally flexible ring in Chapter 4. Instead of using internal coordinates, the Cremer-Pople puckering parameters utilised a discrete Fourier transform to characterise the ring geometry with some amplitudes and phase angles. The general puckering motions, including pseudo-rotation, *i.e.* a set of intramolecular movements of the atoms leading to an indistinguishable conformation from the initial one, can be elucidated using these puckering parameters. To better understand the conformational preferences of complex bi- and polycyclic rings, I applied the concept of unique ring families (URFs) (Kolodzik et al., 2012) to decompose rings into meaningful subgroups, in conjunction with the Cremer-Pople puckering parameters. The coupled substituent orientation can be fully described by

the extension of Cremer-Pople puckering parameters with two additional orientation angles. This new framework provided a means to study flexible ring conformations quantitatively, especially macrocycles.

Furthermore, I proposed models to (i) predict substituent orientations from puckering parameters, (ii) to convert Cremer Pople puckering parameters into endocyclic torsion angles, and (iii) predict substituent exocyclic torsion angles from endocyclic torsion angles. The first model helped understand the influence of substituents upon ring puckering, while the second model provided insights into the torsional changes upon pseudo-rotation and other puckering motions. The latter model explained the local rotational dependence of the substituents' exocyclic torsion angle with respect to neighbouring endocyclic torsion angles. Using these puckering preferences and the proposed models, I developed a knowledge-based conformer sampling method for ring conformations, and showed that it could generate low energy ring conformers efficiently (Chan et al., 2020b).

In Chapter 5, I investigated the conformational entropy of small molecules. Entropy is an important thermodynamic quantity, which provides insights into the stability of molecules. However, it is computationally challenging to evaluate the entropy of flexible molecules with standard quantum mechanics methods. Different approximations have been introduced (Speybroeck et al., 2005; Ellingson et al., 2006; Zheng et al., 2011; Ghahremanpour et al., 2016; Simón-Carballido et al., 2017; Wu et al., 2019), and all required calculation of vibrational, translational, rotational and conformational entropy. Despite that the conformational entropy has the least contribution to the total entropy, the median GFN2-computed conformational entropy was of order of 2.6 kcal/mol at 300 K in my analysis, which should not be neglected (Chan et al., 2020c). The calculation of conformational entropy requires sampling of all thermally-accessible conformers for the calculation, which is thus computationally expensive. To overcome this bottleneck, I studied the factor governing the conformer populations and developed a statistical model to predict the conformational entropy of small molecules rapidly.

In my analysis, I showed that the number of conformers increases logarithmically with the number of degree of freedoms in a molecule. The degrees of freedom of a molecule include the number of acyclic rotatable bonds, the number of terminal methyl groups (which is also well-known as a hindered rotor), and the degree of freedoms in a flexible rings. Inspired from the Cremer-Pople puckering parameters (Cremer and Pople, 1975), the concept of unique ring families (URFs) (Kolodzik et al., 2012), and

the results from Chapter 4, I introduced a new descriptor called *total ring flexibility*, to describe the flexibility of rings. This descriptor showed good correlation with the GFN2-computed conformational entropy in an empirical study (Chan et al., 2020c).

Beyond the number of degrees of freedom in a molecule, the conformational population was also controlled by the chemical functionality and the shape of the molecule. Delocalisation of electrons in functional groups such as amides and esters reduced their conformational flexibility and rendered them planar. A descriptor based on the count of functional groups was included in my model. In addition, the formation of intramolecular interactions such as hydrogen bonds and $\pi - \pi$ stacking gave rise to "folded" structures and a reduction in conformational flexibility. The conformational entropies decreased as a result.

Inspired from the junction analysis (IMGT numbering scheme) in antibody modelling (Monod et al., 2004), I developed a consistent atom numbering scheme to analyse the path characteristics of molecules that consisted of π-π stacking and intramolecular hydrogen bonds. A set of rules were generalised to predict the formation of intramolecular interactions, and a descriptor called *foldability* was introduced to estimate the effect of intramolecular interactions on conformational entropies. The estimated parameters in my analysis matched our expectation that these factors decrease the conformational entropies of small molecules (Chan et al., 2020c).

Using the functional forms and the understanding of contributions to conformer population, I proposed a linear model to predict the conformational entropies of small molecules, and it provided accurate predictions, with a mean absolute error of 4.8 $\mathrm{J/mol \cdot K}$ or under 0.4 kcal/mol at 300 K, outperforming machine learning models based on extended connectivity fingerprints. This work will facilitate the rapid calculation of thermodynamic properties, and in turn advance computer-aided molecular discovery.

6.2 Future Directions

Here, I describe future directions that might build upon and enhance the work presented in this thesis.

6.2.1 Bayesian Optimisation with New Kernels for Conformer Sampling

As mentioned in Chapter 3, the L_2 norm used in the locally periodic kernel is not an ideal distance measure for circular variables. Projection of the circular variables to Cartesian space with an appropriate kernel, or applying an appropriate circular distance measure with the original kernel, will capture the circular natures of the variables and improve the accuracy of the surrogate model. Improvement in surrogate models would lead to a better estimate in the acquisition function and enhance the selection of new query points. The effect of the kernel should be carefully investigated in future.

Additionally, the ring conformations remained fixed throughout the sampling in Chapter 3. To sample ring conformers in BOA framework (Chan et al., 2019), the Cremer-Pople puckering parameters can be used as input variables. A new kernel would also be required to encapsulate the amplitudes and phases couplings. On the other hand, statistical potentials based on puckering preferences can be also embedded into the knowledge-based expected improvement acquisition function for sampling.

6.2.2 General Conformational Preferences

A large set of chemically neutral molecules have been investigated in my thesis, and the effect of charge was not considered. Future work should focus on the effect of charge on conformational preference. Furthermore, as illustrated in Chapter 3, there are occasional discrepancies between the *in vacuo* GFN2-computed low energy structures and the experimentally determined X-ray crystal structures, due to favorable intra- and intermolecular interactions present in the crystal structures. It is also well-known that conformations in solution phase differ from gas phase and crystal structures, because of solvent effects. A better understanding of the preference of intra- and intermolecular interactions in different phases will help generate relevant conformers efficiently for different computational approaches, such as molecular docking and 3D QSAR. This will in turn accelerate the drug design process.

6.2.3 Integration with Conformer Sampling Tools

The torsional preferences and the puckering preferences identified here are not limited to use in the BOA framework. They can be easily embedded into other conformer

sampling tools. For instance, a better sampling grid can be defined for systematic sampling in Confab (O'Boyle et al., 2011b) or docking methods such as AutoDock4 (Morris et al., 2009) and AutoDock Vina (Trott and Olson, 2010). Alternatively, multi-dimensional correlated torsional potentials could be incorporated into the geometry optimisation method used in RDKit's ETKDG framework (Riniker and Landrum, 2015; Wang et al., 2020).

6.2.4 2D Geometry Characterisation for Molecular Properties Predictions

As mentioned in the beginning of my thesis, the discovery process typically starts with screening a large set of molecules of interest. Statistical or machine learning models are frequently used to select potential candidates with desired properties. Some molecular properties, such as solubility and entropy, are conformer-dependent, and classical 2D descriptors are not always sufficient to provide accurate estimates. I showed that the consistent atom numbering scheme used in Chapter 5 provided accurate predictions of the presence of intramolecular interactions from molecular graphs or SMILES strings, and a new descriptor, namely foldability, was developed to predict the conformational entropies of small molecules. This atom numbering scheme can be extended and several descriptors can be derived for other molecular property predictions.

6.3 Final Words

I have studied the conformations of a diverse set of small molecules, including cyclic small molecules and peptides. I developed multiple models to elucidate the conformational preferences of acyclic and cyclic molecules, which I used to develop novel knowledge-based sampling methods. I also explored the use of Bayesian optimisation to search for the lowest energy conformation. I showed that it is more efficient than conventional search methods, and requires orders of magnitude fewer energy evaluation to reach the top candidates. Its flexible framework allows us to incorporate prior knowledge through the kernel and a novel knowledge-based acquisition function. With the increasing availability of computational power and X-ray crystal structures, the proposed analysis and modelling framework can be applied to understand the conformational preferences of more complex structures. This work will lead to better

conformer sampling tools for computer-aided molecular design, and in turn help to accelerate the drug and material discovery process.

Appendix A

Table A.1: 364 Rotatable bond SMARTS patterns and the corresponding periodicity parameters.

	Rotatable bond SMARTS pattern	Parameter
0	[O:1]=[C:2]!@;-[O:3]~[CH0:4]	1
1	[O:1]=[C:2]([N])!@;-[O:3]~[C:4]	1
2	[O:1]=[C:2]!@;-[O:3]~[C:4]	1
3	[O:1]=[C:2]!@;-[O:3]~[!#1:4]	1
4	[$(C=O):1][O:2]!@;-[c:3]~[*:4]	2
5	[$(C=O):1][O:2]!@;-[CX3:3]~[*:4]	1
6	[$(C=O):1][O:2]!@;-[CH1:3][H:4]	1
7	[$(C=O):1][O:2]!@;-[CH2:3]~[C:4]	1
8	[H:1][CX4H1:2]!@;-[O:3][CX4:4]	1
9	[C:1][CH2:2]!@;-[O:3][CX4:4]	1
10	[*:1][CX4:2]!@;-[O:3][$([CX3](=[!O])):4]	1
11	[O:1][CX4:2]!@;-[O:3][CX4:4]	1
12	[*:1][CX4:2]!@;-[O:3][CX4:4]	1
13	[cH1:1][c:2]([cH1])!@;-[O:3][S:4]	2
14	[cH1:1][c:2]([cH0])!@;-[O:3][S:4]	2
15	[cH0:1][c:2]([cH0])!@;-[O:3][S:4]	2
16	[cH1:1][c:2]([cH1])!@;-[O:3][c:4]	1
17	[cH1:1][c:2]([cH0])!@;-[O:3][c:4]	1
18	[cH0:1][c:2]([cH0])!@;-[O:3][c:4]	2
19	[cH0:1][c:2]([cH0])!@;-[O:3][P:4]	2
20	[cH0:1][c:2]([cH0])!@;-[O:3][p:4]	2
21	[cH:1][c:2]([cH])!@;-[O:3][$(C([F])([F])[F]):4]	2
22	[cH0:1][c:2]([cH0])!@;-[O:3][CX4H0:4]	2
23	[a:1][c:2]([a])!@;-[O:3][CX4H0:4]	1
24	[cH1,n:1][c:2]!@;-[O:3][CRH1:4]	2
25	[cH1,n:1][c:2]!@;-[O:3][CH1:4]	2
26	[nX2H0:1][c:2]([cH0])!@;-[O:3][CX4H0:4]	1

Continued on next page

27	[cH0:1][c:2]([nX2])!@;-[O:3][C:4]	1
28	[nX2:1][c:2]([nX2])!@;-[O:3][C:4]	2
29	[nX2:1][c:2]([nX3])!@;-[O:3][C:4]	1
30	[cH1:1][c:2]([nX3])!@;-[O:3][C:4]	1
31	[cH1:1][c:2]([nX2])!@;-[O:3][C:4]	1
32	[$([cH0]([CX3])):1][c:2]([cH1])!@;-[O:3][C:4]	1
33	[cH1:1][c:2](cO)!@;-[O:3][C:4]	1
34	[$(cO):1][c:2](cO)!@;-[O:3][C:4]	2
35	[cH0:1][c:2]([cH0])!@;-[O:3][C:4]	2
36	[cH0:1][c:2]([cH1])!@;-[O:3][C:4]	1
37	[cH1:1][c:2]([cH1])!@;-[O:3][C:4]	2
38	[a:1][c:2]!@;-[O:3][CX3H0:4]	2
39	[aH0:1][c:2]!@;-[OX2:3][!#1:4]	2
40	[!#1:1][CX4H0:2]!@;-[OX2:3][!#1:4]	1
41	[H:1][CX4H1:2]!@;-[OX2:3][!#1:4]	1
42	[C:1][CX4H2:2]!@;-[OX2:3][c:4]	1
43	[c:1][CX4H2:2]!@;-[OX2:3][C:4]	1
44	[C:1][CX4H2:2]!@;-[OX2:3][C:4]	1
45	[c:1][CX4H2:2]!@;-[OX2:3][c:4]	1
46	[!#1:1][CX4H2:2]!@;-[OX2:3][c:4]	1
47	[!#1:1][CX4H2:2]!@;-[OX2:3][C:4]	1
48	[c:1][CX4H2:2]!@;-[OX2:3][!#1:4]	1
49	[C:1][CX4H2:2]!@;-[OX2:3][!#1:4]	1
50	[!#1:1][CX4H2:2]!@;-[OX2:3][!#1:4]	1
51	[!#1:1][CX4:2]!@;-[OX2:3][!#1:4]	2
52	[$([CX3]=O):1][NX3H0:2](C)!@;-[CX4H2:3][C:4]	2
53	[$([CX3]=O):1][NX3H1:2]!@;-[CX4H2:3][C:4]	1
54	[$(S(=O)(=O)):1][NX3H0:2]!@;-[CX4H2:3][!#1:4]	2
55	[$(S(=O)(=O)):1][NX3H1:2]!@;-[CX4H2:3][!#1:4]	1
56	[$(S(=O)(=O)):1][NX3H0:2]!@;-[CX4H1:3][H:4]	2
57	[$(S(=O)(=O)):1][NX3H1:2]!@;-[CX4H1:3][H:4]	1
58	[$(S(=O)(=O)):1][NH1:2]!@;-[c:3][nX2:4]	2
59	[$(S(=O)(=O)):1][NH0:2]!@;-[c:3]([cH1])[cH1:4]	2
60	[$(S(=O)(=O)):1][NH1:2]!@;-[c:3]([cH1])[cH1:4]	1
61	[$(S(=O)(=O)):1][NH0:2]!@;-[c:3]([cH1])[cH0:4]	2
62	[$(S(=O)(=O)):1][NH1:2]!@;-[c:3]([cH1])[cH0:4]	1
63	[$(S(=O)(=O)):1][NH0:2]!@;-[c:3]([cH0])[cH0:4]	2
64	[$(S(=O)(=O)):1][NH1:2]!@;-[c:3]([cH0])[cH0:4]	2
65	[$(S(=O)(=O)):1][N:2]!@;-[c:3][a:4]	1
66	[O-:1][N+:2](=O)!@;-[c:3]([cH,nX2H0])[cH,nX2H0:4]	2
67	[O-:1][N+:2](=O)!@;-[c:3]([cH0])[cH,nX2H0:4]	2
68	[O-:1][N+:2](=O)!@;-[c:3]([cH0])[cH0:4]	2

Continued on next page

69	[O-:1][N+:2](=O)!@;-[c:3][a:4]	2
70	[cH0:1][c:2]([cH0])!@;-[NX3H1:3][C,c:4](~[N,n]...	2
71	[cH0:1][c:2]([cH1])!@;-[NX3H1:3][C,c:4](~[N,n]...	1
72	[cH0:1][c:2]([nX2H0])!@;-[NX3H1:3][C,c:4](~[N,...	2
73	[cH1:1][c:2]([cH1])!@;-[NX3H1:3][C,c:4](~[N,n]...	2
74	[nX2H0:1][c:2]([nX2H0])!@;-[NX3H1:3][C,c:4](~[...	2
75	[nX2H0:1][c:2]([nX3H1])!@;-[NX3H1:3][C,c:4](~[...	2
76	[a:1][a:2]!@;-[NH1:3][C,c:4](~[N,n])(~[N,n])	2
77	[C:1][NH:2]!@;-[C:3](=[NH2:4])[NH2]	2
78	[NH2][C:1](=[NH2])[NH:2]!@;-[CH2:3][C:4]	1
79	[a:1][c:2]!@;-[NX2:3]=[$(C([NX3])n):4]	2
80	[nX2:1][c:2]!@;-[NX2:3]=[$(C([NX3])N):4]	2
81	[cH0:1][c:2]!@;-[NX2:3]=[$(C([NX3])N):4]	2
82	[cH1:1][c:2]!@;-[NX2:3]=[$(C([NX3])N):4]	2
83	[O:1]=[C:2]([NH1])!@;-[NX3H1:3](C=O)[H:4]	1
84	[O:1]=[C:2]!@;-[NX3H1:3](C=O)[H:4]	2
85	[O:1]=[C:2]!@;-[NX3:3](C=O)[*:4]	2
86	[$(C=O):1][NX3H1:2]!@;-[CX3:3]=[NX2:4]	2
87	[$(C=O):1][NX3H0:2]!@;-[CX3:3]=[*H0:4]	2
88	[$(C=O):1][NX3H0:2]!@;-[CX3:3]=[*H1:4]	1
89	[$(C=O):1][NX3H1:2]!@;-[CX3:3]=[*H2:4]	2
90	[$(C=O):1][NX3H1:2]!@;-[CX3:3]=[*H1:4]	1
91	[$(C=O):1][NX3H1:2]!@;-[CX3:3]=[*H0:4]	2
92	[$(C=O):1][NX3H0:2]!@;-[CX3H1:3]=[*:4]	1
93	[$(C=O):1][NX3H1:2]!@;-[CX3H1:3]=[*:4]	1
94	[$([C](=O)):1][NX3H0:2]!@;-[CX4H2:3][$([c]([cH...	2
95	[$(C=O):1][NX3H1:2]!@;-[CX4H2:3][$([c]([cH,nX2...	2
96	[$(C=O):1][NX3H0:2]!@;-[CX4H2:3][!#1:4]	2
97	[$(C=O):1][NX3H1:2]!@;-[CX4H2:3][!#1:4]	1
98	[$(C=O):1][NX3H0:2]!@;-[CX4H1:3][H:4]	2
99	[$(C=O):1][NX3H1:2]!@;-[CX4H1:3][H:4]	1
100	[$(C=O):1][NX3H0:2]!@;-[CX4H0:3][C:4]	1
101	[$(C=O):1][NX3H1:2]!@;-[CX4H0:3][C:4]	1
102	[$(C=O):1][NX3:2]!@;-[!#1:3][!#1:4]	1
103	[$([C](=O)([$([NX3H1]),$([NX3H2])])[NX3H1]):1]...	1
104	[$([C](=O)([$([NX3H1]),$([NX3H2])])[NX3H1]):1]...	1
105	[$([C](=O)):1][NX3H1:2]!@;-[$([a]([nH0,o])):3]...	1
106	[$([C](=O)([$([NX3H1]),$([NX3H2])])[NX3H1]):1]...	2
107	[$([C](=O)):1][NX3H1:2]!@;-[c:3]([cH])[nX2H0:4]	1
108	[$(C=O):1][NX3H0:2]!@;-[c:3]([s,o])[n:4]	1
109	[$(C=O):1][NX3H1:2]!@;-[c:3]([s,o])[n:4]	1
110	[$([C](=O)):1][NX3:2]!@;-[a:3](s)[a:4]	1

Continued on next page

111	[$(C=O):1][NX3:2]!@;-[a:3][nH:4]	1
112	[$(C=O):1][NX3H1:2]!@;-[c:3]([cH0]Cl)[cH:4]	1
113	[$(C=O):1][NX3H1:2]!@;-[c:3]([cH0]F)[cH:4]	1
114	[$(C=O):1][NX3:2]!@;-[$([a]([cH1])):3][$([aH0]...	1
115	[$(C=O):1][NX3:2]!@;-[a:3][aH0:4]	1
116	[$(C=O):1][NX3H1:2]!@;-[c:3]([cH1])[cH1:4]	2
117	[$(C=O):1][NX3H0:2]!@;-[c:3]([cH1])[cH1:4]	2
118	[$(C=O):1][NX3H0:2]!@;-[c:3]([cH0])[cH:4]	2
119	[$(C=O):1][NX3H1:2]!@;-[c:3]([cH0])[cH:4]	1
120	[$(C=O):1][NX3H0:2]!@;-[c:3]([cH0])[cH0:4]	2
121	[$(C=O):1][NX3H1:2]!@;-[c:3]([cH0])[cH0:4]	2
122	[O,S:1]=[C:2]([$([NX3H1]),$([NX3H2])])!@;-[$([...	2
123	[O:1]=[C:2]!@;-[$([NX3](c([nH1])n)):3][H:4]	1
124	[O:1]=[C:2](c)!@;-[$([NX3](c([nX2H0])([nX2H0])...	2
125	[O:1]=[CX3:2]([$([NX3H1]C)])!@;-[NX3H1:3][!#1:4]	1
126	[O:1]=[C:2]([!$([NH1])])!@;-[NX3H1:3]([H:4])[$...	1
127	[O,S:1]=[C:2]([$([NX3H1]),$([NX3H2])])!@;-[$([...	2
128	[O:1]=[C:2]!@;-[NX3H0:3]([a:4])[A]	2
129	[O:1]=[CX3:2](a)!@;-[NX3H0:3][!#1:4]	2
130	[O:1]=[CX3:2]!@;-[NX3H0:3][!#1:4]	2
131	[O:1]=[CX3:2]!@;-[NX3H1:3][!#1:4]	1
132	[CH0:1][NX3:2]([CH0])!@;-[c:3][a:4]	2
133	[CH0:1][NX3:2]([CH1])!@;-[c:3][a:4]	2
134	[cH1,nX2H0:1][c:2]([cH1,nX2H0])!@;-[NX3&r:3][*:4]	2
135	[a:1][c:2]!@;-[NX3H1:3][$([CX4&r]([C;r])([C;r]...	2
136	[cH1:1][c:2]([cH1])!@;-[NX3:3][CX4:4]	2
137	[cH0:1][c:2]([cH,nX2H0])!@;-[NX3H1:3][CX4:4]	1
138	[cH0:1][c:2]([cH,nX2H0])!@;-[NX3H0:3][CX4:4]	1
139	[cH0:1][c:2]([cH0])!@;-[NX3:3][CX4:4]	1
140	[c:1][c:2](c)!@;-[NX3:3][CX4:4]	1
141	[cH1:1][c:2]([cH1])!@;-[NX3:3][a:4]	2
142	[cH1:1][c:2]([cH0])!@;-[NX3:3][a:4]	1
143	[cH0:1][c:2]([cH0])!@;-[NX3:3][a:4]	2
144	[cH0:1][a:2]!@;-[NX3H0:3][$([CX3]=O):4]	2
145	[cH0:1][a:2]!@;-[NX3H1:3][$([CX3]=O):4]	1
146	[nX2H0:1][a:2]([nX2H0])!@;-[NX3H0:3][$([CX3]=O...	1
147	[nX2H0:1][$(a([!nX2H0])([nX2H0])!@;-[NX3H1]):2...	1
148	[nX2H0:1][a:2]!@;-[NX3H1:3][$([CX3]=O):4]	2
149	[a:1][a:2]!@;-[NX3H1:3][$([CX3]=O):4]	2
150	[a:1][a:2]!@;-[NX3:3][CX4H0:4]	1
151	[a:1][a:2]!@;-[NX3:3][!#1:4]	2
152	[O:1]=[CX3:2]!@;-[nX3:3]([aH0:4])([aH0])	2

Continued on next page

153	[O:1]=[CX3:2]!@;-[nX3:3][aH1:4]	2
154	[a:1][CX3:2](=S)!@;-[NX3:3][a:4]	1
155	[!#1:1][CX3:2](=S)!@;-[NX3H0:3][!#1:4]	2
156	[!#1:1][CX3:2](=S)!@;-[NX3H1:3][!#1:4]	2
157	[!#1:1][CH2:2]!@;-[n:3][cH0:4]	2
158	[!#1:1][CH2:2]!@;-[n:3][a:4]	2
159	[cH0:1][n:2]!@;-[CX3H0:3]~[$([n,N](-a)):4]	1
160	[CX4:1][CX4H2:2]!@;-[NX3:3][CX4:4]	1
161	[C:1][CX4H2:2]!@;-[NX3:3][C:4]	1
162	[C:1][CX4:2]!@;-[NX3:3][C:4]	1
163	[!#1:1][CX4H2:2]!@;-[NX3H1:3][!#1:4]	1
164	[!#1:1][CX4H2:2]!@;-[NX3:3][!#1:4]	1
165	[!#1:1][CX4H1:2]!@;-[NX3:3][!#1:4]	1
166	[!#1:1][CX4:2]!@;-[NX3:3][!#1:4]	1
167	[!#1:1][$(S(=O)=O):2]!@;-[nX3:3]([aH1])[aH1:4]	2
168	[!#1:1][$(S(=O)=O):2]!@;-[nX3:3][aH0:4]	2
169	[c:1][S:2](=O)(=O)!@;-[NX2H0-:3]-[*:4]	2
170	[*:1][$(S(=O)=O):2]!@;-[NX3H0&r:3][*:4]	2
171	[C:1][$(S(=O)=O):2]!@;-[NX3H1:3][c:4]	2
172	[C:1][$(S(=O)=O):2]!@;-[NX3H0:3][c:4]	2
173	[c:1][$(S(=O)=O):2]!@;-[NX3H1:3][C:4]	2
174	[c:1][$(S(=O)=O):2]!@;-[NX3H0:3][C:4]	2
175	[c:1][$(S(=O)=O):2]!@;-[NX3H1:3][c:4]	2
176	[c:1][$(S(=O)=O):2]!@;-[NX3H0:3][c:4]	2
177	[C:1][$(S(=O)=O):2]!@;-[NX3H1:3][C:4]	2
178	[C:1][$(S(=O)=O):2]!@;-[NX3H0:3][C:4]	2
179	[*:1][$(S(=O)=O):2]!@;-[NX3H1:3][*:4]	2
180	[*:1][$(S(=O)=O):2]!@;-[NX3H0:3][*:4]	2
181	[!#1:1][CX3:2]!@;-[SX2:3][!#1:4]	2
182	[!#1:1][CX4:2]!@;-[SX2:3][!#1:4]	2
183	[!#1:1][CX3:2]!@;-[SX3:3][!#1:4]	2
184	[!#1:1][CX4:2]!@;-[SX3:3][!#1:4]	1
185	[!#1:1][CX3:2]!@;-[SX4:3][!#1:4]	2
186	[H:1][CX4H1:2]!@;-[SX4:3][!#1:4]	1
187	[!#1:1][CX4:2]!@;-[SX4:3][!#1:4]	1
188	[aH1:1][c:2]([aH1])!@;-[SX2:3][!#1:4]	1
189	[aH1:1][c:2]([aH0])!@;-[SX2:3][*R:4]	1
190	[aH1:1][c:2]([aH0])!@;-[SX2:3][!#1:4]	1
191	[aH0:1][c:2]([aH0])!@;-[SX2:3][!#1:4]	1
192	[!#1:1][c:2]!@;-[SX2:3][!#1:4]	2
193	[!#1:1][c:2]!@;-[SX3:3][!#1:4]	2
194	[aH1:1][c:2]([aH1])!@;-[SX4:3][!#1:4]	2

Continued on next page

195	[aH0:1][c:2]([aH1])!@;-[SX4:3][!#1:4]	1
196	[aH0:1][c:2]([aH0])!@;-[SX4:3][!#1:4]	2
197	[O:1]=[CX3:2]([NH1])!@;-[CH2:3][CX3:4]=O	1
198	[O:1]=[CX3:2]([NH1])!@;-[CH2:3][C:4]	1
199	[$([CX3]([C])([H])):1]=[CX3:2]([C])!@;-[CH2:3]...	1
200	[$([CX3]([C])([H])):1]=[CX3:2]([H])!@;-[CH1:3]...	1
201	[$([CX3]([C])([H])):1]=[CX3:2]([H])!@;-[CH2:3]...	1
202	[N:1][C:2](=O)!@;-[CX4H2:3][CX4H2:4]	1
203	N[C:2](=[O:1])!@;-[CH2:3][N:4]	2
204	[O:1]=[C:2]([O-])!@;-[CX4H1:3][H:4]	1
205	[CX3H2:1]=[CX3:2]!@;-[CX3:3]=[C:4]	1
206	[CX3:1]=[CX3:2]!@;-[CH2:3][OX2:4]	1
207	[CX3:1]=[CX3:2]!@;-[CH1:3](C)[C:4]	1
208	[CX3:1]=[CX3:2]!@;-[CH2:3][C:4]	1
209	[CX3:1]=[CX3:2]!@;-[CH2:3][c:4]	1
210	[CX3:1]=[CX3:2]!@;-[CH2:3][!#1:4]	1
211	[O:1]=[CX3:2](O)!@;-[CX3:3]([$([NH1,NH2,CH2])]...	2
212	[O:1]=[CX3:2]!@;-[CX3:3]=[O:4]	1
213	[CX3R:1]=[CX3R:2]!@;-[CX3:3]=[CX3:4]	2
214	[CX3H0:1]=[CX3H0:2]!@;-[CX3:3]=[CX3H0:4]	1
215	[CX3H0:1]=[CX3H0:2]!@;-[CX3H0:3]=[CX3:4]	2
216	[CX3H0:1]=[CX3:2]!@;-[CX3H0:3]=[CX3:4]	1
217	[CX3H0:1]=[CX3H0:2]!@;-[CX3:3]=[CX3:4]	1
218	[CX3:1]=[CX3:2]!@;-[CX3:3]=[CX3:4]	1
219	[*^2:1]~[C^2:2]([!H])!@;-[C^2:3]~[*^2:4]	2
220	[*^2:1]~[C^2:2]!@;-[C^2:3]~[*^2:4]	2
221	[O:1]=[CX3:2]!@;-[CX4&r3:3]!@[!#1:4]	2
222	[O:1]=[CX3:2]!@;-[CX4H1&r3:3][H:4]	2
223	[OX2:1][CX4H2:2]!@;-[CX4H2:3][N&r:4]	1
224	[OX2:1][CX4H2:2]!@;-[CX4H2:3][N:4]	1
225	[OX2:1][CX4:2]!@;-[CX4:3][N:4]	1
226	[OX2:1][CX4H2:2]!@;-[CX4H2:3][OX2:4]	1
227	[OX2:1][CX4:2]!@;-[CX4:3][OX2:4]	1
228	[!#1:1][CX4&r:2]!@;-[CX4&r:3][!#1:4]	1
229	[!#1:1][CX4H2:2]!@;-[CX4H2:3][!#1:4]	1
230	[!#1:1][CX4:2]!@;-[CX4:3][!#1:4]	1
231	[OX2:1][CX4H2:2]!@;-[CX3:3]([$([NX3H1,NX3H2])]...	2
232	[OH1:1][CX4:2]!@;-[CX3:3]=[O:4]	2
233	[NH1:1][CX4:2]!@;-[CX3:3]=[O:4]	2
234	[O:1][CX4:2]!@;-[CX3:3]=[O:4]	2
235	[N:1][CX4:2]!@;-[CX3:3]=[O:4]	2
236	[C:1][CX4H2:2]!@;-[CX3:3]=[O:4]	1

Continued on next page

237	[c:1][CX4H2:2]!@;-[CX3:3]=[O:4]	1
238	[!#1:1][CX4H2:2]!@;-[CX3:3]=[O:4]	2
239	[c:1][CX4:2]!@;-[CX3:3]=[O:4]	1
240	[C:1][CX4:2]!@;-[CX3:3]=[O:4]	1
241	[!#1:1][CX4:2]!@;-[CX3:3]=[O:4]	1
242	[c:1][CX4:2]!@;-[CX3:3][C:4]	1
243	[C:1][CX4:2]!@;-[CX3:3][c:4]	1
244	[c:1][CX4:2]!@;-[CX3:3][c:4]	2
245	[C:1][CX4:2]!@;-[CX3:3][C:4]	1
246	[!#1:1][CX4:2]!@;-[CX3H0:3][!#1:4]	1
247	[H:1][CX4H1:2]!@;-[CX3:3][!#1:4]	2
248	[!#1:1][CX4H2:2]!@;-[CX3:3][!#1:4]	1
249	[$([cH0]([$([NX3H2]),$([NX3H1])])):1][a:2]!@;-...	2
250	[nH0:1][c&r6:2]([nH0])!@;-[c&r6:3]([nH0])[nH0:4]	2
251	[nH0&r6:1][c&r6:2]([nH0&r6])!@;-[c&r6:3]([cH1&...	2
252	[nH0&r6:1][c&r6:2]([nH0&r6])!@;-[c&r6:3]([cH1&...	2
253	[c:1][c:2]!@;-[c:3][$(c!@c):4]	2
254	[cH0:1][c:2]([cH0])!@;-[c:3]([cH0:4])[cH0]	2
255	[cH0:1][c:2]([cH0])!@;-[c:3]([cH0:4])[cH1]	2
256	[cH0:1][c:2]([cH1])!@;-[c:3]([cH0:4])[cH1]	2
257	[cH0:1][c:2]([cH0])!@;-[c:3]([cH1:4])[cH1]	2
258	[cH0:1][c:2]([cH1])!@;-[c:3]([cH1:4])[cH1]	1
259	[cH1:1][c:2]([cH1])!@;-[c:3]([cH1:4])[cH1]	2
260	[nX2H0:1][c:2]!@;-[c:3][nX3H1:4]	2
261	[nX2H0:1][c:2]([!nX2H0])!@;-[c:3]([!nX2H0])[nX...	1
262	[nX2H0:1][$(c([nX2H0])(a(a)(a))!@;-c[nX2H0]):2...	1
263	[nX2H0:1][$([c&r6](c[OH])):2]!@;-[$([c&r6](c[O...	1
264	[c:1][c:2]!@;-[c:3][s,o,nX3H1:4]	1
265	[cH0:1][c:2]([cH0])!@;-[c:3][nX2H0:4]	2
266	[cH0:1][c:2]!@;-[c:3]([cH0])[nX2H0:4]	2
267	[c:1][c:2]!@;-[c:3]([cH0])[nX2H0:4]	1
268	[c:1][c:2]([cH0])!@;-[c:3][nX2H0:4]	2
269	[cH1:1][c:2]!@;-[c:3]([cH1])[nX2H0:4]	2
270	[c:1][c&r5:2]!@;-[c&r5:3][c:4]	2
271	[c:1][c&r6:2]!@;-[c&r5:3][c:4]	1
272	[c:1][c&r6:2]!@;-[c&r6:3][cH0:4]	2
273	[c:1][c&r6:2]!@;-[c&r6:3][c:4]	1
274	[c:1][c:2]!@;-[c:3][c:4]	1
275	[nX2&r6:1][cH0&r6:2]([cH1&r6])!@;-[CX4H2:3][O!...	1
276	[cH0:1][c:2]!@;-[CX4H0:3][a:4]	1
277	[cH0:1][c:2]!@;-[CX4H0:3][N,O,S:4]	1
278	[cH0:1][c:2]!@;-[CX4H0:3][CX3:4]	1

Continued on next page

279	[cH0:1][c:2]!@;-[CX4H0:3][CX4:4]	1
280	[cH0:1][c:2]!@;-[CX4H0:3][*:4]	1
281	[cH0:1][c:2]!@;-[CX4H2:3][a:4]	2
282	[cH0:1][c:2]!@;-[CX4H2:3][CX3:4]	2
283	[cH1:1][c:2]([cH1])!@;-[CX4H2:3][$([CX4H1]C(=O...	2
284	[cH1:1][c:2]([cH1])!@;-[CX4H2:3][CX4:4]	2
285	[cH0:1][c:2]!@;-[CX4H2:3][CX4:4]	2
286	[cH0:1][c:2]!@;-[CX4H2:3][N,O,S:4]	2
287	[cH0:1][c:2]!@;-[CX4H2:3][!#1:4]	2
288	[cH0:1][c:2]!@;-[CX4H1:3][a:4]	2
289	[cH0:1][c:2]!@;-[CX4H1:3][CX4:4]	2
290	[cH0:1][c:2]!@;-[CX4H1:3][CX3:4]	1
291	[cH0:1][c:2]!@;-[CX4H1:3][N,O,S:4]	1
292	[cH0:1][c:2]!@;-[CX4H1:3][H:4]	2
293	[a:1][c:2]!@;-[CX4H0:3][a:4]	2
294	[a:1][c:2]!@;-[CX4H0:3][CX3:4]	1
295	[a:1][c:2]!@;-[CX4H0:3][CX4:4]	1
296	[a:1][c:2]!@;-[CX4H0:3][N,O:4]	2
297	[a:1][c:2]!@;-[CX4H2:3][a:4]	2
298	[a:1][c:2]!@;-[CX4H2:3][CX3:4]	2
299	[n,o,s:1][c:2]!@;-[CX4H2:3][CX4:4]	1
300	[a:1][c:2]!@;-[CX4H2:3][!#1:4]	2
301	[a:1][c:2]!@;-[CX4H1:3][N,O:4]	1
302	[a:1][c:2]!@;-[CX4H1:3][a:4]	2
303	[a:1][c:2]!@;-[CX4H1:3][H:4]	2
304	[nX2H0:1][c:2]!@;-[C:3](=[N:4])(-[NH1,NH2])	2
305	[a:1][c:2]!@;-[C:3](=[$([N][CX4]):4])([$(N[CX4...	2
306	[a:1][c:2]!@;-[C:3](=[$([NH0][CX4]):4])(-N)	2
307	[a:1][c:2]!@;-[C:3](=[$([N][!#1]):4])([$(N(C)[...	2
308	[a:1][c:2]!@;-[C:3](=[$([N][!#1]):4])([$(N(C)~...	2
309	[a:1][c:2]!@;-[C:3](=[N:4])(-N)	2
310	[O:1]=[C:2]([O-])!@;-[c:3][$(aC(=O)(O)):4]	2
311	[O:1]=[C:2]([O-])!@;-[c:3][nX3H1:4]	2
312	[O:1]=[C:2]([O-])!@;-[c:3][nX2H0:4]	2
313	[O:1]=[C:2]([O-])!@;-[c:3]([cH0])[cH0:4]	1
314	[O:1]=[C:2]([O-])!@;-[c:3]([cH1])[$([cH0][NH1,...	2
315	[O:1]=[C:2]([O-])!@;-[c:3]([cH1])[cH0:4]	2
316	[O:1]=[C:2]([O-])!@;-[c:3]([cH1])[cH1:4]	2
317	[O:1]=[C:2]([O-])!@;-[c:3][a:4]	2
318	[$([c]([NH1,NH2])):1][c:2]!@;-[CX3:3](=[O])=[O:4]	1
319	[$(a[OH1]):1][a:2]!@;-[CX3:3]([NX3H0,CX4H0,c])...	1
320	[$(a[NH1,NH2]):1][a:2]!@;-[CX3:3]([NX3H0,CX4H0...	1

Continued on next page

321	[cH0:1][c:2]([cH1])!@;-[CX3:3](c)=[O:4]	1
322	[cH1:1][c:2]([cH1])!@;-[CX3:3](c)=[O:4]	2
323	[a:1][a:2]!@;-[CX3:3](a)=[O:4]	1
324	[$([cH0](=O)):1][c:2]([cH1])!@;-[CX3:3]([NX3H1...	1
325	[nH0&r6:1][c&r6:2]([cH1&r6])!@;-[C:3]([NH1,NH2...	1
326	[s:1][c:2]!@;-[C:3]([NH1])=[O:4]	2
327	[$([cH0]Cl):1][c:2]([cH1])!@;-[CX3:3]([NX3H1])...	2
328	[$([cH0]F):1][c:2]([cH1])!@;-[CX3:3]([NX3H1])=...	1
329	[$([cH0][OH0]):1][c:2]([cH1])!@;-[C:3](=O)[NH1:4]	1
330	[$([cH0][OH1]):1][c:2]([cH1])!@;-[C:3](=O)[NH1:4]	2
331	[cH0:1][c:2]([cH1])!@;-[CX3:3]([NX3H1])=[O:4]	1
332	[cH0:1][c:2]([cH1])!@;-[CX3:3]([NX3H0])=[O:4]	2
333	[cH1:1][c:2]([cH1])!@;-[C:3]([NH1,NH2])=[O:4]	2
334	[a:1][c:2]!@;-[C:3]([NH0])=[O:4]	1
335	[a:1][c:2]!@;-[C:3]([NH1,NH2])=[O:4]	2
336	[s:1][c:2]([aX2,cH1])!@;-[CX3:3](O)=[O:4]	2
337	[s:1][c:2]([aX2,cH1])!@;-[CX3:3]=[O:4]	1
338	[$([cH0](F)):1][c:2]([cH1])!@;-[CX3:3]([O,N])=...	2
339	[$([cH0]F):1][c:2]([cH1])!@;-[CX3:3]=[O:4]	1
340	[$([cH0](Cl)):1][c:2]([cH1])!@;-[CX3:3]([CX3H]...	1
341	[$([cH0](Cl)):1][c:2]([cH1])!@;-[CX3H:3]=[O:4]	1
342	[$([cH0](Cl)):1][c:2]([cH1])!@;-[CX3:3](O)=[O:4]	2
343	[nX3H1:1][a:2]!@;-[CX3:3]=[O:4]	2
344	[nX2H0:1][c:2]([cH1])!@;-[CX3H1:3]=[O:4]	1
345	[nX2H0&r6:1][c&r6:2]([c&r6])!@;-[CX3:3]([!O])=...	1
346	[nX2H0:1][a:2]([nX2H0])!@;-[CX3:3]=[O:4]	2
347	[$([cH0]!@;-[*^2]):1][c:2]([cH1])!@[CX3:3]=[O:4]	2
348	[cH0:1][c:2]([cH1])!@;-[CX3:3]=[O:4]	2
349	[cH0:1][c:2]([cH0])!@;-[CX3:3]=[O:4]	1
350	[cH1:1][c:2]([cH1])!@;-[CX3:3]([CX3H0])=[O:4]	2
351	[cH1:1][c:2]([cH1])!@;-[CX3:3]=[O:4]	2
352	[a:1][a:2]!@;-[CX3:3]=[O:4]	2
353	[cH1:1][c:2]([nX2])!@;-[CX3:3]=[NX3:4]	2
354	[cH1:1][c:2]([nX3H1])!@;-[CX3:3]=[NX2:4]	1
355	[cH1:1][c:2]([nX2])!@;-[CX3:3]=[NX2:4]	2
356	[cH1:1][c:2]([$([cH0][OH1])])!@;-[CX3:3]=[NX2:4]	1
357	[cH1:1][c:2]([cH0])!@;-[CX3x0:3]=[NX2:4]	2
358	[cH1:1][c:2]([cH1])!@;-[CX3:3]=[NX2:4]	2
359	[cH0:1][c:2]([cH0])!@;-[CX3!r:3]=[NX2!r:4]	2
360	[a:1][c:2]!@;-[CX3:3]=[CX3H0:4]	2
361	[a:1][a:2]!@;-[CX3:3]=[CX3H2:4]	2
362	[a:1][a:2]!@;-[CX3:3]=[CX3H1:4]	2

Continued on next page

363	[*:1][SX2:2]!@;-[SX2:3][*:4]	2
364	sp2-sp2	2
365	sp2-sp3	3
366	sp3-sp3	3

Table A.2: SMARTS patterns and the atom numbers that define the torsion angles. The atom numbers (1-4) and (5-8) define the first and the second torsion angles respectively. The SMARTS pattern in bold is the pattern defined in (Cole et al., 2018)

Pattern Number	SMARTS	1	2	3	4	5	6	7	8
1	[a][a]!@;-[CX3](=[CX3])!@;-[a][a]	0	1	2	4	1	2	4	5
2	[a][a]!@;-[NX3H1]!@;-[CX3](=S)[!#1]	0	1	2	3	1	2	3	5
3	[a][c]!@;-[CX4H0]([CX3,N])!@;-[c][a]	0	1	2	4	1	2	4	5
4	[a][c]!@;-[CX4H1]([N,O,H])!@-[c][a]	0	1	2	4	1	2	4	5
5	[a][c]!@;-[CH2]!@;-[n][a]	0	1	2	3	1	2	3	4
6	[a][c]!@;-[CX4H2]!@;-[OX2][C]	0	1	2	3	1	2	3	4
7	[a][c]!@;-[CX4H1]!@;-[OX2][!#1]	0	1	2	3	1	2	3	4
8	[c][c]!@;-[CX4]!@;-[c][c]	0	1	2	3	1	2	3	4
9	[cH0][c]!@;-[CX4H2]!@;-[a][a]	0	1	2	3	1	2	3	4
10	[cH0][c]([cH0])!@;-[NX3]!@;-[a][a]	0	1	3	4	1	3	4	5
11	[cH1][c]([cH1])!@;-[NX3]([CX4])!@;-[a][a]	0	1	3	5	1	3	5	6
12	[!#1][c]!@;-[SX2]!@;-[c][aH1,aH0]	0	1	2	3	1	2	3	4
13	[!#1][NX3H0]!@;-[C](=O)!@;-[O][CH0]	0	1	2	4	1	2	4	5
14	[CX3,CX4][CX4H2]!@;-[C](=O)!@;-[O]~[C]	0	1	2	4	1	2	4	5
15	[N,O,NH1,OH1][CX4]!@;-[C](=O)!@;-[O]~[C]	0	1	2	4	1	2	4	5
16	**[C,c][NH]!@;-[C](=S)!@;-[NH][C,c]**	0	1	2	4	1	2	4	5
17	[aH1][c]([aH1])!@;-[$(S(=O)=O)]!@;-[NX3H0][*]	0	1	3	4	1	3	4	5
18	[O]=[C]!@;-[O]!@;-[CX4H0][!#1]	0	1	2	3	1	2	3	4
19	[O]=[C]!@;-[NX3H0](A)!@;-[a][cH0]	0	1	2	4	1	2	4	5

Table A.3: Molecules used in BOA comparison.

Target	Number of rotatable bonds
astex_1p62	2
astex_1jd0	3
omegacsd_PAPPHI	3
omegacsd_PEXFED	3
omegapdb_1uwc	3
omegapdb_1d3g	3
omegacsd_YINDII	3
omegacsd_GALMOV	3
omegacsd_KAVLIC	3
astex_1n2j	3
omegacsd_WESZOJ	3
omegacsd_DUXYOK	3
omegacsd_YAHBOY	3
omegapdb_2aw1	3
omegacsd_MINTSA	3
omegacsd_CPIPLA	3
astex_1q1g	3
omegacsd_FABWUA10	3
omegapdb_1syh	3
omegacsd_CEVTUS	3
omegacsd_JUDLUP	3
omegacsd_GEKWAU	3
omegapdb_2evc	3
omegapdb_2flb	3
omegacsd_CMSMOC	3
omegacsd_DENBUT	4
omegacsd_PIKDIW	4
omegacsd_SIHVUA	4
omegacsd_WIKJOP	4
omegacsd_MATSTA10	4
omegapdb_1ofd	4
omegapdb_1y6q	4
omegacsd_MTBPNP	4
omegacsd_FOJZUZ	4
omegacsd_GINKUJ	4
omegacsd_YOPRAW	4
omegacsd_BAJZOB10	4
omegapdb_1qy5	4
omegacsd_BUTDID	4

Continued on next page

Table A.3: Molecules used in BOA comparison.

Target	Number of rotatable bonds
omegacsd_GIHKAJ	4
omegacsd_VUGDEG	4
astex_1tow	4
omegacsd_CAGXEN	4
omegacsd_KEBVES	4
omegacsd_ABHYTZ	4
omegapdb_2g8n	4
omegapdb_1gz8	4
omegacsd_TAPBZO	4
omegapdb_1yc1	4
omegacsd_FOLMEY	4
omegapdb_1uf7	4
omegacsd_CELDEC	4
omegacsd_DENDAC	4
omegacsd_CODYUP10	4
astex_1tz8	4
astex_1opk	4
omegacsd_EHMPYX10	4
omegapdb_1g6c	4
omegacsd_YUKKUK	4
omegacsd_DIGSIV	4
omegapdb_2cbs	4
omegacsd_DAWRAU	4
omegacsd_EITDZL	4
omegacsd_DIZREJ	4
omegapdb_1s8j	4
omegacsd_LETBUH	4
omegacsd_CMPEPI	4
omegapdb_1o3l	4
omegacsd_VOPDIN	4
omegacsd_GEJJUA	4
omegapdb_1rf6	4
omegacsd_LIKMEX	5
omegacsd_DIZWAK	5
omegacsd_YICPIJ	5
omegacsd_DEBBER	5
omegacsd_CETZOQ	5
omegacsd_VEYRAS	5
omegapdb_1v2k	5

Continued on next page

Table A.3: Molecules used in BOA comparison.

Target	Number of rotatable bonds
omegacsd_NBPENC	5
omegacsd_WAVTIW	5
omegacsd_COKTAX	5
omegacsd_BTHYDX	5
omegacsd_WAMZAL	5
omegacsd_BHMPET	5
omegacsd_ACPIXZ	5
omegacsd_VURTOR	5
omegapdb_1uf8	5
omegacsd_CBZHYX	5
omegacsd_LACPAG	5
omegacsd_FOLYIO	5
omegacsd_CLPNXA	5
omegacsd_YAHFAO	5
omegacsd_VIBWIM	5
omegapdb_1w9u	5
omegacsd_KICRIX	5
omegacsd_HEXFEV	5
omegacsd_YUYJEH	5
omegacsd_SILTUC	6
omegacsd_TAPSOS	6
omegacsd_DAFVUB	6
omegapdb_1h1s	6
omegacsd_BENPRL	6
omegacsd_SISYIC	6
omegacsd_CMANPQ	6
omegacsd_FLPNTX10	6
omegapdb_2gss	6
omegacsd_PMEPEN	6
astex_1v48	6
omegacsd_FOYLIO	6
omegacsd_SURREC	6
omegapdb_2j34	6
omegacsd_CFBPBI	6
omegacsd_ACENHT	6
omegacsd_FAHXIV	6
omegacsd_PEPHEX	6
omegacsd_FBPAZD	6
astex_1mzc	6

Continued on next page

Table A.3: Molecules used in BOA comparison.

Target	Number of rotatable bonds
omegacsd_PMBSAN10	6
omegacsd_HALDOL	6
omegacsd_AOPCHY	6
omegacsd_FUJRUX	6
omegacsd_FANRER	6
omegapdb_6prc	6
omegacsd_BEKDIE	6
omegacsd_LEDWAS	6

Table A.4: Molecules that excluded from the analysis (MMFF94).

SMILES	Name
O(c1ccc(Nc2c3ccccc3[nH]c2c2ccccc2)cc1)C	OMEGACSD_PMPAIN
S(c1ccccc1)c1ncc(n1C)[C@H](O)c1ccccc1	OMEGACSD_VUSKID
N(C(C)C)(C(=O)c1ccc(cc1)C)c1c(sc(c1)c1ccccc1)C(=O)O	OMEGAPDB_1yvx
N1C(=O)/C(=C(\CCC(=O)OC)/c2ccccc2)/C=C1c1ccccc1	OMEGACSD_FAXPUP
O=C(CCC)OC[C@@H](OC(=O)CCC)CO[P@@](=O)([O-])O [C@@H]1 [C@H](O)[C@H](O)[C@@H] (OP(=O)([O-])[O-])[C@H](O)[C@H]1O	DB4_4MXP_A

Table A.5: Molecules that excluded from the analysis (GFN2).

SMILES	Name
COc1c(ccc(c1)Cc1c2cc(c(cc2ccn1)OC)OC)OC	OMEGACSD_MVERIQ
O=P([O-])([O-])O[P@](=O)([O-])OCCc1c(C)c(c(s1) [C@H](O)CO)Cc1cnc(C)nc1N	1U0_4KXY_A
O=C(N[O-])[C@@H](CCCC[NH2+]Cc1ccc(cc1)F) C[C@@H](OC)c1ccc(cc1)F	0LX_4DV8_A
c1cc(c(cc1C(c1ccc(c(c1)C)OC[C@@H](C(C)(C)C)O) (CC)CC)C)O[C@H](CCO)CO	YR4_3AUN_A

Table A.6: RMSD: Wilcoxon signed-rank test on each method pair. Case 1: the lowest energy conformations found by either BOA or uniform random search is used as reference conformation. Molecules with three or fewer rotatable bonds ($N_{\text{rotor}} : 1, 2, 3$) and molecules with five or more rotatable bonds ($N_{\text{rotor}} : 5, 6$) are grouped together respectively due to small sample size. Case 2: the lowest energy conformation found by Confab is used as reference conformation. Molecules with three or fewer rotatable bonds ($N_{\text{rotor}} : 1, 2, 3$) are grouped together due to small sample size. The p-values are rounded to 2 significant figures.

		N_{rotor}			
Method Pairs	Case	1,2,3	4	5	6
EI-Uniform	1	$\mathbf{3.7 \times 10^{-10}}$	$\mathbf{6.5 \times 10^{-5}}$	$\mathbf{1.0 \times 10^{-3}}$	
EI-Confab	1	**0.04**	**0.04**	0.32	
EI-LCB	1	**0.03**	0.67	1.0	
LCB-Confab	1	0.64	0.24	0.73	
LCB-Uniform	1	$\mathbf{2.6 \times 10^{-7}}$	$\mathbf{6.1 \times 10^{-6}}$	$\mathbf{5.5 \times 10^{-5}}$	
Confab-Uniform	1	$\mathbf{3.1 \times 10^{-9}}$	$\mathbf{8.7 \times 10^{-8}}$	$\mathbf{1.5 \times 10^{-4}}$	
EI-LCB	2	0.03	0.66	0.96	0.76
EI-Uniform	2	$\mathbf{3.7 \times 10^{-10}}$	$\mathbf{6.5 \times 10^{-5}}$	0.06	$\mathbf{3.1 \times 10^{-3}}$
LCB-Uniform	2	$\mathbf{2.6 \times 10^{-7}}$	$\mathbf{6.1 \times 10^{-6}}$	$\mathbf{3.1 \times 10^{-3}}$	$\mathbf{4.8 \times 10^{-3}}$

Table A.7: TFD: Wilcoxon signed-rank test on each method pair.Case 1: the lowest energy conformations found by either BOA or uniform random search is used as reference conformation. Molecules with three or fewer rotatable bonds ($N_{\text{rotor}} : 1, 2, 3$) and molecules with five or more rotatable bonds ($N_{\text{rotor}} : 5, 6$) are grouped together respectively due to small sample size. Case 2: the lowest energy conformation found by Confab is used as reference conformation. Molecules with three or fewer rotatable bonds ($N_{\text{rotor}} : 1, 2, 3$) are grouped together due to small sample size. The p-values are rounded to 2 significant figures.

		N_{rotor}			
Method Pairs	Case	1,2,3	4	5	6
EI-Uniform	1	$\mathbf{1.3 \times 10^{-8}}$	$\mathbf{2.1 \times 10^{-5}}$	$\mathbf{3.1 \times 10^{-4}}$	
EI-Confab	1	0.70	0.14	0.14	
EI-LCB	1	0.07	0.94	0.74	
LCB-Confab	1	0.14	0.16	0.26	
LCB-Uniform	1	$\mathbf{1.0 \times 10^{-6}}$	$\mathbf{2.1 \times 10^{-5}}$	4.9×10^{-4}	
Confab-Uniform	1	$\mathbf{9.9 \times 10^{-10}}$	$\mathbf{5.0 \times 10^{-6}}$	$\mathbf{9.0 \times 10^{-6}}$	
EI-LCB	2	0.07	0.94	0.61	0.91
EI-Uniform	2	$\mathbf{1.3 \times 10^{-8}}$	$\mathbf{2.1 \times 10^{-5}}$	**0.01**	**0.01**
LCB-Uniform	2	$\mathbf{1.0 \times 10^{-6}}$	$\mathbf{2.1 \times 10^{-5}}$	$\mathbf{5.1 \times 10^{-3}}$	**0.04**

Table A.8: Wilcoxon signed rank test versus number of rotatable bonds with MMFF94 as energy function. BOA-EI and BOKEI are the Bayesian optimization with standard expected improvement and knowledge-based expected improvement respectively. GA represents the Genetic algorithm. We tested whether BOKEI found lower energy conformations than the two other methods, GA and BOA-EI, in the search (*i.e.* one-sided test). R (3.4.4) (R Core Team, 2020) was used to perform the hypothesis test. As the number of molecules with more than 10 rotatable bonds were small, we grouped the molecules with 11-13 and 14-18 rotatable bonds together when performing statistical test. The test showed the energy difference (between BOKEI and BOA-EI) was statistically significant ($p < 0.01$) across all rotatable bonds. The energy difference (between BOKEI and GA) were statistically significant up to ten rotatable bonds.

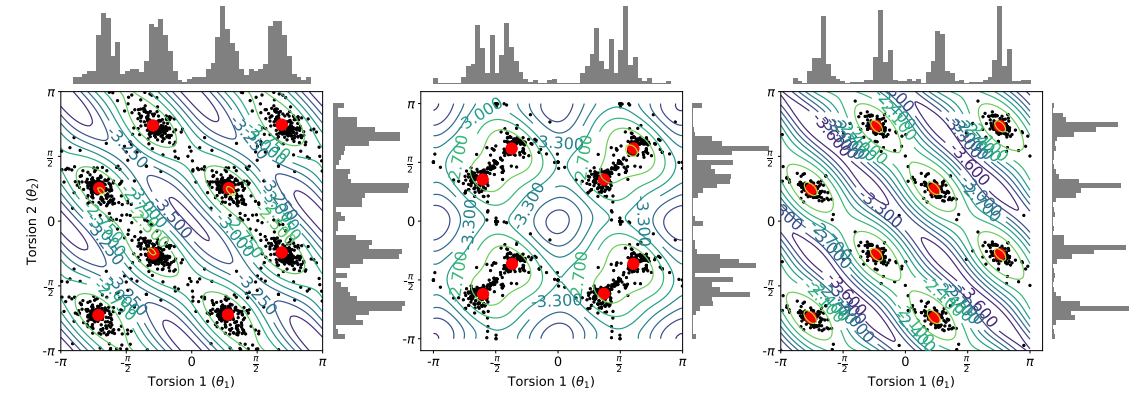
Number of rotatable bonds	BOKEI	BOA-EI	GA
2	NA	**3.82e-05**	**3.82e-06**
3	NA	**1.50e-08**	**1.23e-10**
4	NA	**3.04e-06**	**2.16e-13**
5	NA	**3.42e-11**	**3.54e-16**
6	NA	**9.84e-09**	**3.66e-13**
7	NA	**8.15e-06**	**1.05e-11**
8	NA	**1.14e-07**	**1.66e-05**
9	NA	**1.25e-06**	**1.75e-05**
10	NA	**2.21e-05**	**4.17e-07**
11-13	NA	**4.00e-03**	0.06
14-18	NA	**1.00e-03**	1

Table A.9: Wilcoxon signed rank test across number of rotatable bonds on all stochastic search algorithms with GFN2 as energy function. BOKEI and BOA-EI stand for Bayesian optimization with standard expected improvement and knowledge-based expected improvement respectively. We would like to test whether BOKEI found lower energy conformation than BOA-EI in the search (*i.e.* one-sided test). We performed the statistical test in R (3.4.4) (R Core Team, 2020). As the number of molecules with more than 10 rotatable bonds were small, we grouped the molecules with 11-13 together when performing statistical test. The test showed the energy difference (between BOKEI and BOA-EI) was statistically significant ($p < 0.01$) across all rotatable bonds.

Number of rotatable bonds	BOKEI	BOA-EI
2	NA	**7.90e-04**
3	NA	**1.41e-06**
4	NA	**3.41e-08**
5	NA	**9.31e-07**
6	NA	**1.21e-04**
7	NA	**2.02e-06**
8	NA	**9.84e-07**
9	NA	**1.34e-04**
10	NA	**2.34e-04**
11-13	NA	**6.14e-03**

Table A.10: Higher order correlated torsion SMARTS pattern and the atom numbers that define the torsion angles. Atoms (1-4), (5-8) and (9-12) defines the first, second and third torsion angles respectively.

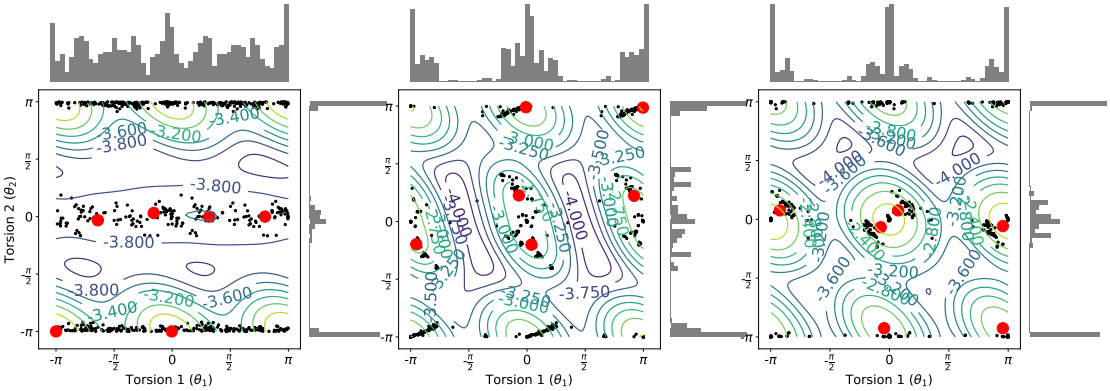
SMARTS	1	2	3	4	5	6	7	8	9	10	11	12
[#1][N](c(c)c)!@;-[C](=S)!@;-[NH1]!@;-[C](=O)	0	1	5	7	1	5	7	8	5	7	8	9



Cluster 1: ω: 0.125 μ: -2.24 ν: -2.28 κ_1: 1.5 κ_2: 1.5 κ_3: -1.32
Cluster 2: ω: 0.125 μ: -2.22 ν: 0.8 κ_1: 1.5 κ_2: 1.5 κ_3: -1.32
Cluster 3: ω: 0.125 μ: -0.92 ν: -0.79 κ_1: 1.5 κ_2: 1.49 κ_3: -1.32
Cluster 4: ω: 0.125 μ: -0.93 ν: 2.32 κ_1: 1.5 κ_2: 1.49 κ_3: -1.31
Cluster 5: ω: 0.125 μ: 0.87 ν: -2.27 κ_1: 1.5 κ_2: 1.5 κ_3: -1.31
Cluster 6: ω: 0.125 μ: 0.89 ν: 0.8 κ_1: 1.5 κ_2: 1.5 κ_3: -1.31
Cluster 7: ω: 0.125 μ: 2.16 ν: -0.76 κ_1: 1.5 κ_2: 1.49 κ_3: -1.31
Cluster 8: ω: 0.125 μ: 2.16 ν: 2.34 κ_1: 1.5 κ_2: 1.49 κ_3: -1.32

Cluster 1: ω: 0.115 μ: -1.89 ν: -1.95 κ_1: 1.52 κ_2: 1.52 κ_3: -1.27
Cluster 2: ω: 0.115 μ: -1.9 ν: 1.11 κ_1: 1.53 κ_2: 1.52 κ_3: -1.27
Cluster 3: ω: 0.135 μ: -1.16 ν: -1.14 κ_1: 1.51 κ_2: 1.51 κ_3: -1.29
Cluster 4: ω: 0.14 μ: -1.17 ν: 1.94 κ_1: 1.52 κ_2: 1.51 κ_3: -1.3
Cluster 5: ω: 0.11 μ: 1.16 ν: -1.94 κ_1: 1.52 κ_2: 1.53 κ_3: -1.26
Cluster 6: ω: 0.1 μ: 1.16 ν: 1.11 κ_1: 1.53 κ_2: 1.51 κ_3: -1.26
Cluster 7: ω: 0.14 μ: 1.9 ν: -1.14 κ_1: 1.52 κ_2: 1.51 κ_3: -1.3
Cluster 8: ω: 0.15 μ: 1.9 ν: 1.94 κ_1: 1.51 κ_2: 1.5 κ_3: -1.3

Cluster 1: ω: 0.125 μ: -2.39 ν: -2.34 κ_1: 1.88 κ_2: 1.89 κ_3: -1.81
Cluster 2: ω: 0.125 μ: -2.38 ν: 0.77 κ_1: 1.88 κ_2: 1.89 κ_3: -1.81
Cluster 3: ω: 0.125 μ: -0.74 ν: -0.8 κ_1: 1.88 κ_2: 1.88 κ_3: -1.81
Cluster 4: ω: 0.125 μ: -0.74 ν: 2.31 κ_1: 1.88 κ_2: 1.88 κ_3: -1.81
Cluster 5: ω: 0.125 μ: 0.73 ν: -2.34 κ_1: 1.88 κ_2: 1.89 κ_3: -1.81
Cluster 6: ω: 0.125 μ: 0.73 ν: 0.78 κ_1: 1.88 κ_2: 1.88 κ_3: -1.81
Cluster 7: ω: 0.125 μ: 2.38 ν: -0.8 κ_1: 1.88 κ_2: 1.88 κ_3: -1.81
Cluster 8: ω: 0.125 μ: 2.38 ν: 2.31 κ_1: 1.88 κ_2: 1.88 κ_3: -1.81

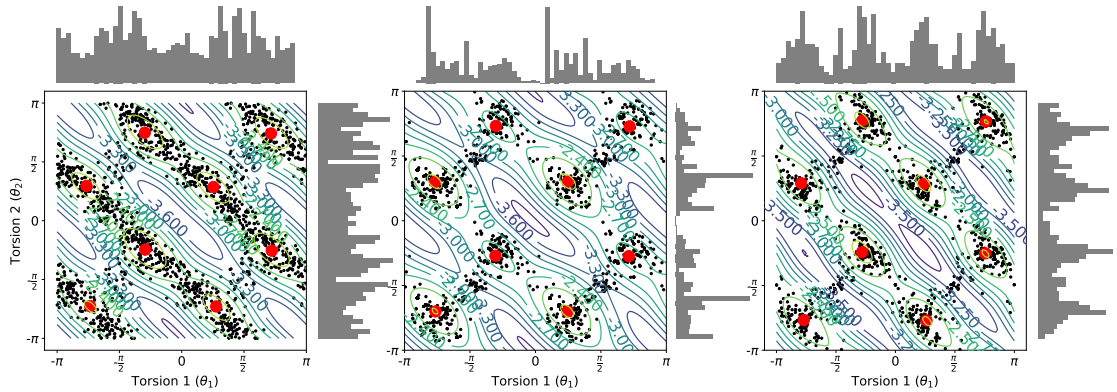


Cluster 1: ω: 0.3 μ: -3.14 ν: -3.14 κ_1: 0.76 κ_2: 1.3 κ_3: -0.6
Cluster 2: ω: 0.3 μ: -0.01 ν: -3.14 κ_1: 0.75 κ_2: 1.3 κ_3: -0.57
Cluster 3: ω: 0.1 μ: -2.01 ν: -0.1 κ_1: 0.92 κ_2: 0.9 κ_3: -0.26
Cluster 4: ω: 0.1 μ: -0.5 ν: 0.1 κ_1: 0.92 κ_2: 0.9 κ_3: -0.26
Cluster 5: ω: 0.1 μ: 1.0 ν: 0 κ_1: 0.93 κ_2: 0.9 κ_3: -0.26
Cluster 6: ω: 0.1 μ: 2.51 ν: 0 κ_1: 0.92 κ_2: 0.9 κ_3: -0.26

Cluster 1: ω: 0.28 μ: 3.13 ν: 3.1 κ_1: 1.14 κ_2: 1.36 κ_3: -0.82
Cluster 2: ω: 0.28 μ: -0.03 ν: 3.11 κ_1: 1.13 κ_2: 1.35 κ_3: -0.79
Cluster 3: ω: 0.11 μ: 2.89 ν: 0.7 κ_1: 1.5 κ_2: 1.38 κ_3: -1.25
Cluster 4: ω: 0.11 μ: -2.99 ν: -0.62 κ_1: 1.52 κ_2: 1.4 κ_3: -1.24
Cluster 5: ω: 0.11 μ: -0.22 ν: 0.71 κ_1: 1.4 κ_2: 1.28 κ_3: -1.18
Cluster 6: ω: 0.11 μ: 0.12 ν: -0.63 κ_1: 1.41 κ_2: 1.29 κ_3: -1.18

Cluster 1: ω: 0.09 μ: -2.9 ν: 0.24 κ_1: 1.96 κ_2: 1.96 κ_3: -1.3
Cluster 2: ω: 0.12 μ: -0.23 ν: -0.21 κ_1: 1.96 κ_2: 1.95 κ_3: -1.35
Cluster 3: ω: 0.11 μ: 0.23 ν: 0.24 κ_1: 1.96 κ_2: 1.96 κ_3: -1.34
Cluster 4: ω: 0.14 μ: 3.0 ν: -0.18 κ_1: 1.96 κ_2: 1.96 κ_3: -1.4
Cluster 5: ω: 0.27 μ: -0.14 ν: -2.91 κ_1: 1.47 κ_2: 1.48 κ_3: -1.3
Cluster 6: ω: 0.27 μ: 3 ν: -2.91 κ_1: 1.44 κ_2: 1.45 κ_3: -1.3

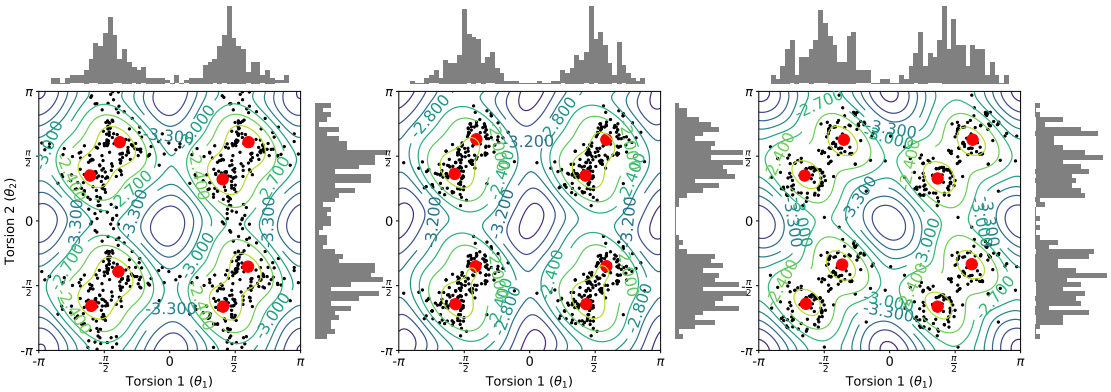
Figure A.1: Mixture models for correlated torsion. The contour plot indicates the log density of the mixture model and the points (in red) mark the mean location for the components.



Cluster 1: ω: 0.11 μ: -2.31 ν: -2.28 κ_1: 1.63 κ_2: 1.63 κ_3: -1.51
Cluster 2: ω: 0.12 μ: -2.4 ν: 0.93 κ_1: 1.62 κ_2: 1.63 κ_3: -1.52
Cluster 3: ω: 0.14 μ: -0.92 ν: -0.77 κ_1: 1.62 κ_2: 1.61 κ_3: -1.58
Cluster 4: ω: 0.13 μ: -0.93 ν: 2.35 κ_1: 1.63 κ_2: 1.62 κ_3: -1.53
Cluster 5: ω: 0.12 μ: 0.87 ν: -2.29 κ_1: 1.63 κ_2: 1.62 κ_3: -1.53
Cluster 6: ω: 0.125 μ: 0.8 ν: 0.9 κ_1: 1.62 κ_2: 1.63 κ_3: -1.54
Cluster 7: ω: 0.13 μ: 2.27 ν: -0.8 κ_1: 1.61 κ_2: 1.62 κ_3: -1.56
Cluster 8: ω: 0.125 μ: 2.25 ν: 2.33 κ_1: 1.63 κ_2: 1.63 κ_3: -1.51

Cluster 1: ω: 0.16 μ: -2.4 ν: -2.2 κ_1: 1.63 κ_2: 1.65 κ_3: -1.45
Cluster 2: ω: 0.16 μ: -2.41 ν: 0.94 κ_1: 1.65 κ_2: 1.66 κ_3: -1.46
Cluster 3: ω: 0.09 μ: -0.96 ν: -0.85 κ_1: 1.74 κ_2: 1.73 κ_3: -1.28
Cluster 4: ω: 0.09 μ: -0.95 ν: 2.3 κ_1: 1.74 κ_2: 1.73 κ_3: -1.28
Cluster 5: ω: 0.16 μ: 0.77 ν: -2.2 κ_1: 1.66 κ_2: 1.65 κ_3: -1.46
Cluster 6: ω: 0.16 μ: 0.78 ν: 0.95 κ_1: 1.65 κ_2: 1.67 κ_3: -1.47
Cluster 7: ω: 0.09 μ: 2.24 ν: -0.86 κ_1: 1.73 κ_2: 1.74 κ_3: -1.26
Cluster 8: ω: 0.09 μ: 2.26 ν: 2.29 κ_1: 1.73 κ_2: 1.74 κ_3: -1.27

Cluster 1: ω: 0.12 μ: -2.42 ν: -2.41 κ_1: 1.68 κ_2: 1.68 κ_3: -1.41
Cluster 2: ω: 0.12 μ: -2.49 ν: 0.91 κ_1: 1.68 κ_2: 1.67 κ_3: -1.42
Cluster 3: ω: 0.135 μ: -0.88 ν: -0.76 κ_1: 1.67 κ_2: 1.66 κ_3: -1.44
Cluster 4: ω: 0.125 μ: -0.87 ν: 2.43 κ_1: 1.69 κ_2: 1.67 κ_3: -1.43
Cluster 5: ω: 0.125 μ: 0.82 ν: -2.42 κ_1: 1.67 κ_2: 1.67 κ_3: -1.43
Cluster 6: ω: 0.13 μ: 0.75 ν: 0.893 κ_1: 1.66 κ_2: 1.67 κ_3: -1.43
Cluster 7: ω: 0.125 μ: 2.37 ν: -0.79 κ_1: 1.67 κ_2: 1.67 κ_3: -1.43
Cluster 8: ω: 0.12 μ: 2.39 ν: 2.41 κ_1: 1.69 κ_2: 1.68 κ_3: -1.41

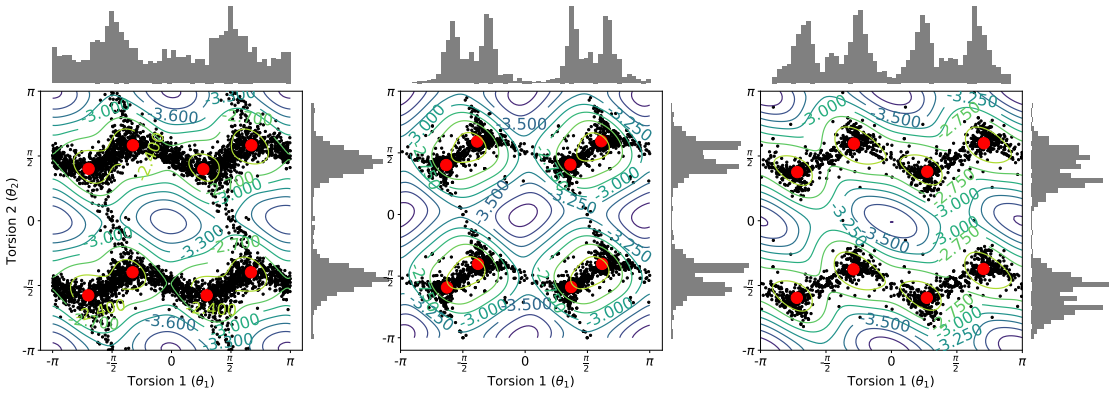


Cluster 1: ω: 0.125 μ: -1.88 ν: -2.06 κ_1: 1.66 κ_2: 1.64 κ_3: -1.43
Cluster 2: ω: 0.12 μ: -1.91 ν: 1.1 κ_1: 1.67 κ_2: 1.66 κ_3: -1.43
Cluster 3: ω: 0.13 μ: -1.2 ν: 1.91 κ_1: 1.67 κ_2: 1.66 κ_3: -1.44
Cluster 4: ω: 0.125 μ: -1.23 ν: -1.23 κ_1: 1.67 κ_2: 1.65 κ_3: -1.45
Cluster 5: ω: 0.125 μ: 1.28 ν: -2.08 κ_1: 1.68 κ_2: 1.65 κ_3: -1.44
Cluster 6: ω: 0.125 μ: 1.27 ν: 1.01 κ_1: 1.69 κ_2: 1.67 κ_3: -1.43
Cluster 7: ω: 0.12 μ: 1.88 ν: -1.11 κ_1: 1.69 κ_2: 1.68 κ_3: -1.44
Cluster 8: ω: 0.13 μ: 1.89 ν: 1.91 κ_1: 1.69 κ_2: 1.68 κ_3: -1.44

Cluster 1: ω: 0.14 μ: -1.78 ν: -2.02 κ_1: 1.86 κ_2: 1.85 κ_3: -1.71
Cluster 2: ω: 0.14 μ: -1.8 ν: 1.14 κ_1: 1.86 κ_2: 1.86 κ_3: -1.71
Cluster 3: ω: 0.11 μ: -1.28 ν: 1.98 κ_1: 1.89 κ_2: 1.87 κ_3: -1.68
Cluster 4: ω: 0.11 μ: -1.3 ν: -1.1 κ_1: 1.9 κ_2: 1.86 κ_3: -1.68
Cluster 5: ω: 0.14 μ: 1.36 ν: -2.02 κ_1: 1.87 κ_2: 1.86 κ_3: -1.72
Cluster 6: ω: 0.13 μ: 1.34 ν: 1.1 κ_1: 1.89 κ_2: 1.89 κ_3: -1.71
Cluster 7: ω: 0.11 μ: 1.84 ν: -1.1 κ_1: 1.9 κ_2: 1.88 κ_3: -1.68
Cluster 8: ω: 0.12 μ: 1.84 ν: 1.97 κ_1: 1.88 κ_2: 1.87 κ_3: -1.69

Cluster 1: ω: 0.125 μ: -2.01 ν: -2.01 κ_1: 1.67 κ_2: 1.67 κ_3: -1.42
Cluster 2: ω: 0.125 μ: -2.04 ν: 1.1 κ_1: 1.67 κ_2: 1.68 κ_3: -1.42
Cluster 3: ω: 0.125 μ: -1.1 ν: 1.97 κ_1: 1.69 κ_2: 1.68 κ_3: -1.43
Cluster 4: ω: 0.125 μ: -1.14 ν: -1.06 κ_1: 1.69 κ_2: 1.67 κ_3: -1.43
Cluster 5: ω: 0.125 μ: 1.15 ν: -2.08 κ_1: 1.68 κ_2: 1.67 κ_3: -1.42
Cluster 6: ω: 0.125 μ: 1.15 ν: 1.03 κ_1: 1.69 κ_2: 1.69 κ_3: -1.42
Cluster 7: ω: 0.125 μ: 1.97 ν: -1.05 κ_1: 1.69 κ_2: 1.68 κ_3: -1.41
Cluster 8: ω: 0.125 μ: 1.99 ν: 1.97 κ_1: 1.68 κ_2: 1.68 κ_3: -1.41

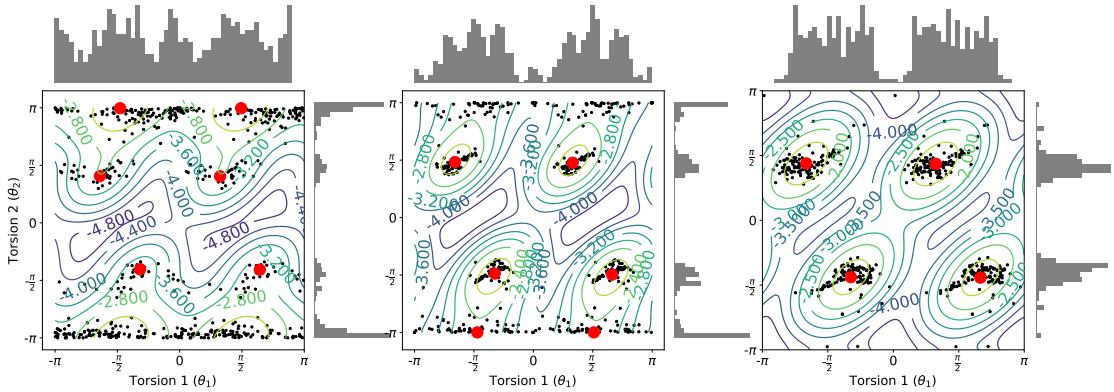
Figure A.1: Mixture models for correlated torsions. The contour plot indicates the log density of the mixture model and the points (in red) mark the mean location for the components. (Continued)



Cluster 1: ω: 0.13 μ: -2.2 ν: -1.81 κ_1: 1.66 κ_2: 1.64 κ_3: -1.11
Cluster 2: ω: 0.12 μ: -2.19 ν: 1.26 κ_1: 1.67 κ_2: 1.66 κ_3: -1.11
Cluster 3: ω: 0.12 μ: -1.03 ν: -1.25 κ_1: 1.67 κ_2: 1.66 κ_3: -1.07
Cluster 4: ω: 0.13 μ: -1.02 ν: 1.83 κ_1: 1.67 κ_2: 1.65 κ_3: -1.08
Cluster 5: ω: 0.125 μ: 0.93 ν: -1.81 κ_1: 1.68 κ_2: 1.65 κ_3: -1.08
Cluster 6: ω: 0.12 μ: 0.84 ν: 1.25 κ_1: 1.69 κ_2: 1.67 κ_3: -1.05
Cluster 7: ω: 0.125 μ: 2.1 ν: -1.25 κ_1: 1.69 κ_2: 1.68 κ_3: -1.06
Cluster 8: ω: 0.13 μ: 2.11 ν: 1.83 κ_1: 1.69 κ_2: 1.68 κ_3: -1.07

Cluster 1: ω: 0.12 μ: -1.97 ν: -1.86 κ_1: 1.44 κ_2: 1.56 κ_3: -1.19
Cluster 2: ω: 0.11 μ: -1.99 ν: 1.26 κ_1: 1.47 κ_2: 1.59 κ_3: -1.16
Cluster 3: ω: 0.13 μ: -1.19 ν: -1.26 κ_1: 1.46 κ_2: 1.56 κ_3: -1.21
Cluster 4: ω: 0.14 μ: -1.21 ν: 1.86 κ_1: 1.43 κ_2: 1.52 κ_3: -1.22
Cluster 5: ω: 0.12 μ: 1.16 ν: -1.85 κ_1: 1.45 κ_2: 1.56 κ_3: -1.18
Cluster 6: ω: 0.11 μ: 1.14 ν: 1.27 κ_1: 1.48 κ_2: 1.59 κ_3: -1.14
Cluster 7: ω: 0.13 μ: 1.94 ν: -1.26 κ_1: 1.46 κ_2: 1.55 κ_3: -1.21
Cluster 8: ω: 0.14 μ: 1.92 ν: 1.87 κ_1: 1.42 κ_2: 1.51 κ_3: -1.22

Cluster 1: ω: 0.12 μ: -2.27 ν: -1.87 κ_1: 1.42 κ_2: 1.57 κ_3: -1.2
Cluster 2: ω: 0.12 μ: -2.26 ν: 1.18 κ_1: 1.42 κ_2: 1.56 κ_3: -1.19
Cluster 3: ω: 0.13 μ: -0.91 ν: -1.18 κ_1: 1.42 κ_2: 1.54 κ_3: -1.21
Cluster 4: ω: 0.13 μ: -0.9 ν: 1.87 κ_1: 1.41 κ_2: 1.54 κ_3: -1.22
Cluster 5: ω: 0.12 μ: 0.85 ν: -1.87 κ_1: 1.41 κ_2: 1.56 κ_3: -1.19
Cluster 6: ω: 0.12 μ: 0.86 ν: 1.19 κ_1: 1.41 κ_2: 1.56 κ_3: -1.19
Cluster 7: ω: 0.13 μ: 2.21 ν: -1.17 κ_1: 1.41 κ_2: 1.54 κ_3: -1.21
Cluster 8: ω: 0.13 μ: 2.22 ν: 1.88 κ_1: 1.41 κ_2: 1.54 κ_3: -1.21

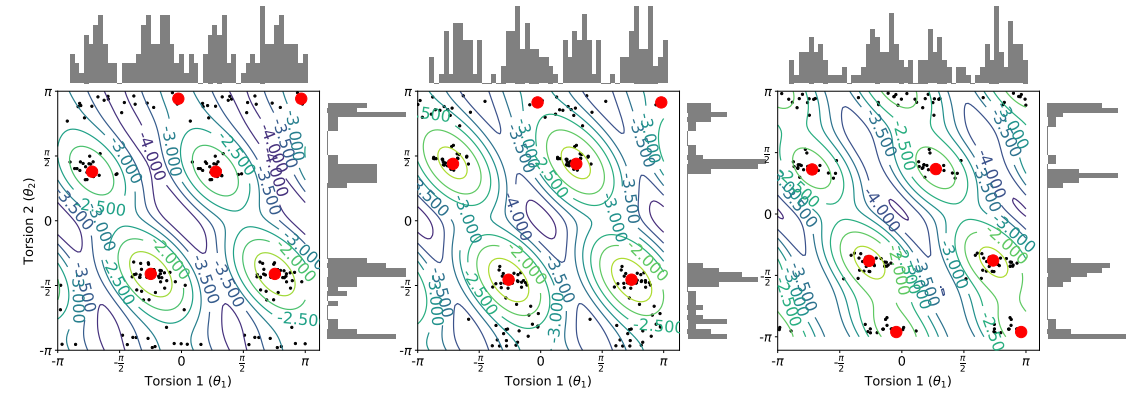


Cluster 1: ω: 0.06 μ: -2.01 ν: 1.28 κ_1: 1.78 κ_2: 1.77 κ_3: -1.08
Cluster 2: ω: 0.05 μ: -1 ν: -1.27 κ_1: 1.78 κ_2: 1.77 κ_3: -1.06
Cluster 3: ω: 0.06 μ: 1.02 ν: 1.27 κ_1: 1.78 κ_2: 1.77 κ_3: -1.08
Cluster 4: ω: 0.05 μ: 2.02 ν: -1.28 κ_1: 1.78 κ_2: 1.77 κ_3: -1.06
Cluster 5: ω: 0.39 μ: -1.5 ν: 3.13 κ_1: 1.02 κ_2: 1.71 κ_3: -0.49
Cluster 6: ω: 0.39 μ: 1.55 ν: 3.13 κ_1: 1.02 κ_2: 1.71 κ_3: -0.49

Cluster 1: ω: 0.145 μ: -2.07 ν: 1.52 κ_1: 1.74 κ_2: 1.7 κ_3: -1.37
Cluster 2: ω: 0.155 μ: -1.02 ν: -1.53 κ_1: 1.74 κ_2: 1.7 κ_3: -1.37
Cluster 3: ω: 0.145 μ: 1.04 ν: 1.5 κ_1: 1.74 κ_2: 1.71 κ_3: -1.37
Cluster 4: ω: 0.155 μ: 2.08 ν: -1.56 κ_1: 1.73 κ_2: 1.7 κ_3: -1.37
Cluster 5: ω: 0.2 μ: -1.48 ν: -3.14 κ_1: 1.0 κ_2: 1.45 κ_3: -0.74
Cluster 6: ω: 0.2 μ: 1.6 ν: -3.14 κ_1: 0.99 κ_2: 1.45 κ_3: -0.74

Cluster 1: ω: 0.265 μ: -2.09 ν: 1.38 κ_1: 1.67 κ_2: 1.71 κ_3: -1.58
Cluster 2: ω: 0.235 μ: -1.02 ν: -1.38 κ_1: 1.7 κ_2: 1.73 κ_3: -1.56
Cluster 3: ω: 0.265 μ: 1.02 ν: 1.37 κ_1: 1.68 κ_2: 1.71 κ_3: -1.59
Cluster 4: ω: 0.235 μ: 2.09 ν: -1.39 κ_1: 1.7 κ_2: 1.73 κ_3: -1.58

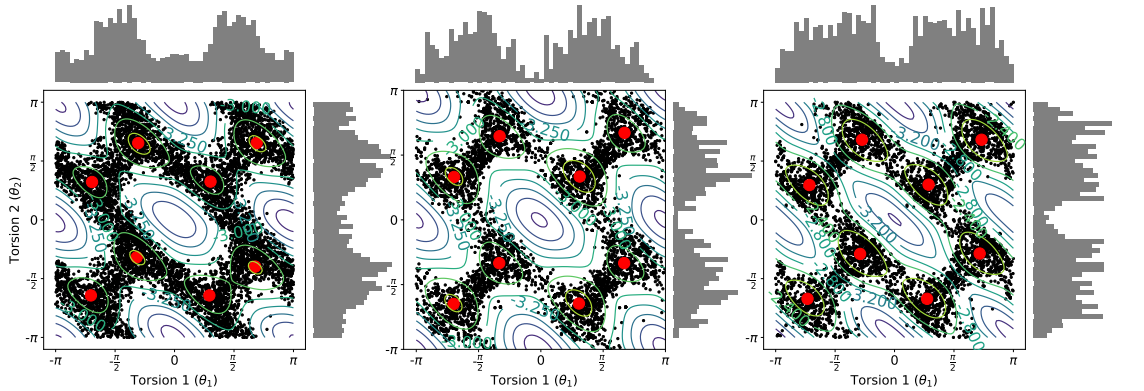
Figure A.1: Mixture models for correlated torsions. The contour plot indicates the log density of the mixture model and the points (in red) mark the mean location for the components.(Continued)



Cluster 1: ω: 0.135 μ: 3.04 ν: 2.96 κ_1: 1.05 κ_2: 1.3 κ_3: -0.92
Cluster 2: ω: 0.135 μ: -0.08 ν: 2.96 κ_1: 1.04 κ_2: 1.28 κ_3: -0.92
Cluster 3: ω: 0.145 μ: -2.27 ν: 1.19 κ_1: 1.92 κ_2: 1.89 κ_3: -1.43
Cluster 4: ω: 0.22 μ: -0.78 ν: -1.29 κ_1: 1.87 κ_2: 1.84 κ_3: -1.61
Cluster 5: ω: 0.15 μ: 0.87 ν: 1.19 κ_1: 1.92 κ_2: 1.89 κ_3: -1.43
Cluster 6: ω: 0.215 μ: 2.37 ν: -1.29 κ_1: 1.86 κ_2: 1.84 κ_3: -1.6

Cluster 1: ω: 0.1 μ: 3.08 ν: 2.87 κ_1: 1.11 κ_2: 1.27 κ_3: -1.04
Cluster 2: ω: 0.1 μ: -0.08 ν: 2.87 κ_1: 1.11 κ_2: 1.27 κ_3: -1.04
Cluster 3: ω: 0.18 μ: -2.24 ν: 1.38 κ_1: 1.91 κ_2: 1.84 κ_3: -1.56
Cluster 4: ω: 0.22 μ: -0.82 ν: -1.43 κ_1: 1.87 κ_2: 1.79 κ_3: -1.61
Cluster 5: ω: 0.18 μ: 0.91 ν: 1.38 κ_1: 1.91 κ_2: 1.84 κ_3: -1.56
Cluster 6: ω: 0.22 μ: 2.32 ν: -1.43 κ_1: 1.87 κ_2: 1.79 κ_3: -1.61

Cluster 1: ω: 0.21 μ: 3.02 ν: -3.02 κ_1: 1.48 κ_2: 1.55 κ_3: -1.43
Cluster 2: ω: 0.21 μ: -0.14 ν: -3.02 κ_1: 1.47 κ_2: 1.54 κ_3: -1.43
Cluster 3: ω: 0.125 μ: -2.27 ν: 1.14 κ_1: 1.91 κ_2: 1.9 κ_3: -1.5
Cluster 4: ω: 0.165 μ: -0.83 ν: -1.2 κ_1: 1.9 κ_2: 1.92 κ_3: -1.64
Cluster 5: ω: 0.125 μ: 0.86 ν: 1.15 κ_1: 1.92 κ_2: 1.92 κ_3: -1.5
Cluster 6: ω: 0.165 μ: 2.31 ν: -1.19 κ_1: 1.89 κ_2: 1.89 κ_3: -1.63



Cluster 1: ω: 0.12 μ: -2.21 ν: -2.02 κ_1: 1.39 κ_2: 1.39 κ_3: -1.08
Cluster 2: ω: 0.11 μ: -2.18 ν: 1.01 κ_1: 1.4 κ_2: 1.41 κ_3: -1.07
Cluster 3: ω: 0.13 μ: -1.0 ν: -1 κ_1: 1.41 κ_2: 1.4 κ_3: -1.12
Cluster 4: ω: 0.14 μ: -0.96 ν: 2.05 κ_1: 1.4 κ_2: 1.38 κ_3: -1.12
Cluster 5: ω: 0.12 μ: 0.92 ν: -2.02 κ_1: 1.39 κ_2: 1.4 κ_3: -1.07
Cluster 6: ω: 0.11 μ: 0.95 ν: 1.02 κ_1: 1.41 κ_2: 1.41 κ_3: -1.07
Cluster 7: ω: 0.13 μ: 2.13 ν: -1.25 κ_1: 1.41 κ_2: 1.39 κ_3: -1.11
Cluster 8: ω: 0.14 μ: 2.17 ν: 2.05 κ_1: 1.38 κ_2: 1.37 κ_3: -1.11

Cluster 1: ω: 0.135 μ: -2.2 ν: -2.04 κ_1: 1.39 κ_2: 1.41 κ_3: -1.11
Cluster 2: ω: 0.13 μ: -2.19 ν: 1.04 κ_1: 1.42 κ_2: 1.43 κ_3: -1.11
Cluster 3: ω: 0.115 μ: -1.05 ν: -1.05 κ_1: 1.44 κ_2: 1.44 κ_3: -1.07
Cluster 4: ω: 0.12 μ: -1.04 ν: 2.03 κ_1: 1.43 κ_2: 1.42 κ_3: -1.08
Cluster 5: ω: 0.14 μ: 0.96 ν: -2.03 κ_1: 1.4 κ_2: 1.41 κ_3: -1.12
Cluster 6: ω: 0.135 μ: 0.98 ν: 1.05 κ_1: 1.42 κ_2: 1.53 κ_3: -1.12
Cluster 7: ω: 0.11 μ: 2.1 ν: -1.05 κ_1: 1.44 κ_2: 1.43 κ_3: -1.06
Cluster 8: ω: 0.115 μ: 2.11 ν: 2.11 κ_1: 1.42 κ_2: 1.42 κ_3: -1.06

Cluster 1: ω: 0.12 μ: -2.3 ν: -2.11 κ_1: 1.84 κ_2: 1.85 κ_3: -1.54
Cluster 2: ω: 0.12 μ: -2.25 ν: 0.93 κ_1: 1.86 κ_2: 1.86 κ_3: -1.56
Cluster 3: ω: 0.13 μ: -0.91 ν: -0.91 κ_1: 1.86 κ_2: 1.86 κ_3: -1.6
Cluster 4: ω: 0.13 μ: -0.86 ν: 2.14 κ_1: 1.85 κ_2: 1.84 κ_3: -1.58
Cluster 5: ω: 0.125 μ: 0.86 ν: -2.11 κ_1: 1.85 κ_2: 1.85 κ_3: -1.56
Cluster 6: ω: 0.12 μ: 0.9 ν: 0.94 κ_1: 1.86 κ_2: 1.86 κ_3: -1.58
Cluster 7: ω: 0.125 μ: 2.25 ν: -0.9 κ_1: 1.85 κ_2: 1.85 κ_3: -1.58
Cluster 8: ω: 0.13 μ: 2.3 ν: 2.14 κ_1: 1.85 κ_2: 1.84 κ_3: -1.57

Figure A.1: Mixture models for correlated torsion. The contour plot indicates the log density of the mixture model and the points (in red) mark the mean location for the components. (Continued)

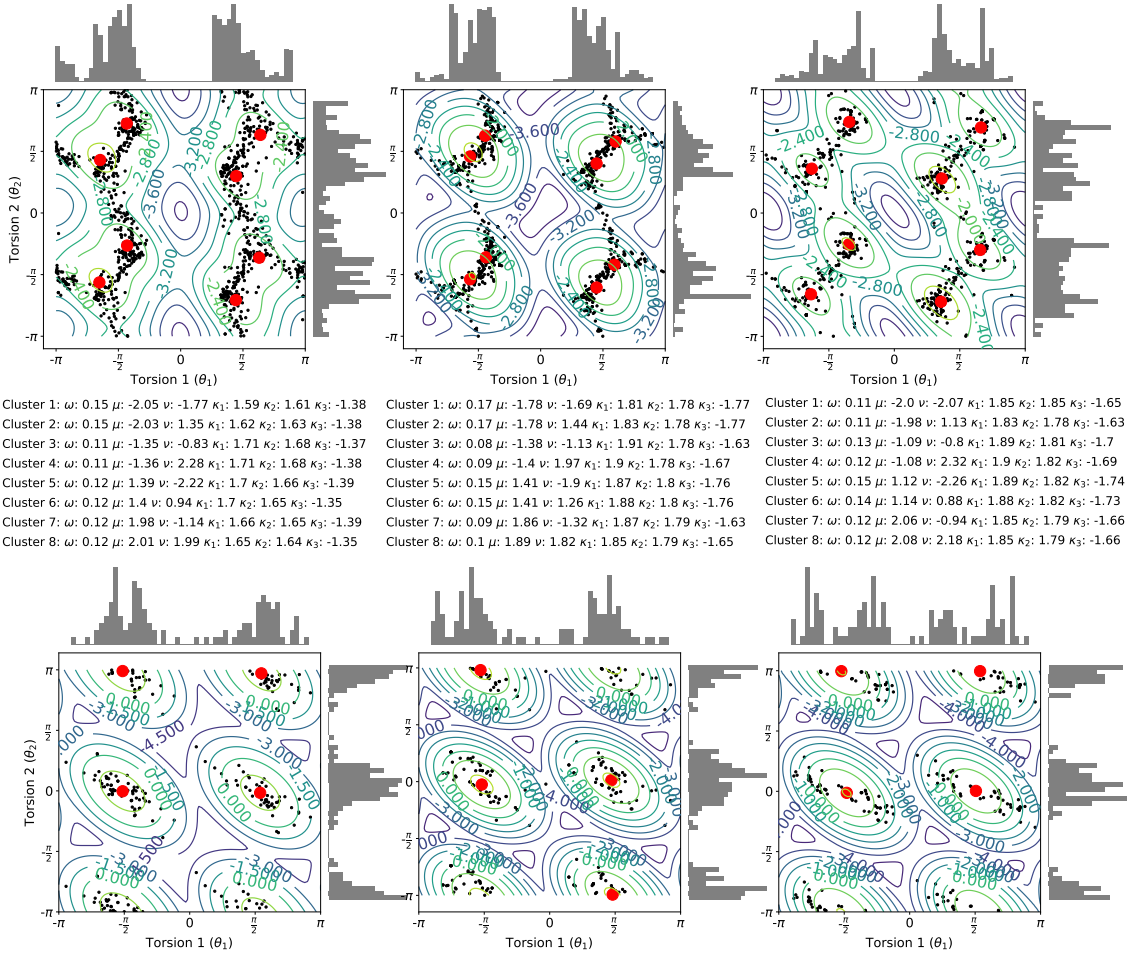
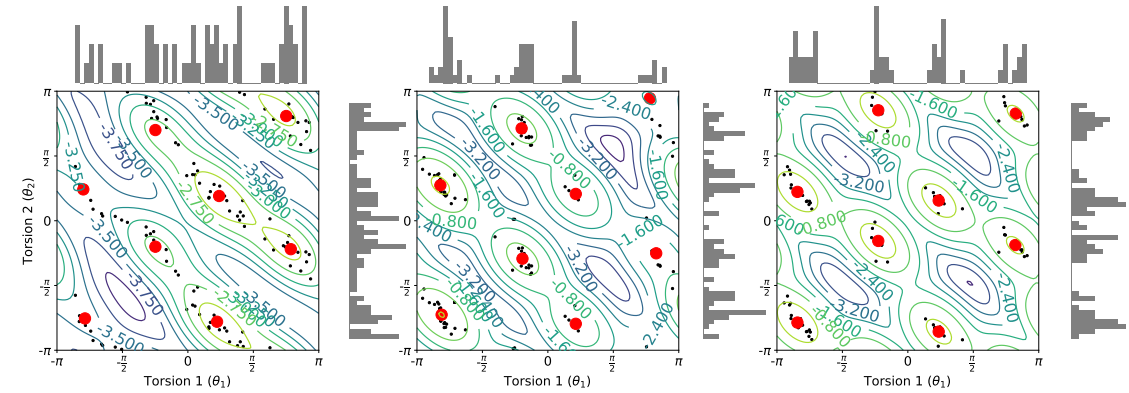


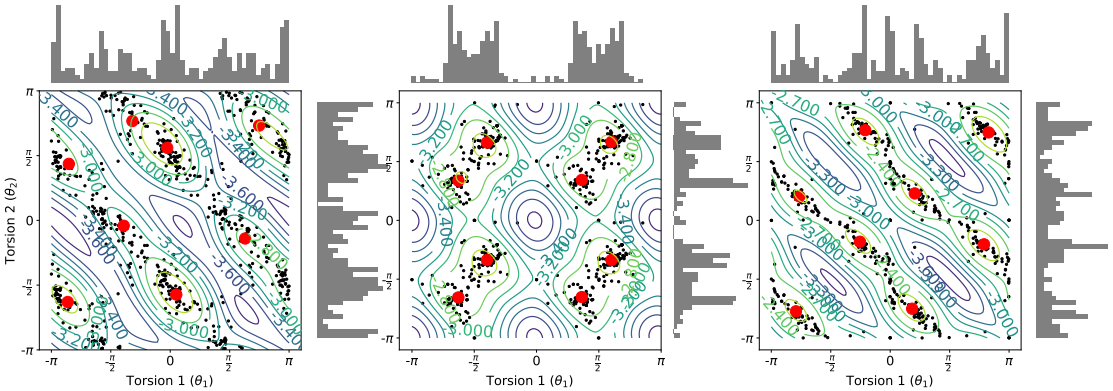
Figure A.1: Mixture models for correlated torsion. The contour plot indicates the log density of the mixture model and the points (in red) mark the mean location for the components. (Continued)



Cluster 1: ω: 0.05 μ: -2.47 ν: -2.37 κ_1: 1.42 κ_2: 1.4 κ_3: -0.8
Cluster 2: ω: 0.06 μ: -2.51 ν: 0.76 κ_1: 1.41 κ_2: 1.4 κ_3: -0.84
Cluster 3: ω: 0.12 μ: -0.78 ν: -0.63 κ_1: 1.34 κ_2: 1.34 κ_3: -1.02
Cluster 4: ω: 0.12 μ: -0.78 ν: 2.2 κ_1: 1.34 κ_2: 1.34 κ_3: -1.03
Cluster 5: ω: 0.17 μ: 0.7 ν: -2.45 κ_1: 1.27 κ_2: 1.28 κ_3: -1.14
Cluster 6: ω: 0.18 μ: 0.75 ν: 0.6 κ_1: 1.27 κ_2: 1.27 κ_3: -1.15
Cluster 7: ω: 0.15 μ: 2.47 ν: -0.69 κ_1: 1.33 κ_2: 1.33 κ_3: -1.1
Cluster 8: ω: 0.15 μ: 2.36 ν: 2.54 κ_1: 1.32 κ_2: 1.32 κ_3: -1.1

Cluster 1: ω: 0.21 μ: -2.55 ν: -2.28 κ_1: 2.96 κ_2: 2.97 κ_3: -2.56
Cluster 2: ω: 0.23 μ: -2.58 ν: 0.86 κ_1: 2.97 κ_2: 2.97 κ_3: -2.56
Cluster 3: ω: 0.14 μ: -0.61 ν: -0.92 κ_1: 2.98 κ_2: 2.97 κ_3: -2.42
Cluster 4: ω: 0.14 μ: -0.63 ν: 2.24 κ_1: 2.94 κ_2: 2.93 κ_3: -2.42
Cluster 5: ω: 0.1 μ: 0.67 ν: -2.5 κ_1: 2.98 κ_2: 2.97 κ_3: -2.26
Cluster 6: ω: 0.1 μ: 0.67 ν: 0.65 κ_1: 2.93 κ_2: 2.93 κ_3: -2.26
Cluster 7: ω: 0.05 μ: 2.6 ν: -0.79 κ_1: 2.91 κ_2: 2.91 κ_3: -2.1
Cluster 8: ω: 0.03 μ: 2.45 ν: 2.97 κ_1: 2.91 κ_2: 2.91 κ_3: -2.1

Cluster 1: ω: 0.16 μ: -2.65 ν: -2.47 κ_1: 2.82 κ_2: 2.8 κ_3: -2.72
Cluster 2: ω: 0.16 μ: -2.65 ν: 0.7 κ_1: 2.82 κ_2: 2.8 κ_3: -2.72
Cluster 3: ω: 0.12 μ: -0.71 ν: -0.49 κ_1: 2.92 κ_2: 2.92 κ_3: -2.81
Cluster 4: ω: 0.12 μ: -0.71 ν: 2.68 κ_1: 2.94 κ_2: 2.92 κ_3: -2.81
Cluster 5: ω: 0.12 μ: 0.74 ν: -2.69 κ_1: 2.82 κ_2: 2.8 κ_3: -2.72
Cluster 6: ω: 0.12 μ: 0.74 ν: 0.49 κ_1: 2.91 κ_2: 2.91 κ_3: -2.79
Cluster 7: ω: 0.1 μ: 2.58 ν: -0.59 κ_1: 2.91 κ_2: 2.91 κ_3: -2.67
Cluster 8: ω: 0.1 μ: 2.59 ν: 2.6 κ_1: 2.91 κ_2: 2.91 κ_3: -2.66



Cluster 1: ω: 0.16 μ: -2.71 ν: -1.98 κ_1: 1.23 κ_2: 1.23 κ_3: -1.09
Cluster 2: ω: 0.1 μ: -2.68 ν: 1.37 κ_1: 1.32 κ_2: 1.32 κ_3: -0.95
Cluster 3: ω: 0.09 μ: -1.23 ν: -0.13 κ_1: 1.35 κ_2: 1.35 κ_3: -0.9
Cluster 4: ω: 0.05 μ: -1 ν: 2.41 κ_1: 1.36 κ_2: 1.37 κ_3: -0.86
Cluster 5: ω: 0.18 μ: 0.16 ν: -1.8 κ_1: 1.22 κ_2: 1.23 κ_3: -1.1
Cluster 6: ω: 0.16 μ: -0.08 ν: 1.76 κ_1: 1.28 κ_2: 1.28 κ_3: -1.06
Cluster 7: ω: 0.12 μ: 1.98 ν: -0.44 κ_1: 1.27 κ_2: 1.26 κ_3: -1.06
Cluster 8: ω: 0.14 μ: 2.35 ν: 2.3 κ_1: 1.23 κ_2: 1.23 κ_3: -1.06

Cluster 1: ω: 0.12 μ: -1.97 ν: -2.06 κ_1: 1.34 κ_2: 1.31 κ_3: -1.08
Cluster 2: ω: 0.12 μ: -1.97 ν: 1.06 κ_1: 1.36 κ_2: 1.33 κ_3: -1.08
Cluster 3: ω: 0.13 μ: -1.25 ν: -1.06 κ_1: 1.36 κ_2: 1.33 κ_3: -1.08
Cluster 4: ω: 0.13 μ: -1.25 ν: 2.08 κ_1: 1.37 κ_2: 1.31 κ_3: -1.08
Cluster 5: ω: 0.12 μ: 1.14 ν: -2.05 κ_1: 1.34 κ_2: 1.32 κ_3: -1.07
Cluster 6: ω: 0.12 μ: 1.14 ν: 1.08 κ_1: 1.36 κ_2: 1.34 κ_3: -1.06
Cluster 7: ω: 0.13 μ: 1.89 ν: -1.06 κ_1: 1.35 κ_2: 1.32 κ_3: -1.08
Cluster 8: ω: 0.13 μ: 1.89 ν: 2.096 κ_1: 1.33 κ_2: 1.3 κ_3: -1.09

Cluster 1: ω: 0.125 μ: -2.49 ν: -2.43 κ_1: 1.63 κ_2: 1.64 κ_3: -1.44
Cluster 2: ω: 0.11 μ: -2.4 ν: 0.64 κ_1: 1.66 κ_2: 1.66 κ_3: -1.43
Cluster 3: ω: 0.12 μ: -0.8 ν: -0.57 κ_1: 1.64 κ_2: 1.66 κ_3: -1.46
Cluster 4: ω: 0.125 μ: -0.66 ν: 2.41 κ_1: 1.64 κ_2: 1.65 κ_3: -1.47
Cluster 5: ω: 0.13 μ: 0.58 ν: -2.36 κ_1: 1.64 κ_2: 1.63 κ_3: -1.48
Cluster 6: ω: 0.12 μ: 0.65 ν: 0.72 κ_1: 1.66 κ_2: 1.66 κ_3: -1.48
Cluster 7: ω: 0.13 μ: 2.48 ν: -0.64 κ_1: 1.65 κ_2: 1.65 κ_3: -1.48
Cluster 8: ω: 0.14 μ: 2.6 ν: 2.35 κ_1: 1.64 κ_2: 1.62 κ_3: -1.49

Figure A.1: Mixture models for correlated torsions. The contour plot indicates the log density of the mixture model and the points (in red) mark the mean location for the components.(Continued)

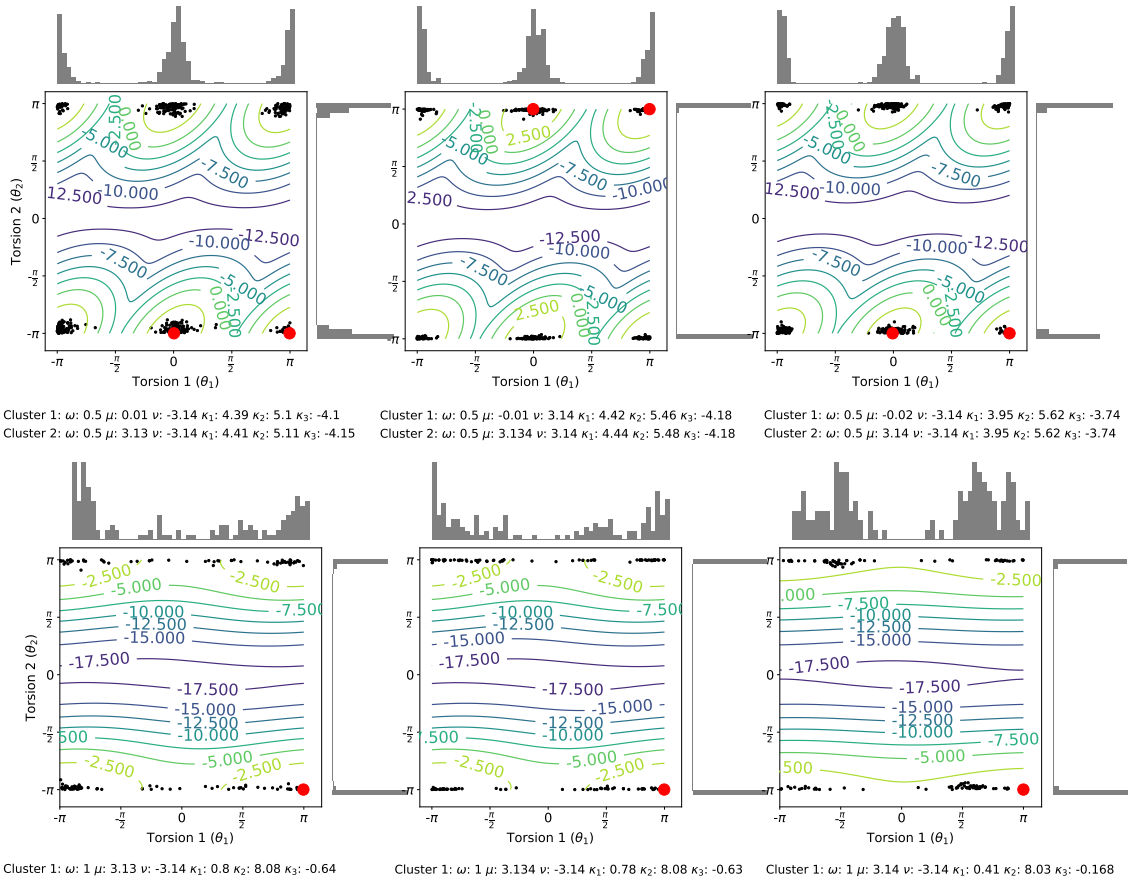


Figure A.1: Mixture models for correlated torsion. The contour plot indicates the log density of a mixture model and the points (in red) mark the mean location for the components.(Continued)

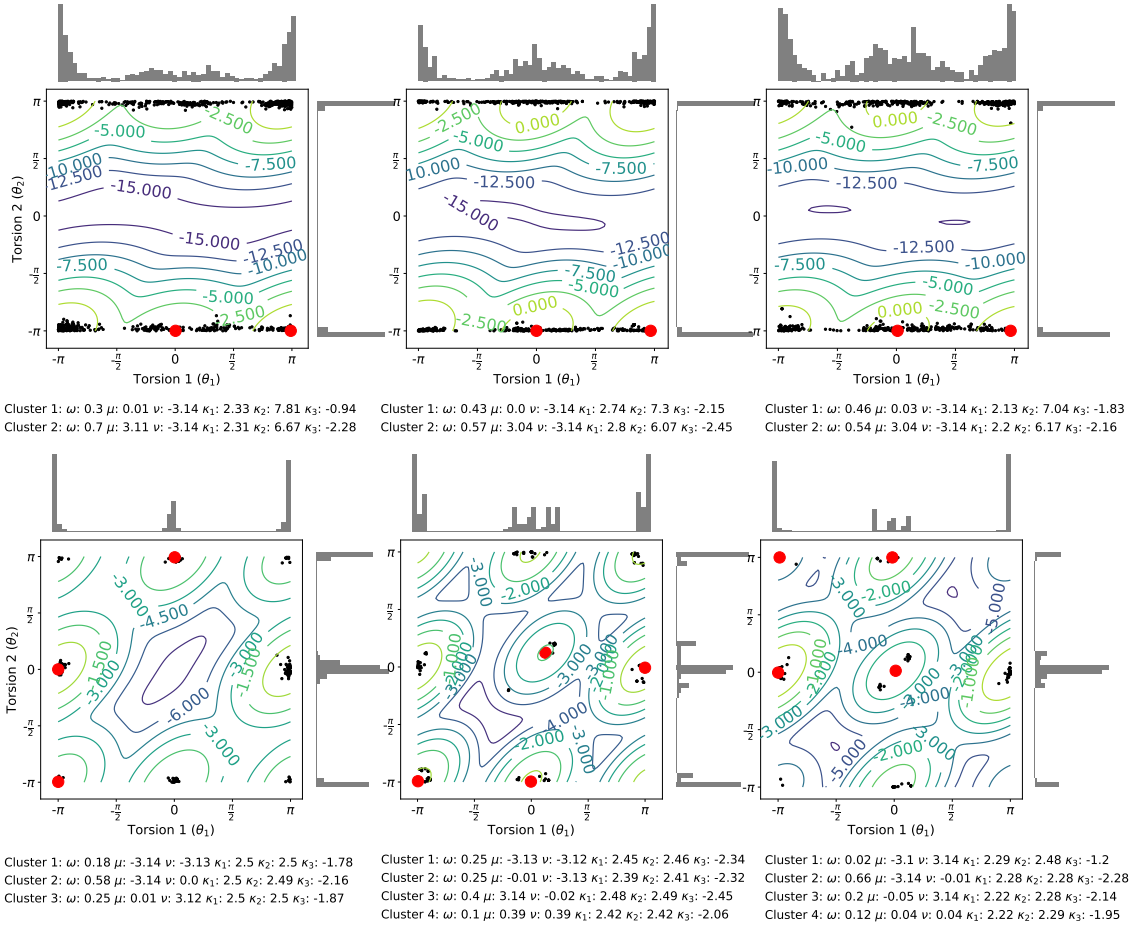
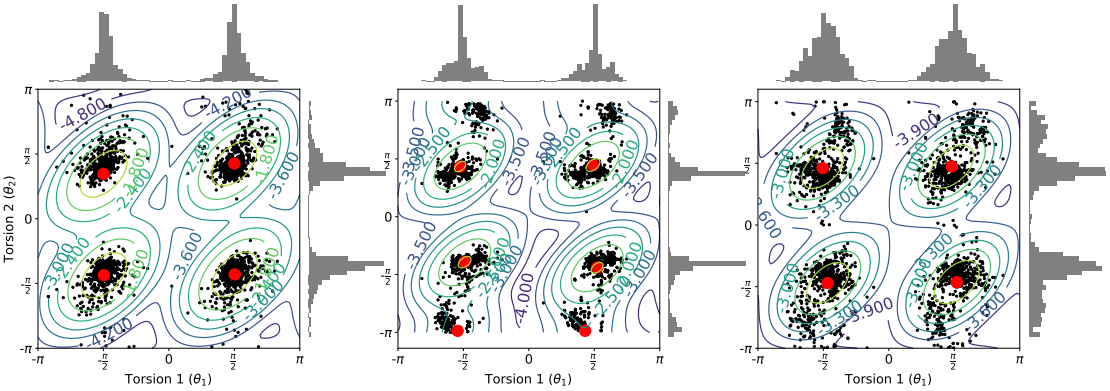


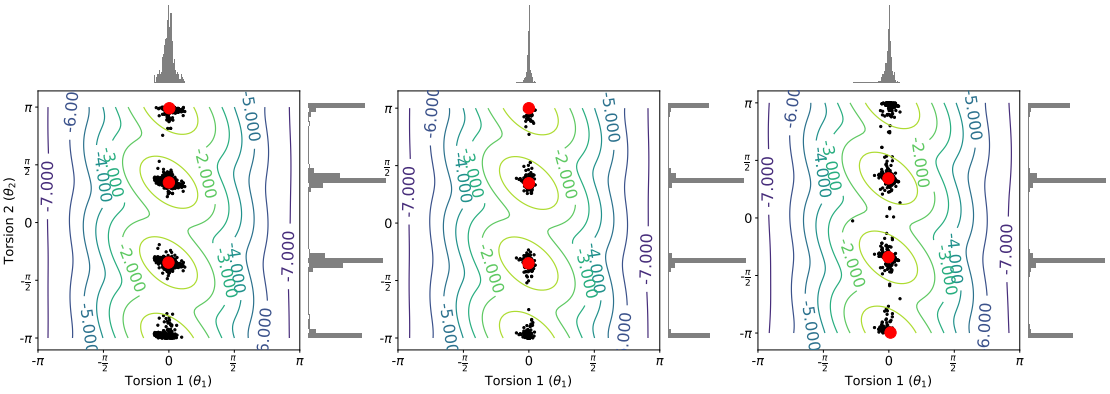
Figure A.1: Mixture models for correlated torsion. The contour plot indicates the log density of a mixture model and the points (in red) mark the mean location for the components.(Continued)



Cluster 1: ω: 0.25 μ: -1.56 ν: -1.37 κ_1: 1.91 κ_2: 1.89 κ_3: -1.83
Cluster 2: ω: 0.25 μ: -1.57 ν: 1.09 κ_1: 1.91 κ_2: 1.89 κ_3: -1.88
Cluster 3: ω: 0.25 μ: 1.58 ν: -1.35 κ_1: 1.9 κ_2: 1.88 κ_3: -1.84
Cluster 4: ω: 0.25 μ: 1.57 ν: 1.34 κ_1: 1.91 κ_2: 1.87 κ_3: -1.84

Cluster 1: ω: 0.08 μ: -1.71 ν: -3.1 κ_1: 1.91 κ_2: 1.91 κ_3: -1.23
Cluster 2: ω: 0.21 μ: -1.54 ν: -1.23 κ_1: 1.91 κ_2: 1.86 κ_3: -1.85
Cluster 3: ω: 0.21 μ: -1.65 ν: 1.36 κ_1: 1.9 κ_2: 1.85 κ_3: -1.86
Cluster 4: ω: 0.07 μ: 1.35 ν: -3.1 κ_1: 1.91 κ_2: 1.91 κ_3: -1.08
Cluster 5: ω: 0.21 μ: 1.64 ν: -1.38 κ_1: 1.92 κ_2: 1.86 κ_3: -1.83
Cluster 6: ω: 0.22 μ: 1.54 ν: 1.4 κ_1: 1.9 κ_2: 1.84 κ_3: -1.82

Cluster 1: ω: 0.25 μ: -1.47 ν: -1.48 κ_1: 1.25 κ_2: 1.13 κ_3: -1.14
Cluster 2: ω: 0.25 μ: -1.59 ν: 1.45 κ_1: 1.26 κ_2: 1.14 κ_3: -1.12
Cluster 3: ω: 0.25 μ: 1.63 ν: -1.46 κ_1: 1.28 κ_2: 1.14 κ_3: -1.15
Cluster 4: ω: 0.25 μ: 1.51 ν: 1.49 κ_1: 1.27 κ_2: 1.14 κ_3: -1.15

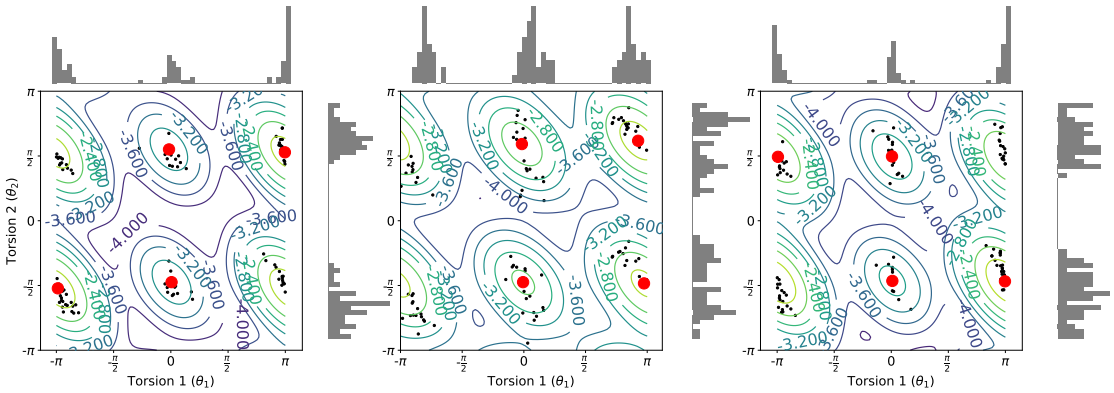


Cluster 1: ω: 0.333 μ: 0.01 ν: 3.11 κ_1: 1.98 κ_2: 1.97 κ_3: -1.95
Cluster 2: ω: 0.333 μ: 0 ν: 1.09 κ_1: 1.98 κ_2: 1.98 κ_3: -1.96
Cluster 3: ω: 0.334 μ: -0.01 ν: -1.09 κ_1: 1.98 κ_2: 1.98 κ_3: -1.96

Cluster 1: ω: 0.333 μ: 0 ν: 3.14 κ_1: 2 κ_2: 1.99 κ_3: -1.98
Cluster 2: ω: 0.333 μ: 0 ν: 1.09 κ_1: 2 κ_2: 1.99 κ_3: -1.98
Cluster 3: ω: 0.334 μ: -0.01 ν: -1.1 κ_1: 2 κ_2: 1.99 κ_3: -1.99

Cluster 1: ω: 0.33 μ: -0.01 ν: -1.07 κ_1: 2 κ_2: 1.98 κ_3: -1.98
Cluster 2: ω: 0.333 μ: -0.01 ν: 1.08 κ_1: 1.99 κ_2: 1.98 κ_3: -1.97
Cluster 3: ω: 0.334 μ: 0.04 ν: -3.13 κ_1: 1.99 κ_2: 1.98 κ_3: -1.97

Figure A.1: Mixture models for correlated torsions. The contour plot indicates the log density of the mixture model and the points (in red) mark the mean location for the components. (Continued)



Cluster 1: ω: 0.38 μ: -3.1 ν: -1.64 κ_1: 1.4 κ_2: 1.38 κ_3: -1.33
Cluster 2: ω: 0.34 μ: 3.13 ν: 1.66 κ_1: 1.44 κ_2: 1.42 κ_3: -1.31
Cluster 3: ω: 0.14 μ: 0.02 ν: -1.49 κ_1: 1.47 κ_2: 1.46 κ_3: -0.99
Cluster 4: ω: 0.14 μ: -0.05 ν: 1.73 κ_1: 1.47 κ_2: 1.45 κ_3: -0.97

Cluster 1: ω: 0.29 μ: 3.06 ν: -1.52 κ_1: 1.31 κ_2: 1.25 κ_3: -1.17
Cluster 2: ω: 0.31 μ: 2.91 ν: 1.94 κ_1: 1.32 κ_2: 1.27 κ_3: -1.19
Cluster 3: ω: 0.18 μ: -0.02 ν: -1.49 κ_1: 1.44 κ_2: 1.36 κ_3: -1.03
Cluster 4: ω: 0.22 μ: -0.05 ν: 1.86 κ_1: 1.44 κ_2: 1.3 κ_3: -1.05

Cluster 1: ω: 0.41 μ: 3.13 ν: -1.47 κ_1: 1.45 κ_2: 1.39 κ_3: -1.25
Cluster 2: ω: 0.31 μ: -3.12 ν: 1.55 κ_1: 1.47 κ_2: 1.44 κ_3: -1.16
Cluster 3: ω: 0.13 μ: 0.03 ν: -1.46 κ_1: 1.48 κ_2: 1.47 κ_3: -0.89
Cluster 4: ω: 0.15 μ: 0.02 ν: 1.56 κ_1: 1.48 κ_2: 1.46 κ_3: -0.92

Figure A.1: Mixture models for correlated torsions. The contour plot indicates the log density of the mixture model and the points (in red) mark the mean location for the components. (Continued)

Appendix B

Table B.1: Reference Bond Lengths. Aromatic atoms are represented by lower case.

Bond	Ring Size, N	Bond Length (Å)
C—C	5	1.535
C=C	5	1.339
c—c	5	1.403
C—C	6	1.531
C=C	6	1.336
c—c	6	1.388
C—C	7	1.530
C=C	7	1.339
c—c	7	1.390

Table B.2: Reference Bond Angles. Aromatic atoms are represented by lower case.

Bond Angle	Ring Size, N	Bond Angle (rad)
C—C—C	5	1.816
C=C—C	5	1.904
c—c—C	5	1.859
C—C—C	6	1.931
C=C—C	6	2.132
c—c—C	6	2.097
C—C—C	7	1.979
C=C—C	7	2.188
c—c—C	7	2.256

Table B.3: Partitioning of a Projected Planar Ring into Segments S_1, S_2 and S_3. The number of atoms, bonds, and angles in each of the three segments are given for N-membered rings of varying size.

	# Atoms in Segment			# Bonds in Segment			# Angles in Segment		
N	S_1	S_2	S_3	S_1	S_2	S_3	S_1	S_2	S_3
5	3	2	3	2	1	2	1	0	1
6	3	3	3	2	2	2	1	1	1
7	3	4	3	2	3	2	1	2	1
8	4	3	4	3	2	3	2	1	2
9	4	4	4	3	3	3	2	2	2
10	4	5	4	3	4	3	2	3	2
11	5	4	5	4	3	4	3	2	3
12	5	5	5	4	4	4	3	3	3
13	5	6	5	4	5	4	3	4	3
14	6	5	6	5	4	5	4	3	4
15	6	6	6	5	5	5	4	4	4
16	6	7	6	5	6	5	4	5	4

Table B.4: Model parameters for predicting the α substituent orientation angle of carbonyl functional group at a given position. Each model's performance is given in terms of circular correlation, R^2_{circ}, mean angular error (MAE), and standard deviation (s.d.) of angular error.

Substituent	Ring Size (Cluster)	Model Parameters (Pos, A, B, C, D, E, F, G)	Performance (R^2_{circ}, MAE, s.d.)
Carbonyl	5 (Envelope)	(1, 1.572, -1.185, -0.016, -0.001, -0.051) (2, 1.576, -1.217, -0.000, 0.045, -0.020) (3, 1.572, -1.249, 0.079, 0.013, 0.114)	(0.990, 0.026, 0.028) (0.986, 0.023, 0.028) (0.994, 0.022, 0.016)
Carbonyl	6 (Chair)	(1, 0.534, -1.343, 0.084, 3.302, -0.071, -0.006, -4.443), (2, 1.473, -1.340, -0.059, 0.995, 0.318, 0.005, 0.417)	(0.974, 0.032, 0.033), (0.958, 0.029, 0.032)
Carbonyl	6 (Chair)	(1, 0.880, -1.359, 0.014, -3.185, 0.129, 0.165, -1.568), (2, 1.603, -1.264, 0.075, 0.725, -0.002, 0.244, -0.694)	(0.974, 0.028, 0.026), (0.949, 0.034, 0.037)
Carbonyl	6 (Boat)	(1, 1.577, -1.104, 0.007, -1.270, 0.015, 0.015, -1.107) (2, 1.575, -1.104, 0.004, 1.427, 0.009, 0.018, -0.489)	(0.997, 0.035, 0.031), (0.997, 0.032, 0.035),
Carbonyl	12 (CCCC-DDDD, Sub-cluster 1)	(1, -1.840, -0.376, -0.708, -0.225, -1.071, 0.056, -0.324, -6.098, -0.156, -0.508, -0.061, 0.411, 1.803, 0.901, 0.043, 0.182, -1.001, -0.4201, 0.481) (4, 2.353, -0.379, -0.730, 2.795, -1.124, 0.061, -0.345, -0.959, -0.356, 0.451, 0.063, 0.422, 0.995, 3.512, 0.026, 0.160, 0.933, 0.436, 3.163) (7, 3.2801, -0.3783, -0.7161, 2.929, -1.1372, 0.0725, -0.3533, 0.4037, -0.2139, -0.3711, 0.0275, 0.3781, 0.6678, -0.3031, 0.0196, 0.1622, 1.2203, 1.7649, 1.1696) (10, 1.291, -0.347, -0.703, -1.845, -1.192, 0.025, -0.367, 0.538, -0.244, 0.551, 0.070, 0.289, -0.370, 1.066, 0.013, 0.288, -0.079, 1.466, 2.157)	(0.995, 0.012, 0.009) (0.994, 0.013, 0.010) (0.994, 0.013, 0.010) (0.994, 0.013, 0.010)

Table B.4: Model parameters for predicting the α substituent orientation angle of carbonyl functional group at a given position. Each model's performance is given in terms of circular correlation, R^2_{circ}, mean angular error (MAE), and standard deviation (s.d.) of angular error.

Substituent	Ring Size (Cluster)	Model Parameters (Pos, A, B, C, D, E, F, G)	Performance (R^2_{circ}, MAE, s.d.)
Carbonyl	12 (CCCC-DDDD, Sub-cluster 2)	(1, 0.519, -0.375, -0.586, -3.332, -0.861, 0.037, -0.113, -0.823, -0.028, -0.770, 0.017, -0.213, -0.267, -0.017, -0.057, 0.632, -1.086, 0.101, 1.713) (4,1.550, -0.376, -0.453, -0.953, -0.988, 0.038, -0.144, -0.422, -0.095, 0.701, 0.011, 0.187, 0.104, -0.339, -0.045, 0.619, -0.216, 2.095, -2.561) (7, 0.838, -0.380, -0.534, 0.988, -0.895, 0.044, -0.169, -3.195, -0.011, -0.725, 0.021, 1.069, 1.567, -0.493, -0.035, -1.400, -0.209, 4.376, -0.236) (10, 0.079, -0.377, -0.457, -2.470, -0.951, 0.044, -0.072, -2.296, 0.008, 0.801, 0.016, 0.797,0.442, -0.342, -0.049, -0.985, -1.147, 2.500, 4.496)	(0.999, 0.010, 0.010) (0.999, 0.011, 0.010) (0.999, 0.011, 0.010) (1.000, 0.010, 0.009)

Table B.5: Model parameters for predicting β orientation angle of carbonyl group at a given position. Each model's performance is given in terms of circular correlation, R^2_{circ}, mean angular error (MAE), and standard deviation (s.d.) of angular error.

Substituents	Ring Size (Cluster)	Model Parameters (Pos, A, B, C, D, E, F, G)	Performance (R^2_{circ}, MAE, s.d.)
Carbonyl	5 (Envelope)	(1, 0.001, -0.000, 0.002, 0.011, 0.114) (2, 0.001, -0.002, 0.002, 0.010, 0.118) (3, 0.012, -0.23, 0.025, 0.128, 0.037)	(0.410, 0.017, 0.15) (0.620, 0.014, 0.012) (0.337, 0.016, 0.012)
Carbonyl	6 (Chair)	(1, 0.471, 0.038, 0.240, -1.749, 0.064, 0.154, 1.614), (2, -0.278, -0.001, -0.175, 0.929, -0.001, 0.0743, -0.778)	(0.514, 0.021, 0.022), (0.570, 0.015, 0.018)
Carbonyl	6 (Chair)	(1, -0.490, -0.072, -0.276, -1.919, 0.175, 0.159, -1.844), (2, -0.405, 0.019, 0.245, -1.427, 0.018, 0.216, -1.243)	(0.514, 0.021, 0.022), (0.570, 0.0315, 0.018)
Carbonyl	6 (Boat)	(1, 0.002, -0.007, 0.008, -0.102, 0.015, 0.211, -1.237) (2, -0.003, 0.004, -0.003, 0.008, -0.010, 0.213, 0.429)	(0.929, 0.028, 0.025), (0.934, 0.027, 0.025),

Table B.5: Model parameters for predicting β orientation angle of carbonyl group at a given position. Each model's performance is given in terms of circular correlation, R^2_{circ}, mean angular error (MAE), and standard deviation (s.d.) of angular error.

Substituents	Ring Size (Cluster)	Model Parameters (Pos, A, B, C, D, E, F, G)	Performance (R^2_{circ}, MAE, s.d.)
Carbonyl	12 (CCCC-DDDD, Sub-cluster 1)	(1, 21.764, 0.302, -0.3618, 8.760, -0.883, -1.148, -0.176, 32.280, 0.492, -5.984, 0.267, -5.709, -7.149, -24.586, 0.0054, -2.873, 9.648, -20.494, 1.896) (4, -19.002, 0.341, -0.043, -4.879, -0.356, -1.190, 0.021, -30.778, 1.332, 6.009, 0.052, -5.867, 7.729, -21.298, 0.083 ,-2.792, -7.935, -20.719, -5.047) (7, 22.466, 0.405, -0.336, 4.221, 0.835, -1.300, 0.346, 37.008, 0.456, -6.296, 0.166, -6.015, -9.446, -24.509, 0.092, -2.601, 9.392 -20.931, 0.889) (10, -12.859, 0.337, -0.527, -1.170, -0.230, -1.192, 0.054, -22.646, -0.112, 6.009, 0.072, -5.897, 6.250, -21.932, 0.079, -2.597, -5.140, -19.561, -3.109)	(0.908, 0.067, 0.051), (0.914, 0.066, 0.048), (0.912, 0.065, 0.049), (0.909, 0.068, 0.048)
Carbonyl	12 (CCCC-DDDD, Sub-cluster 2)	(1, -5.850, -0.188, -1.256, -19.314, 0.870, -0.393, -0.101, 1.369, 1.258, 0.132, 0.155, -3.474, -4.628, 2.818, 0.028, 5.924, -3.474, -4.628, 19.583) (4, -9.498, -0.134, -0.689, -8.553, 0.414, -0.423, -0.245, -13.320, 0.997, 0.024, 0.127, 0.373, 2.656, -4.144, 0.110, -1.111, -5.818, 8.279, 17.277) (7, -3.751, -0.167,-0.916, -15.416, 0.366, -0.395, -0.127, 4.390, 1.110, 0.290, 0.215, 0.602, -4.754, 1.873, 0.166, -1.458, -3.511, -1.313, 20.958) (10, -7.111, -0.126, -0.985, -5.345, 0.519, -0.430, -0.055, -11.107, 1.507, -0.035, 0.141, -1.687, 3.053, -1.491, 0.081, 1.585, -4.188, -3.969, 24.305)	(0.994, 0.041, 0.044), (0.994, 0.043, 0.042), (0.993, 0.046, 0.046), (0.995, 0.040, 0.044)

Table B.6: Model parameters for predicting endocyclic torsion angles. Each model's performance is given in terms of circular correlation, R^2_{circ}, mean angular error (MAE), and standard deviation (s.d.) of angular error.

Ring Size	Cluster	Model Parameters	Performance (R^2_{circ}, MAE, s.d.)
5	Cluster 1 (Envelope)	(0, 0.000, -1.694, -0.543) , (1 , 0.000, 1.035, 1.433), (2, 0.000, 0.001, -1.803), (3, 0.000, -1.070, 1.483), (4, 0.000, 1.710, -0.575)	(0.999, 0.012, 0.018), (0.999, 0.016, 0.023), (0.999, 0.016, 0.023), (0.999, 0.027, 0.033), (0.999, 0.024, 0.029)
6	Cluster 1 (Chair)	(0, 0.141, -1.155, -0.642, -1.454) , (1, -0.116, -1.111, -0.628, 1.483), (2, 0.160, -1.112, -0.625, -1.412), (3, -0.249, -1.085, -0.632, 1.285), (4, 0.327, -1.029, -0.674, -1.174), (5, -0.268, -1.098, -0.735, 1.264)	(0.971, 0.019, 0.018), (0.964, 0.021, 0.018), (0.953, 0.021, 0.019), (0.938, 0.021, 0.024), (0.889, 0.036, 0.027), (0.917, 0.033, 0.027)
6	Cluster 2 (Chair)	(0, -0.160, -1.155, -0.646, -1.419) , (1, 0.114, -1.084, -0.626, 1.486), (2, -0.174, -1.120, -0.638, -1.385), (3, 0.251, -1.072, -0.617, 1.2806), (4, -0.314, -1.026, -0.659, -1.195), (5, 0.279, -1.067, -0.720, 1.240)	(0.974, 0.018, 0.016), (0.968, 0.021, 0.017), (0.958, 0.022, 0.019), (0.933, 0.022, 0.025), (0.902, 0.034, 0.028), (0.926, 0.031, 0.026)
6	Cluster 3 (Boat)	(0, -0.002, -1.251, -0.707, -1.585) (1, -0.003, -1.251, -0.715, 1.616) (2, 0.001, -1.24, -0.704, -1.658) (3, -0.003, -1.246, -0.741, 1.641) (4, -0.003, -1.299, -0.759, -1.541) (5, 0.003, -1.302, -0.702, 1.549)	(0.999, 0.025, 0.028), (0.999, 0.030, 0.022), (0.999, 0.030, 0.027), (0.999, 0.028, 0.034), (0.999, 0.041, 0.033), (0.998, 0.038, 0.037),

Table B.6: Model parameters for predicting endocyclic torsion angles. Each model's performance is given in terms of circular correlation, R^2_{circ}, mean angular error (MAE), and standard deviation of angular error.

Ring Size	Cluster	Model Parameters	Performance (R^2_{circ}, MAE, s.d.)
12	Sub-cluster 1 (CCCC-DDDD)	(1, 1.654 , 0.168, -1.333, -0.834, -0.457, -1.491, 0.803, 0.240, -0.310, -8.318),	(0.884, 0.124, 0.094),
		(2, -1.429, -1.161, 0.731, -0.595, -0.872, -0.211 , -1.001, -1.585, -1.998, 7.736),	(0.824, 0.126, 0.089),
		(4, 2.323, 0.127, -0.773, -0.975, -1.130, -1.390, 0.988, 0.886, 1.102, 8.203),	(0.876, 0.137, 0.096),
		(5, -2.195, -1.191, 0.978, -0.023, -0.437, -0.202, -1.543, -1.046, -0.402, -8.151),	(0.796, 0.132, 0.090),
		(7, 1.452, 0.258, -1.043, -0.790, 1.035, -1.636, 1.637, 0.031, 1.319, -8.982),	(0.875, 0.135, 0.099),
		(8, -1.401, -1.272, 1.519, -0.618, 1.064 , -0.183, -1.375, -1.642, -0.812, 8.332),	(0, 779, 0.129, 0.095),
		(10, 2.5168, 0.088, -1.469, -0.924, 1.127, -1.362, 1.438, 1.020, 0.411, 8.112),	(0.883, 0.133, 0.093),
		(11, -2.380, -1.100, 1.382, 0.099, 0.931, -0.195, -0.863, -1.102, -2.154, -7.874)	(0.774, 0.126, 0.090),
12	Sub-cluster 2 (CCCC-DDDD)	(0, 0.985, -0.187, -0.850, -1.433, -1.600, -0.564, -0.205, 0.050, 0.317, -0.656),	(0.997, 0.037, 0.042),
		(1, -1.384, -0.385, 0.214, -0.664, -0.888, -0.456, -1.228, -1.861, -1.2400, 0.881),	(0.993, 0.045, 0.050),
		(3, 1.106, -0.178, -1.112, -1.577, -1.837, -0.532 , -1.225 , 0.298, -1.126, 1.296),	(0.997, 0.047, 0.043),
		(4, -1.214, -0.416, -1.187, -0.738, -1.785, -0.470, -1.270, -1.461, -2.301, -2.131),	(0.989, 0.060, 0.057),
		(6, 0.971,-0.160, -0.911, -1.441, -1.631, -0.568, 0.030, 0.037, -0.552, -1.201),	(0.997, 0.042, 0.039),
		(7, -1.234, -0.384, 0.519, -0.796, -1.232, -0.443, -1.094, -1.940, -1.385, 0.822),	(0.992, 0.051, 0.051),
		(9, 1.095, -0.170 , -1.380, -1.731, -1.315, -0.568, -0.158, 0.401, 0.033, 0.538),	(0.997, 0.040, 0.041)
		(10, -1.413, -0.385, 0.095, -0.547, -1.553, -0.426, -0.891, -1.304 , -1.357, -0.557)	(0.993, 0.048,0.048)

Table B.7: Model parameters for predicting substituent exocyclic torsion angles. Each model's performance is given in terms of circular correlation, R^2_{circ}, mean angular error, MAE, and standard deviation (s.d.) of angular error.

Substituent(s)	Model Parameters (Intercept, Slope)	(R^2_{circ}, MAE, s.d.)
Carbonyl (C=O)	$(-\pi,\ 1)$	(0.998,0.040,0.039)
Methyl (CH3)	$(-\frac{2\pi}{3},\ 1)$, $(-\frac{5\pi}{6},\ 6)$, $(-\pi,\ 1)$, $(\frac{5\pi}{6},\ 1)$, $(\frac{2\pi}{3},\ 1)$	(0.997,0.052,0.039),(0.994,0.061,0.041),(0.993,0.029,0.028), (0.998,0.056,0.039), (0.998,0.047,0.038)
Hydroxyl (OH)	$(-\frac{2\pi}{3},\ 1)$, $(\frac{2\pi}{3},\ 1)$	(0.998,0.054,0.038),(0.998,0.057,0.040)
Alkoxy (-O-)	$(-\frac{2\pi}{3},\ 1)$, $(-\pi,\ 1)$, $(\frac{2\pi}{3},\ 1)$	(0.998,0.044,0.039),(0.980,0.038,0.037),(0.998,0.048,0.041)
Bulky carbon (-CH0, CH1, CH2)	$(-\frac{2\pi}{3},\ 1)$, $(-\frac{5\pi}{6},\ 6)$, $(\frac{5\pi}{6},\ 1)$, $(\frac{2\pi}{3},\ 1)$	(0.996, 0.064, 0.052),(0.972, 0.063, 0.036), (0.975, 0.082, 0.049),(0.996, 0.068, 0.054)
Halogen (-F)	$(-\frac{2\pi}{3},\ 1)$, $(\frac{2\pi}{3},\ 1)$	(0.998,0.046,0.033), (0.998,0.048,0.036)
Halogen (-Cl)	$(-\frac{2\pi}{3},\ 1)$, $(\frac{2\pi}{3},\ 1)$	(0.996,0.060,0.048), (0.997,0.054,0.043)
Halogen (-Br)	$(-\frac{2\pi}{3},\ 1)$, $(\frac{2\pi}{3},\ 1)$	(0.998,0.057,0.045), (0.998,0.052,0.052)

Table B.8: Molecules in the benchmarks set.

Name	SMILES
cyclopentane	C1CCCC1
cyclopentanol	C1CCCC1O
cyclopentene	C1=CCCC1
2-methylcyclopent-2-en-1-ol	CC1=CCCC1O
cyclopent-2-ene-1,5-dione	C1C=CC(=O)C1=O
cyclohexane	C1CCCCC1
methylcyclohexane	CC1CCCCC1
4,4-dimethylcyclohexanone	O=C1CCC(C)(C)CC1
cyclohexene	C1=CCCCC1
1-methyl-1-cyclohexene	CC1=CCCCC1
cyclohexa-1,3-diene	C1=CC=CCC1
cyclohexa-1,4-diene	C1=CCC=CC1
2,5-cyclohexadienone	C1C=CC(=O)C=C1
cycloheptane	C1CCCCCC1
cycloheptanone	O=C1CCCCCC1
cycloheptene	C1=CCCCCC1
cyclohept-2-en-1-one	O=C1C=CCCCC1
cyclohepta-1,3-diene	C1=CC=CCCC1
cyclohepta-1,4-diene	C1=CCC=CCC1
6-methylcyclohepta-1,4-diene	C1=CCC=CCC1C

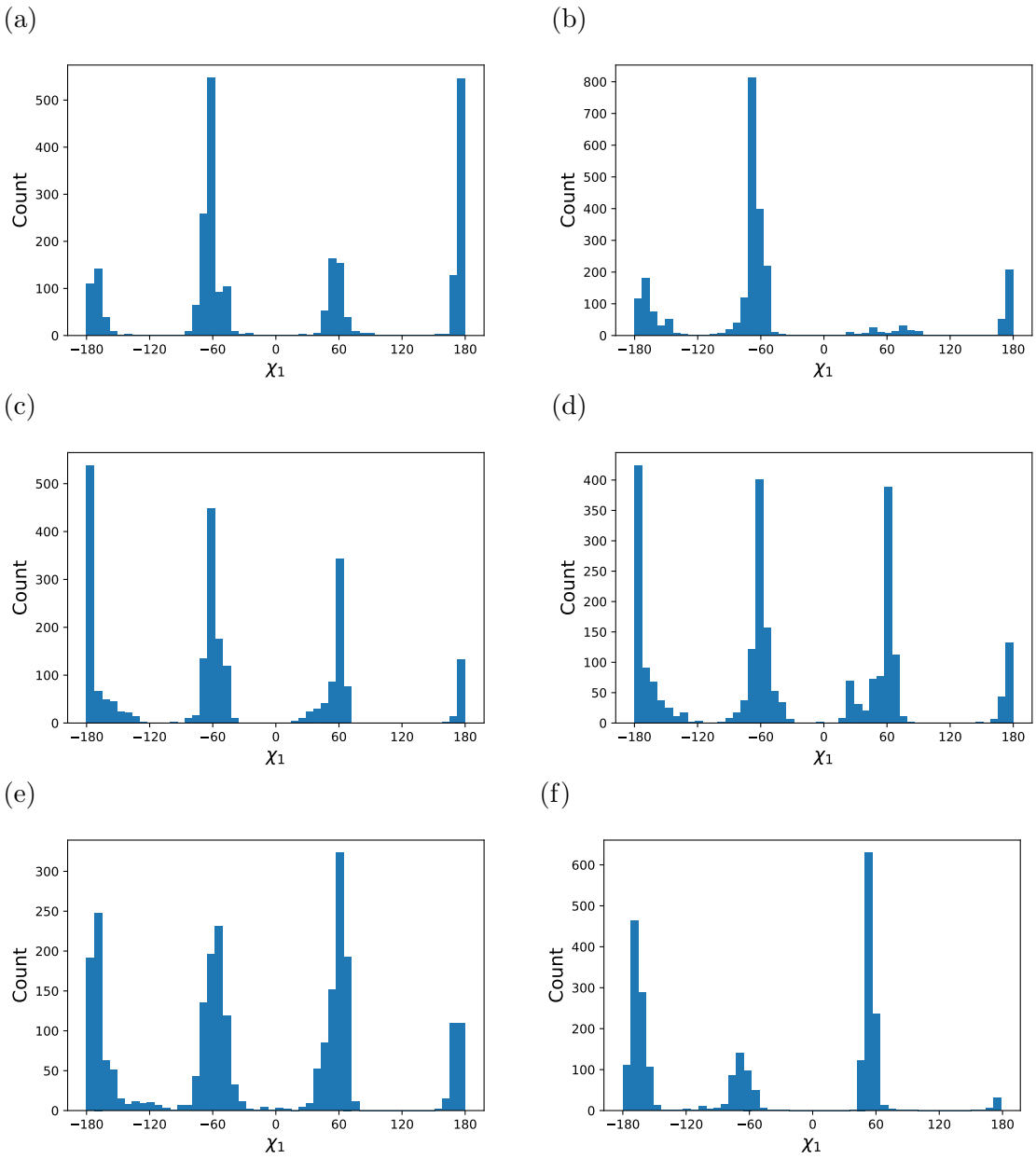


Figure B.1: Histogram of χ_1 angles for 12 amino acids. (a) Valine; (b) Leucine; (c) Phenylalanine; (d) Tyrosine; (e) Tryptophan; (f) Cysteine; (g) Serine; (h) Threonine ; (i) Histidine; (j) Lysine; (k) Aspartic acid; (l) Glutamic Acid. The χ_1 angles of all amino acid side chains in CTPs exhibit multimodal, which are similar to the one observed in normal protein side chains.

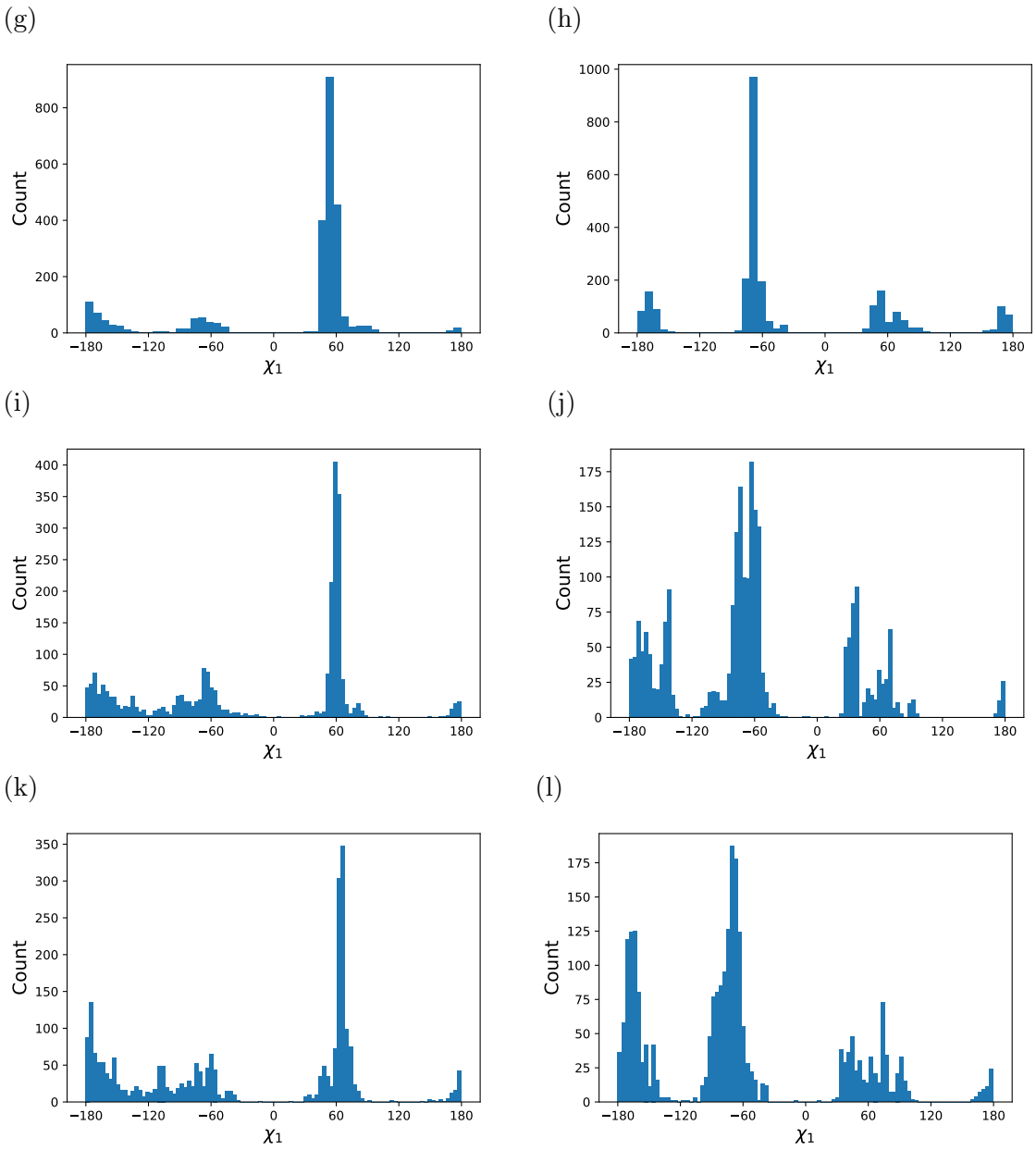


Figure B.1: Histogram of χ_1 angles for 12 amino acids. (a) Valine; (b) Leucine; (c) Phenylalanine; (d) Tyrosine; (e) Tryptophan; (f) Cysteine; (g) Serine; (h) Threonine ; (i) Histidine; (j) Lysine; (k) Aspartic acid; (l) Glutamic Acid. The χ_1 angles of all amino acid side chains in CTPs exhibit multimodal, which are similar to the one observed in normal protein side chains.

Table B.9: RMSD and TFD values between sampled conformations and reference conformations. The TFD values are calculated using conformation with the lowest RMSD value.

Molecule	RMSD	TFD
cyclopentane	0.020	0.058
cyclopentanol	0.075	0.167
cyclopentene	0.053	0.039
2-methylcyclopent-2-en-1-ol	0.069	0.013
cyclopent-2-ene-1,5-dione	0.072	0.031
cyclohexane	0.057	0.063
methylcyclohexane	0.111	0.029
4,4-dimethylcyclohexanone	0.164	0.049
cyclohexene	0.073	0.012
1-methyl-1-cyclohexene	0.066	0.067
cyclohexa-1,3-diene	0.079	0.048
cyclohexa-1,4-diene	0.088	0.015
2,5-cyclohexadienone	0.085	0.013
cycloheptane	0.115	0.055
cycloheptanone	0.130	0.007
cycloheptene	0.053	0.003
cyclohept-2-en-1-one	0.166	0.099
cyclohepta-1,3-diene	0.090	0.090
cyclohepta-1,4-diene	0.104	0.044
6-methylcyclohepta-1,4-diene	0.168	0.050

Appendix C

Table C.1: GFN2-computed conformational entropies of 70 cyclic molecules

	SMILES	RF	Entropy
0	C1CCCCC1	3.0	5.792
1	C1=CCCCC1	2.0	5.825
2	C1=CCC=CC1	1.0	-0.000
3	C1=CCCC=C1	1.0	5.763
4	O=C1CCCCC1	3.0	7.116
5	C1CCCCCCC1	5.0	28.186
6	C1CCCNCCC1	5.0	24.067
7	C1CCNCCNC1	5.0	24.388
8	C1=C\CCCCCC/1	4.0	12.114
9	OC1CCCCCCC1	5.0	27.033
10	c1ccccc1	0.0	-0.000
11	c1ccncc1	0.0	-0.000
12	c1cncnc1	0.0	-0.000
13	C1CCCC1	2.0	19.145
14	C1=CCCC1	1.0	5.763
15	C1CCCCCC1	4.0	22.105
16	c1ccc2c(c1)CCCC2	2.0	6.996
17	C1CCCCCCCC1	6.0	29.889
18	C1CCCCCCCCC1	7.0	37.845
19	C1CCCCCCCCCC1	8.0	41.701
20	C1CCCCCCCCCCC1	9.0	23.576
21	C1CCCCCCCCCCCCC1	11.0	37.458
22	C1CCCCCCCCCCCCCCC1	13.0	42.363
23	c1ccc2c(c1)Cc1ccccc1-2	0.0	-0.000
24	C1CCC2=C(C1)CCCC2	4.0	11.441
25	C1=C\C2=C(/CCCC/1)CCCCCCC2	8.0	22.727
26	c1ccc2c(c1)CCCCCC2	4.0	11.571
27	C1=C\c2ccccc2CCCC/1	3.0	11.115

Continued on next page

	SMILES	RF	Entropy
28	c1ccc2c(c1)CCCc1ccccc1C2	2.0	9.115
29	c1ccc2c(c1)CCCCc1ccccc1-2	2.0	0.683
30	C1CCOC1	2.0	5.763
31	C1COCCOCCO1	6.0	18.070
32	C1COCCOCCOCCO1	9.0	36.030
33	C1COCCOCCOCCOCCO1	12.0	49.560
34	C1COCCO1	3.0	5.807
35	C1CCC2(CC1)CCCC2	5.0	16.420
36	C1CCC2(CC1)CCCCC2	6.0	11.827
37	O=C1CCCC1	2.0	5.763
38	O=C1CCCCCC1	4.0	16.918
39	OC1CCCCCC1	4.0	22.619
40	NC1CCCCCC1	4.0	27.206
41	NC1CCCCC1	3.0	15.512
42	NC1CCCCCCC1	5.0	31.236
43	OC1CCCCCCCC1	6.0	27.601
44	C1CCCC2(CC1)CCCCCC2	8.0	13.360
45	C1CCC1	1.0	5.763
46	C1COC1	1.0	-0.000
47	C1CNCC2=C(C1)C1=C2CCCNCN1	7.0	21.922
48	C1CC[C@H]2CCCC[C@@H]2C1	0.0	0.015
49	C1CC[C@@H]2CCC[C@H]2C1	0.0	0.027
50	C1CC[C@H]2CNCC[C@@H]2C1	0.0	0.022
51	C1CC[C@H]2C[C@H]3CCCC[C@@H]3C[C@@H]2C1	0.0	0.015
52	C1CC[C@@H]2[C@@H](C1)C[C@H]1CCCC[C@H]12	1.5	5.594
53	C1CCC2=C(C1)C[C@@H]1CCCC[C@H]21	3.5	14.517
54	c1ccc2c(c1)C[C@@H]1CCCC[C@H]21	1.5	6.775
55	C1C[C@H]2CC[C@@H]1C2	0.0	-0.000
56	C1CC[C@H]2CNC[C@@H](C1)C2	5.0	12.943
57	C1CC[C@@H]2CCC[C@H](C1)C2	5.0	9.699
58	c1ccc2c(c1)CCCC[C@H]1CCCC[C@@H]21	3.0	3.362
59	c1ccc2c(c1)CCCC[C@H]1CCCC[C@H]21	0.0	1.142
60	c1ccc2c(c1)CCCC[C@H]1CCCC[C@@H]1CCC2	3.0	15.566
61	O=C1CCC[C@H]1O	2.0	2.414
62	C1CC2CCC1CC2	9.0	5.763
63	C1C[C@H]2CCC[C@@H](C1)C2	4.0	5.479
64	C1CC2CCCC(C1)CCC2	15.0	5.763
65	CC1(C)[C@H]2CC[C@]1(C)C(=O)C2	2.0	27.403
66	C1CC[C@H]2C[C@H]2CC1	2.5	0.915
67	C1CC[C@H]2[C@@H](CC1)[C@@H]1CCCC[C@H]21	4.0	12.946
68	C1CCC[C@@H]2[C@H](CC1)[C@H]1CCCCC[C@@H]21	7.5	7.075

Continued on next page

	SMILES	RF	Entropy
69	C1CNCN[C@@H]2[C@H](C1)[C@H]1CNCCC[C@H]12	6.0	9.765

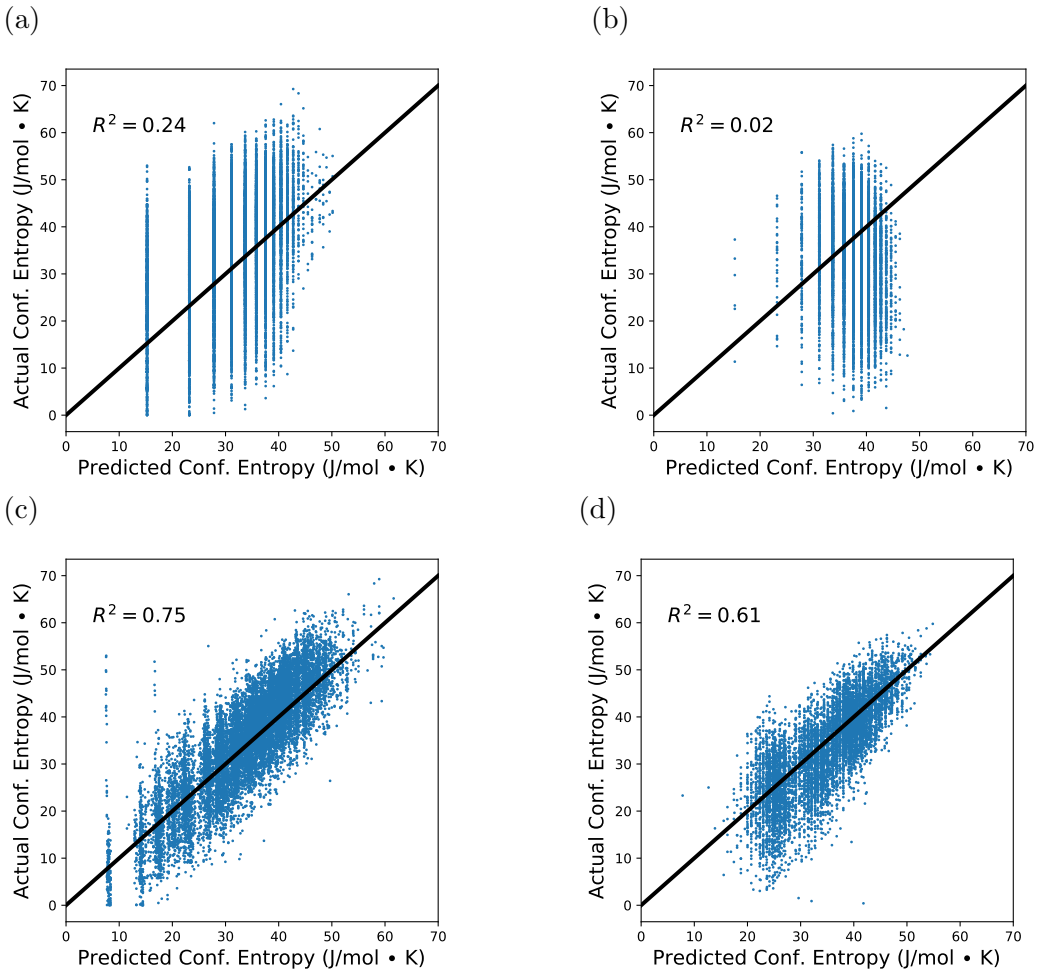
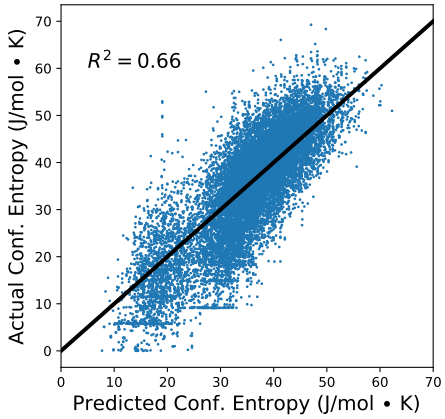
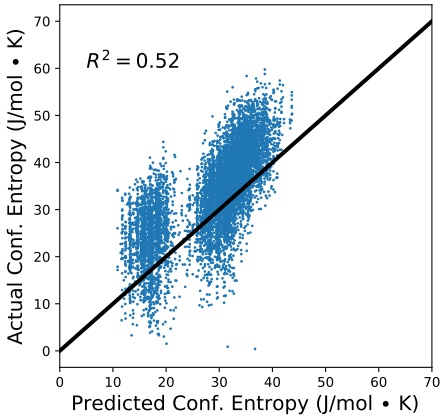


Figure C.1: Correlation between predicted entropies and GFN2-computed entropies on ZINC-I (left) and peptide (right) test sets. (a)-(b) LR-1 (linear model with number of rotatable bonds as sole descriptor); (c)-(d) LR-Best; (e)-(f) LASSO; (g)-(h) Ridge Regression ; (i)-(j) Kernel Ridge Regression (KRR); (k)-(l) Neural Network (NN).

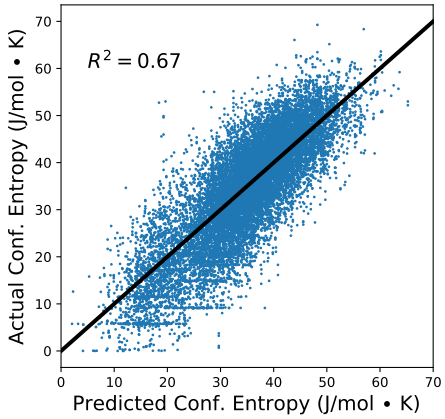
(e)



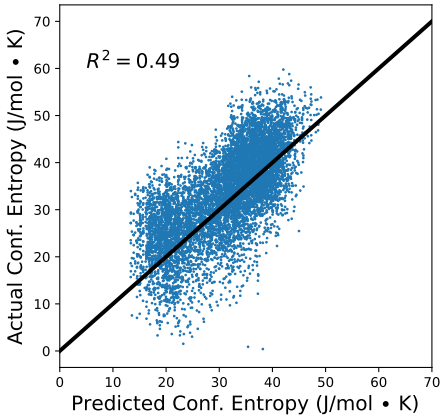
(f)



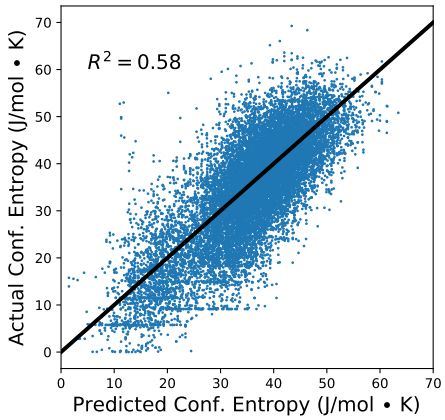
(g)



(h)



(i)



(j)

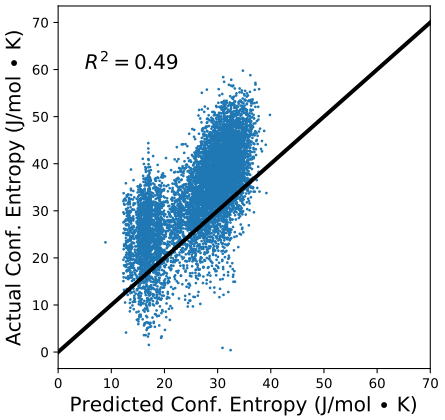
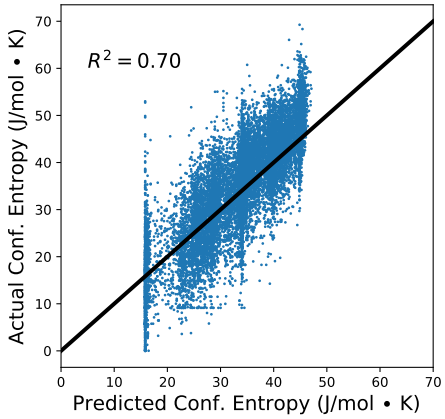


Figure C.1: (Continued)

(k)



(l)

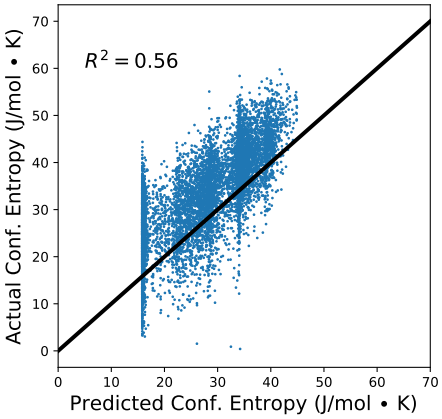


Figure C.1: (Continued)

Bibliography

Al-Jallal, N. A., Al-Kahtani, A. A., and El-Azhary, A. A. (2005). Conformational Study of the Structure of Free 18-Crown-6. *The Journal of Physical Chemistry A*, 109(16):3694–3703.

Alibay, I. and Bryce, R. A. (2019). Ring Puckering Landscapes of Glycosaminoglycan-Related Monosaccharides from Molecular Dynamics Simulations. *Journal of Chemical Information and Modeling*, 59(11):4729–4741.

Altona, C. and Sundaralingam, M. (1972). Conformational analysis of the sugar ring in nucleosides and nucleotides. New description using the concept of pseudorotation. *Journal of the American Chemical Society*, 94(23):8205–8212.

Anet, F. A. L. and Cheng, A. K. (1975). Conformation of cyclohexadecane. *Journal of the American Chemical Society*, 97(9):2420–2424.

Appavoo, S. D., Huh, S., Diaz, D. B., and Yudin, A. K. (2019). Conformational Control of Macrocycles by Remote Structural Modification. *Chemical Reviews*, 119(17):9724–9752.

Armstrong, M. S., Morris, G. M., Finn, P. W., Sharma, R., Moretti, L., Cooper, R. I., and Richards, W. G. (2010). Electroshape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. *Journal of Computer-Aided Molecular Design*, 24(9).

Ash, S., Cline, M. A., Homer, R. W., Hurst, T., and Smith, G. B. (1997). SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation. *Journal of Chemical Information and Computer Sciences*, 37(1):71–79.

Ballester, P. J. and Richards, W. G. (2007). Ultrafast shape recognition for similarity search in molecular databases. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 463(2081):1307–1321.

Bannwarth, C., Ehlert, S., and Grimme, S. (2019). GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *Journal of Chemical Theory and Computation*, 15(3):1652–1671.

Becke, A. D. (1988). Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A*, 38:3098–3100.

Becke, A. D. (1992). Density-functional thermochemistry. I. The effect of the exchange-only gradient correction. *The Journal of Chemical Physics*, 96(3):2155–2160.

Begel, S., Puchta, R., and van Eldik, R. (2014). Host-guest complexes of calix[4]tubes - prediction of ion selectivity by quantum chemical calculations VI. *Journal of Molecular Modeling*, 20(4).

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242.

Bilton, C., Allen, F. H., Shields, G. P., and Howard, J. A. K. (2000). Intramolecular hydrogen bonds: common motifs, probabilities of formation and implications for supramolecular organization. *Acta Crystallographica Section B*, 56(5):849–856.

Bocian, D. F., Pickett, H. M., Rounds, T. C., and Strauss, H. L. (1975). Conformations of cycloheptane. *Journal of the American Chemical Society*, 97(4):687–695.

Born, M. and Oppenheimer, R. (1927). Zur quantentheorie der molekeln. *Annalen der Physik*, 389(20):457–484.

Brain, Z. E. and Addicoat, M. A. (2011). Optimization of a genetic algorithm for searching molecular conformer space. *The Journal of Chemical Physics*, 135(17):174106.

Brochu, E., Cora, V. M., and De Freitas, N. (2010). A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. *arXiv preprint arXiv:1012.2599*.

Brooks, B. R., Brooks III, C. L., Mackerell Jr., A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post,

C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M., and Karplus, M. (2009). CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry*, 30(10):1545–1614.

CCG (2018). Molecular Operating Environment.

Chan, L., Hutchison, G. R., and Morris, G. M. (2019). Bayesian optimization for conformer generation. *Journal of Cheminformatics*, 11(1):32.

Chan, L., Hutchison, G. R., and Morris, G. M. (2020a). BOKEI: Bayesian Optimization using Knowledge of Correlated Torsions and Expected Improvement for Conformer Generation. *Physical Chemistry Chemical Physics*, 22:5211–5219.

Chan, L., Hutchison, G. R., and Morris, G. M. (2020b). Understanding Ring Puckering in Small Molecules and Cyclic Peptides. *ChemRxiv 10.26434/chemrxiv.12999938*.

Chan, L., Morris, G. M., and Hutchison, G. R. (2020c). Understanding Conformational Entropy in Small Molecules. *ChemRxiv :10.26434/chemrxiv.12671027.v1*.

Chang, C. A., Chen, W., and Gilson, M. K. (2007). Ligand configurational entropy and protein binding. *Proceedings of the National Academy of Sciences*, 104(5):1534–1539.

Chang, C.-E., Chen, W., and Gilson, M. K. (2005). Evaluating the Accuracy of the Quasiharmonic Approximation. *Journal of Chemical Theory and Computation*, 1(5):1017–1028.

Chang, G., Guida, W. C., and Still, W. C. (1989). An internal-coordinate Monte Carlo method for searching conformational space. *Journal of the American Chemical Society*, 111(12):4379–4386.

Chollet, F. (2015). Keras. `https://keras.io`.

Christensen, A. S., Kubař, T., Cui, Q., and Elstner, M. (2016). Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications. *Chemical Reviews*, 116(9):5301–5337.

Cole, J. C., Korb, O., McCabe, P., Read, M. G., and Taylor, R. (2018). Knowledge-Based Conformer Generation Using the Cambridge Structural Database. *Journal of Chemical Information and Modeling*, 58(3):615–629.

Coutsias, E. A., Lexa, K. W., Wester, M. J., Pollock, S. N., and Jacobson, M. P. (2016). Exhaustive Conformational Sampling of Complex Fused Ring Macrocycles Using Inverse Kinematics. *Journal of Chemical Theory and Computation*, 12(9):4674–4687.

Cremer, D. (1980). A General Definition of Ring Substituent Positions. *Israel Journal of Chemistry*, 20(1-2):12–19.

Cremer, D. and Pople, J. A. (1975). General definition of ring puckering coordinates. *Journal of the American Chemical Society*, 97(6):1354–1358.

Dalke (2014). *MACCS Structural Keys*.

Davies, G. J., Planas, A., and Rovira, C. (2012). Conformational Analyses of the Reaction Coordinate of Glycosidases. *Accounts of Chemical Research*, 45(2):308–316.

Daylight (2011). *SMARTS Theory Manual Daylight Chemical Information Systems*.

de Leeuw, F. A., Kampen, P. N. V., Altona, C., Díez, E., and Esteban, A. L. (1984). Relationships between torsion angles and ring-puckering coordinates: Part III. Application to heterocyclic puckered five-membered rings. *Journal of Molecular Structure*, 125(1):67 – 88.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Devereux, C., Smith, J. S., Davis, K. K., Barros, K., Zubatyuk, R., Isayev, O., and Roitberg, A. E. (2020). Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *Journal of Chemical Theory and Computation*, 16(7):4192–4202.

Dewar, M. J. S., Zoebisch, E. G., Healy, E. F., and Stewart, J. J. P. (1985). Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *Journal of the American Chemical Society*, 107(13):3902–3909.

Dill, K. A. and Chan, H. S. (1997). From Levinthal to pathways to funnels. *Nature Structural & Molecular Biology*, 4(1):10–19.

Dragojlovic, V. (2015). Conformational analysis of cycloalkanes. *ChemTexts*, 1(3):14.

Driggers, E. M., Hale, S. P., Lee, J., and Terrett, N. K. (2008). The exploration of macrocycles for drug discovery–an underexploited structural class. *Nature Reviews. Drug Discovery*, 7(7):608—624.

Dunbrack, R. L. (2002). Rotamer Libraries in the 21st Century. *Current Opinion in Structural Biology*, 12(4):431 – 440.

Duvenaud, D. (2014). *Automatic model construction with Gaussian processes*. PhD thesis, University of Cambridge.

Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints. In *Advances in Neural Information Processing Systems 28*, pages 2224–2232.

Ebejer, J.-P., Morris, G. M., and Deane, C. M. (2012). Freely Available Conformer Generation Methods: How Good Are They? *Journal of Chemical Information and Modeling*, 52(5):1146–1158.

Edman, P. (1959). Chemistry of amino acids and peptides. *Annual Review of Biochemistry*, 28(1):69–96.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32(2):407–499.

El-Azhary, A. A. and Al-Kahtani, A. A. (2005). Conformational Study of the Structure of 12-crown-4-Alkali Metal Cation Complexes. *The Journal of Physical Chemistry A*, 109(35):8041–8048.

Ellingson, B. A., Lynch, V. A., Mielke, S. L., and Truhlar, D. G. (2006). Statistical thermodynamics of bond torsional modes: Tests of separable, almost-separable, and improved Pitzer–Gwinn approximations. *The Journal of Chemical Physics*, 125(8):084305.

Fisher, N. I. and Lee, A. J. (1983). A Correlation Coefficient for Circular Data. *Biometrika*, 70(2):327–332.

Flachsenberg, F., Andresen, N., and Rarey, M. (2017). RingDecomposerLib: An Open-Source Implementation of Unique Ring Families and Other Cycle Bases. *Journal of Chemical Information and Modeling*, 57(2):122–126.

Flores-Ortega, A., Casanovas, J., Zanuy, D., Nussinov, R., and Alemán, C. (2007). Conformations of Proline Analogues Having Double Bonds in the Ring. *The Journal of Physical Chemistry B*, 111(19):5475–5482.

Folmsbee, D. and Hutchison, G. (2020). Assessing conformer energies using electronic structure and machine learning methods. *International Journal of Quantum Chemistry*, page e26381.

Forli, S. and Botta, M. (2007). Lennard-Jones Potential and Dummy Atom Settings to Overcome the AUTODOCK Limitation in Treating Flexible Ring Systems. *Journal of Chemical Information and Modeling*, 47(4):1481–1492.

Friedrich, N.-O., de Bruyn Kops, C., Flachsenberg, F., Sommer, K., Rarey, M., and Kirchmair, J. (2017). Benchmarking Commercial Conformer Ensemble Generators. *Journal of Chemical Information and Modeling*, 57(11):2719–2728.

Friedrich, N.-O., Flachsenberg, F., Meyder, A., Sommer, K., Kirchmair, J., and Rarey, M. (2019). Conformator: A Novel Method for the Generation of Conformer Ensembles. *Journal of Chemical Information and Modeling*, 59(2):731–742.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M. P., Overington, J. P., Papadatos, G., Smit, I., and Leach, A. R. (2016). The ChEMBL database in 2017. *Nucleic Acids Research*, 45(D1):D945–D954.

Gelbart, M. A., Snoek, J., and Adams, R. P. (2014). Bayesian Optimization with Unknown Constraints. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, page 250–259.

Ghahremanpour, M. M., van Maaren, P. J., Ditz, J. C., Lindh, R., and van der Spoel, D. (2016). Large-scale calculations of gas phase thermochemistry: Enthalpy of formation, standard entropy, and heat capacity. *The Journal of Chemical Physics*, 145(11):114305.

Giordanetto, F. and Kihlberg, J. (2014). Macrocyclic Drugs and Clinical Candidates: What Can Medicinal Chemists Learn from Their Properties? *Journal of Medicinal Chemistry*, 57(2):278–295.

Gobbi, A. and Poppinger, D. (1998). Genetic optimization of combinatorial libraries. *Biotechnology and Bioengineering*, 61(1):47–54.

Gong, H.-Y., Wang, D.-X., Zheng, Q.-Y., and Wang, M.-X. (2009). Highly selective complexation of metal ions by the self-tuning tetraazacalixpyridine macrocycles. *Tetrahedron*, 65(1):87 – 92.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

GPyOpt (2016). GPyOpt: A Bayesian Optimization framework in Python. `http://github.com/SheffieldML/GPyOpt`.

Gražulis, S., Chateigner, D., Downs, R. T., Yokochi, A. F. T., Quirós, M., Lutterotti, L., Manakova, E., Butkus, J., Moeck, P., and Le Bail, A. (2009). Crystallography Open Database – an open-access collection of crystal structures. *Journal of Applied Crystallography*, 42(4):726–729.

Gražulis, S., Daškevič, A., Merkys, A., Chateigner, D., Lutterotti, L., Quirós, M., Serebryanaya, N. R., Moeck, P., Downs, R. T., and Le Bail, A. (2012). Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Research*, 40(D1):D420–D427.

Griffiths, R.-R. and Hernández-Lobato, J. M. (2020). Constrained Bayesian optimization for automatic chemical design using variational autoencoders. *Chemical Science*, 11:577–586.

Grimme, S. (2012). Supramolecular Binding Thermodynamics by Dispersion-Corrected Density Functional Theory. *Chemistry – A European Journal*, 18(32):9955–9964.

Grimme, S. (2019). Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations. *Journal of Chemical Theory and Computation*, 15(5):2847–2862.

Grimme, S., Bannwarth, C., and Shushkov, P. (2017). A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1–86). *Journal of Chemical Theory and Computation*, 13(5):1989–2009.

Groom, C. R., Bruno, I. J., Lightfoot, M. P., and Ward, S. C. (2016). The Cambridge Structural Database. *Acta Crystallographica Section B*, 72(2):171–179.

Guba, W., Meyder, A., Rarey, M., and Hert, J. (2016). Torsion Library Reloaded: A New Version of Expert-Derived SMARTS Rules for Assessing Conformations of Small Molecules. *Journal of Chemical Information and Modeling*, 56(1):1–5.

Gutsche, C. D. and Bauer, L. J. (1985). Calixarenes. 13. The conformational properties of calix[4]arenes, calix[6]arenes, calix[8]arenes, and oxacalixarenes. *Journal of the American Chemical Society*, 107(21):6052–6059.

Haasnoot, C. A. G. (1992). The conformation of six-membered rings described by puckering coordinates derived from endocyclic torsion angles. *Journal of the American Chemical Society*, 114(3):882–887.

Halgren, T. A. (1992). The representation of van der Waals (vdW) interactions in molecular mechanics force fields: potential form, combination rules, and vdW parameters. *Journal of the American Chemical Society*, 114(20):7827–7843.

Halgren, T. A. (1996). Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry*, 17(5-6):490–519.

Halgren, T. A. and Nachbar, R. B. (1996). Merck molecular force field. IV. conformational energies and geometries for MMFF94. *Journal of Computational Chemistry*, 17(5-6):587–615.

Hancock, R. D. and Martell, A. E. (1989). Ligand design for selective complexation of metal ions in aqueous solution. *Chemical Reviews*, 89(8):1875–1914.

Hansen, N. (2006). *The CMA Evolution Strategy: A Comparing Review*, pages 75–102. Springer Berlin Heidelberg.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Hartshorn, M. J., Verdonk, M. L., Chessari, G., Brewerton, S. C., Mooij, W. T. M., Mortenson, P. N., and Murray, C. W. (2007). Diverse, High-Quality Test Set for the Validation of Protein-Ligand Docking Performance. *Journal of Medicinal Chemistry*, 50(4):726–741.

Havel, T. F., Kuntz, I. D., and Crippen, B. (1985). The theory and practice of distance geometry. *Bulletin of Mathematical Biology*, 47(1):157–157.

Hawkins, P. C. D. (2017). Conformation Generation: The State of the Art. *Journal of Chemical Information and Modeling*, 57(8):1747–1756.

Hawkins, P. C. D., Skillman, A. G., Warren, G. L., Ellingson, B. A., and Stahl, M. T. (2010). Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *Journal of Chemical Information and Modeling*, 50(4):572–584.

Head, M. S., Given, J. A., and Gilson, M. K. (1997). "Mining Minima": Direct Computation of Conformational Free Energy. *The Journal of Physical Chemistry A*, 101(8):1609–1618.

Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., and Pletnev, I. (2013). InChI - the worldwide chemical structure identifier standard. *Journal of Cheminformatics*, 5(1).

Hill, A. D. and Reilly, P. J. (2007). Puckering Coordinates of Monocyclic Rings by Triangular Decomposition. *Journal of Chemical Information and Modeling*, 47(3):1031–1035.

Hill, T. L. (1986). *An Introduction to Statistical Thermodynamics*. Courier Corporation.

Hofmann, T., Schölkopf, B., and Smola, A. J. (2008). Kernel methods in machine learning. *The Annals of Statistics*, pages 1171–1220.

Hohenberg, P. and Kohn, W. (1964). Inhomogeneous Electron Gas. *Physical Review*, 136:B864–B871.

Hyttinen, N. and Prisle, N. L. (2020). Improving Solubility and Activity Estimates of Multifunctional Atmospheric Organics by Selecting Conformers in COSMOtherm. *The Journal of Physical Chemistry A*, 124(23):4801–4812.

Häse, F., Roch, L. M., Kreisbeck, C., and Aspuru-Guzik, A. (2018). Phoenics: A Bayesian Optimizer for Chemistry. *ACS Central Science*, 4(9):1134–1145.

Imrie, F., Bradley, A. R., van der Schaar, M., and Deane, C. M. (2020). Deep Generative Models for 3D Linker Design. *Journal of Chemical Information and Modeling*, 60(4):1983–1995.

Ionescu, A. R., Bérces, A., Zgierski, M. Z., Whitfield, D. M., and Nukada, T. (2005). Conformational Pathways of Saturated Six-Membered Rings. A Static and Dynamical Density Functional Study. *The Journal of Physical Chemistry A*, 109(36):8096–8105.

Irikura, K. K. (1998). Appendix B: Essential Statistical Thermodynamics. *Computational Thermochemistry*, pages 402–418.

Irikura, K. K. (2020). How much does a methyl rotor (internal rotation) contribute to the entropy? https://cccbdb.nist.gov/methylrotor.asp.

Jammalamadaka, S. R. and Sengupta, A. (2001). *Topics in circular statistics*. World Scientific.

Jones, J. E. (1924). On the determination of molecular fields. —II. From the equation of state of a gas. *Proceedings of the Royal Society of London. Series A*, 106(738):463–477.

Kaminský, J. and Jensen, F. (2007). Force Field Modeling of Amino Acid Conformational Energies. *Journal of Chemical Theory and Computation*, 3(5):1774–1788.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. (2018). Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*.

Kanal, I. Y., Keith, J. A., and Hutchison, G. R. (2018). A sobering assessment of small-molecule force field methods for low energy conformer predictions. *International Journal of Quantum Chemistry*, 118(5):e25512.

Kilpatrick, J. E., Pitzer, K. S., and Spitzer, R. (1947). The Thermodynamics and Molecular Structure of Cyclopentane[1]. *Journal of the American Chemical Society*, 69(10):2483–2488.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., and Bolton, E. E. (2018). PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research*, 47(D1):D1102–D1109.

Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization.

Kitchen, D. B., Decornez, H., Furr, J. R., and Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery*, 3(11):935–949.

Kolodzik, A., Urbaczek, S., and Rarey, M. (2012). Unique Ring Families: A Chemically Meaningful Description of Molecular Ring Topologies. *Journal of Chemical Information and Modeling*, 52(8):2013–2021.

Kuhn, B., Mohr, P., and Stahl, M. (2010). Intramolecular Hydrogen Bonding in Medicinal Chemistry. *Journal of Medicinal Chemistry*, 53(6):2601–2611.

Labute, P. (2010). LowModeMD—Implicit Low-Mode Velocity Filtering Applied to Conformational Search of Macrocycles and Protein Loops. *Journal of Chemical Information and Modeling*, 50(5):792–800.

Landrum, G. (2018). RDKit: Open-Source Cheminformatics.

Law, H. C., Zhao, P., Chan, L., Huang, J., and Sejdinovic, D. (2019). Hyperparameter Learning via Distributional Transfer. In *Advances in Neural Information Processing Systems*, pages 6804–6815.

Lee, A. A., Brenner, M. P., and Colwell, L. J. (2016). Predicting protein–ligand affinity with a random matrix framework. *Proceedings of the National Academy of Sciences*, 113(48):13564–13569.

Lee, C., Yang, W., and Parr, R. G. (1988). Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37:785–789.

Levinthal, C. (1968). Are there pathways for protein folding? *Journal de chimie physique*, 65:44–45.

Loiseau, N., Gomis, J.-M., Santolini, J., Delaforge, M., and André, F. (2003). Predicting the conformational states of cyclic tetrapeptides. *Biopolymers*, 69(3):363–385.

Lyu, S., Beiranvand, N., Freindorf, M., and Kraka, E. (2019). Interplay of Ring Puckering and Hydrogen Bonding in Deoxyribonucleosides. *The Journal of Physical Chemistry A*, 123(32):7087–7103.

Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., and Svetnik, V. (2015). Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. *Journal of Chemical Information and Modeling*, 55(2):263–274.

Maas, A. L., Hannun, A. Y., and Ng, A. Y. (2013). Rectifier Nonlinearities Improve Neural Network Acoustic Models. In *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*.

Mardia, K. V. and Frellsen, J. (2012). *Statistics of Bivariate von Mises Distributions*, pages 159–178. Springer Berlin Heidelberg.

Mardia, K. V., Hughes, G., Taylor, C. C., and Singh, H. (2008). A Multivariate Von Mises Distribution with Applications to Bioinformatics. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 36(1):99–109.

Mardia, K. V. and Jupp, P. E. (2009). *Directional statistics*. John Wiley & Sons.

Mardia, K. V., Taylor, C. C., and Subramaniam, G. K. (2007). Protein Bioinformatics and Mixtures of Bivariate von Mises Distributions for Angular Data. *Biometrics*, 63(2):505–512.

Marsault, E. and Peterson, M. L. (2011). Macrocycles Are Great Cycles: Applications, Opportunities, and Challenges of Synthetic Macrocycles in Drug Discovery. *Journal of Medicinal Chemistry*, 54(7):1961–2004.

Mayes, H. B., Broadbelt, L. J., and Beckham, G. T. (2014). How Sugars Pucker: Electronic Structure Calculations Map the Kinetic Landscape of Five Biologically Paramount Monosaccharides and Their Implications for Enzymatic Catalysis. *Journal of the American Chemical Society*, 136(3):1008–1022.

McCabe, P., Korb, O., and Cole, J. (2014). Kernel Density Estimation Applied to Bond Length, Bond Angle, and Torsion Angle Distributions. *Journal of Chemical Information and Modeling*, 54(5):1284–1288.

Mekenyan, O., Dimitrov, D., Nikolova, N., and Karabunarliev, S. (1999). Conformational Coverage by a Genetic Algorithm. *Journal of Chemical Information and Computer Sciences*, 39(6):997–1016.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Minary, P., Tuckerman, M. E., and Martyna, G. J. (2004). Long Time Molecular Dynamics for Enhanced Conformational Sampling in Biomolecular Systems. *Physical Review Letters*, 93:150201.

Mitchell, M. (1998). *An Introduction to Genetic Algorithms*. MIT Press.

Mockus, J., Tiesis, V., and Zilinskas, A. (1978). The application of Bayesian methods for seeking the extremum. *Towards Global Optimization*, 2(117-129):2.

Monod, M. Y., Giudicelli, V., Chaume, D., and Lefranc, M.-P. (2004). IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V–J and V–D–J JUNCTIONs. *Bioinformatics*, 20:i379–i385.

Morgan, S., Grootendorst, P., Lexchin, J., Cunningham, C., and Greyson, D. (2011). The cost of drug development: A systematic review. *Health Policy*, 100(1):4 – 17.

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., and Olson, A. J. (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*, 19(14):1639–1662.

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., and Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry*, 30(16):2785–2791.

Moss, G. (1996). Basic terminology of stereochemistry (IUPAC Recommendations 1996). *Pure and Applied Chemistry*, 68(12):2193–2222.

Nair, V. and Hinton, G. E. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, page 807–814.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. (2011a). Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1):33.

O'Boyle, N. M., Morley, C., and Hutchison, G. R. (2008). Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chemistry Central Journal*, 2(1):5.

O'Boyle, N. M., Vandermeersch, T., Flynn, C. J., Maguire, A. R., and Hutchison, G. R. (2011b). Confab - Systematic generation of diverse low-energy conformers. *Journal of Cheminformatics*, 3(1):8.

Pakes, P. W., Rounds, T. C., and Strauss, H. L. (1981). Conformations of cyclooctane and some related oxocanes. *The Journal of Physical Chemistry*, 85(17):2469–2475.

Paoloni, L., Rampino, S., and Barone, V. (2019). Potential-Energy Surfaces for Ring-Puckering Motions of Flexible Cyclic Molecules through Cremer–Pople Coordinates: Computation, Analysis, and Fitting. *Journal of Chemical Theory and Computation*, 15(7):4280–4294.

Pawar, D. M., Smith, S. V., Mark, H. L., Odom, R. M., and Noe, E. A. (1998). Conformational Study of Cyclodecane and Substituted Cyclodecanes by Dynamic NMR Spectroscopy and Computational Methods. *Journal of the American Chemical Society*, 120(41):10715–10720.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peter, C., Oostenbrink, C., van Dorp, A., and van Gunsteren, W. F. (2004). Estimating entropies from molecular dynamics simulations. *The Journal of Chemical Physics*, 120(6):2652–2661.

Petersson, G. A. and Al-Laham, M. A. (1991). A complete basis set model chemistry. II. Open-shell systems and the total energies of the first-row atoms. *The Journal of Chemical Physics*, 94(9):6081–6090.

Pracht, P., Bohle, F., and Grimme, S. (2020). Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Physical Chemistry Chemical Physics*, 22:7169–7192.

R Core Team (2020). R: A Language and Environment for Statistical Computing.

Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., and Koes, D. R. (2017). Protein–Ligand Scoring with Convolutional Neural Networks. *Journal of Chemical Information and Modeling*, 57(4):942–957.

Rahimi, A. and Recht, B. (2008). Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems 20*, pages 1177–1184.

Rai, B. K., Sresht, V., Yang, Q., Unwalla, R., Tu, M., Mathiowetz, A. M., and Bakken, G. A. (2019). Comprehensive Assessment of Torsional Strain in Crystal Structures of Small Molecules and Protein-Ligand Complexes using ab Initio Calculations. *Journal Of Chemical Information And Modeling*, 59(10):4195–4208.

Ralaivola, L., Swamidass, S. J., Saigo, H., and Baldi, P. (2005). Graph kernels for chemical informatics. *Neural Networks*, 18(8):1093 – 1110.

Ramachandran, G., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7(1):95 – 99.

Rappe, A. K., Casewit, C. J., Colwell, K. S., Goddard, W. A., and Skiff, W. M. (1992). UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society*, 114(25):10024–10035.

Riniker, S. and Landrum, G. A. (2015). Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *Journal of Chemical Information and Modeling*, 55(12):2562–2574.

Rogers, D. and Hahn, M. (2010). Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986a). *Learning Internal Representations by Error Propagation*, page 318–362. MIT Press.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986b). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Sahu, S., Rani Sahoo, P., Patel, S., and Mishra, B. (2011). Oxidation of thiourea and substituted thioureas: a review. *Journal of Sulfur Chemistry*, 32(2):171–197.

Salehinejad, H., Sankar, S., Barfett, J., Colak, E., and Valaee, S. (2017). Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078*.

Schneider, N., Lowe, D. M., Sayle, R. A., and Landrum, G. A. (2015). Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity. *Journal of Chemical Information and Modeling*, 55(1):39–53.

Schrödinger, LLC (2015). The PyMOL Molecular Graphics System.

Schrödinger, LLC (2020). Maestro, Schrödinger.

Schulz-Gasch, T., Schärfer, C., Guba, W., and Rarey, M. (2012). TFD: Torsion Fingerprints As a New Measure To Compare Small Molecule Conformations. *Journal of Chemical Information and Modeling*, 52(6):1499–1512.

Schwab, C. H. (2010). Conformations and 3D pharmacophore searching. *Drug Discovery Today: Technologies*, 7(4):e245 – e253.

Seabold, S. and Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. In *9th Python in Science Conference*.

Sega, M., Autieri, E., and Pederiva, F. (2011). Pickett angles and Cremer–Pople coordinates as collective variables for the enhanced sampling of six-membered ring conformations. *Molecular Physics*, 109(1):141–148.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):60.

Simón-Carballido, L., Bao, J. L., Alves, T. V., Meana-Pañeda, R., Truhlar, D. G., and Fernández-Ramos, A. (2017). Anharmonicity of Coupled Torsions: The Extended Two-Dimensional Torsion Method and Its Use To Assess More Approximate Methods. *Journal of Chemical Theory and Computation*, 13(8):3478–3492.

Sindhikara, D., Spronk, S. A., Day, T., Borrelli, K., Cheney, D. L., and Posy, S. L. (2017). Improving Accuracy, Diversity, and Speed with Prime Macrocycle Conformational Sampling. *Journal of Chemical Information and Modeling*, 57(8):1881–1894.

Singh, H., Hnizdo, V., and Demchuk, E. (2002). Probabilistic Model for Two Dependent Circular Variables. *Biometrika*, 89(3):719–723.

Slater, J. C. (1951). A Simplification of the Hartree-Fock Method. *Physical Review*, 81:385–390.

Sliwoski, G., Kothiwale, S., Meiler, J., and Lowe, E. W. (2014). Computational Methods in Drug Discovery. *Pharmacological Reviews*, 66(1):334–395.

Smith, J. S., Isayev, O., and Roitberg, A. E. (2017). ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical Science*, 8:3192–3203.

Smith, J. S., Nebgen, B. T., Zubatyuk, R., Lubbers, N., Devereux, C., Barros, K., Tretiak, S., Isayev, O., and Roitberg, A. E. (2019). Approaching coupled cluster

accuracy with a general-purpose neural network potential through transfer learning. *Nature Communications*, 10(1):2903.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, pages 2951–2959.

Snoek, J., Rippel, O., Swersky, K., Kiros, R., Satish, N., Sundaram, N., Patwary, M., Prabhat, M., and Adams, R. (2015). Scalable Bayesian Optimization using Deep Neural Networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, pages 2171–2180.

Spellmeyer, D. C., Wong, A. K., Bower, M. J., and Blaney, J. M. (1997). Conformational analysis using distance geometry methods. *Journal of Molecular Graphics and Modelling*, 15(1):18 – 36.

Sperandio, O., Souaille, M., Delfaud, F., Miteva, M. A., and Villoutreix, B. O. (2009). MED-3DMC: A new tool to generate 3D conformation ensembles of small molecules with a Monte Carlo sampling of the conformational space. *European Journal of Medicinal Chemistry*, 44(4):1405 – 1409.

Speybroeck, V. V., Vansteenkiste, P., Neck, D. V., and Waroquier, M. (2005). Why does the uncoupled hindered rotor model work well for the thermodynamics of *n*-alkanes? *Chemical Physics Letters*, 402(4):479 – 484.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. (2009). Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. *arXiv preprint arXiv:0912.3995*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Stephens, P. J., Devlin, F. J., Chabalowski, C. F., and Frisch, M. J. (1994). Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *The Journal of Physical Chemistry*, 98(45):11623–11627.

Sterling, T. and Irwin, J. J. (2015). ZINC 15 – Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337.

Stewart, J. J. P. (1989). Optimization of parameters for semiempirical methods I. Method. *Journal of Computational Chemistry*, 10(2):209–220.

Suárez, E., Díaz, N., and Suárez, D. (2011). Entropy Calculations of Single Molecules by Combining the Rigid–Rotor and Harmonic-Oscillator Approximations with Conformational Entropy Estimations from Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation*, 7(8):2638–2653.

Supady, A., Blum, V., and Baldauf, C. (2015). First-Principles Molecular Structure Search with a Genetic Algorithm. *Journal of Chemical Information and Modeling*, 55(11):2338–2348.

Thiel, W. (2014). Semiempirical quantum–chemical methods. *WIREs Computational Molecular Science*, 4(2):145–157.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. J. and Taylor, J. (2011). The Solution Path of the Generalized Lasso. *Annals of Statistics*, 39(3):1335–1371.

Ton, J.-F., Chan, L., Teh, Y. W., and Sejdinovic, D. (2019). Noise Contrastive Meta-Learning for Conditional Density Estimation using Kernel Mean Embeddings. *arXiv e-prints*, page arXiv:1906.02236.

Trott, O. and Olson, A. J. (2010). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461.

Tsujishita, H. and Hirono, S. (1997). CAMDAS: an automated conformational analysis system using molecular dynamics. *Journal of Computer-Aided Molecular Design*, 11(3):305–315.

Vainio, M. J. and Johnson, M. S. (2007). Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *Journal of Chemical Information and Modeling*, 47(6):2462–2474.

van der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science Engineering*, 13(2):22–30.

Vansteenkiste, P., Van Speybroeck, V., Marin, G. B., and Waroquier, M. (2003). Ab Initio Calculation of Entropy and Heat Capacity of Gas-Phase n-Alkanes Using Internal Rotations. *The Journal of Physical Chemistry A*, 107(17):3139–3145.

Verma, J. (2010). 3D-QSAR in Drug Design - A Review. *Current Topics in Medicinal Chemistry*, 10(1):95–115.

Villar, E. A., Beglov, D., Chennamadhavuni, S., Porco, J. A., Kozakov, D., Vajda, S., and Whitty, A. (2014). How proteins bind macrocycles. *Nature chemical biology*, 10:723 – 731.

Wales, D. J. and Doye, J. P. K. (1997). Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *The Journal of Physical Chemistry A*, 101(28):5111–5116.

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and Testing of a General AMBER Force Field. *Journal of Computational Chemistry*, 25(9):1157–1174.

Wang, S., Witek, J., Landrum, G. A., and Riniker, S. (2020). Improving Conformer Generation for Small Rings and Macrocycles Based on Distance Geometry and Experimental Torsional-Angle Preferences. *Journal of Chemical Information and Modeling*, 60(4):2044–2058.

Watts, K. S., Dalal, P., Tebben, A. J., Cheney, D. L., and Shelley, J. C. (2014). Macrocycle Conformational Sampling with MacroModel. *Journal of Chemical Information and Modeling*, 54(10):2680–2696.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36.

Wenzel, J., Matter, H., and Schmidt, F. (2019). Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *Journal of Chemical Information and Modeling*, 59(3):1253–1268.

Wicker, J. G. P. and Cooper, R. I. (2015). Will it crystallise? Predicting crystallinity of molecular materials. *CrystEngComm*, 17:1927–1934.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian Processes for Machine Learning*, volume 2. MIT press Cambridge, MA.

Williams, C. K. I. and Seeger, M. (2001). Using the Nyström Method to Speed Up Kernel Machines. In *Advances in Neural Information Processing Systems 13*, pages 682–688.

Wilson, S. R., Cui, W., Moskowitz, J. W., and Schmidt, K. E. (1991). Applications of simulated annealing to the conformational analysis of flexible molecules. *Journal of Computational Chemistry*, 12(3):342–349.

Wiswesser, W. J. (1954). *A Line-formula Chemical Notation*. Crowell.

Wu, J., Ning, H., Xu, X., and Ren, W. (2019). Accurate entropy calculation for large flexible hydrocarbons using a multi-structural 2-dimensional torsion method. *Physical Chemistry Chemical Physics*, 21(19):10003–10010.

Zamyatnin, A. (1972). Protein volume in solution. *Progress in Biophysics and Molecular Biology*, 24:107 – 123.

Zheng, J., Yu, T., Papajak, E., Alecu, I. M., Mielke, S. L., and Truhlar, D. G. (2011). Practical methods for including torsional anharmonicity in thermochemical calculations on complex molecules: The internal-coordinate multi-structural approximation. *Physical Chemistry Chemical Physics*, 13:10885–10907.

Zwanzig, R., Szabo, A., and Bagchi, B. (1992). Levinthal's paradox. *Proceedings of the National Academy of Sciences of the United States of America*, 89(1):20–22.