UNIVERSITÀ DEGLI STUDI DI TRENTO

Facoltà di Scienze Matematiche, Fisiche e Naturali

Dipartimento di Fisica

Tesi di Dottorato di Ricerca in Fisica
Ph.D. Thesis in Physics

Development of Free Energy calculation methods for the study of monosaccharides conformation in computer simulations

Supervisors:
Prof. Francesco Pederiva
Dr. Marcello Sega

Ph.D. candidate:
Emmanuel Autieri

Dottorato di Ricerca in Fisica, XXIV ciclo
Trento, 20th December 2011

Contents

List of Figures

List of Tables

Introduction

Jauchzet, frohlocket! auf, preiset die Tage,
Rühmet, was heute der Höchste getan!
Lasset das Zagen, verbannet die Klage,
Stimmet voll Jauchzen und Fröhlichkeit an!
Dienet dem Höchsten mit herrlichen Chören,
Lasst uns den Namen des Herrschers verehren!

J.S.Bach
Weihnachts-Oratorium (BWV 248I), 1.Coro

Cyclic and heterocyclic compounds are ubiquitous in nature, and occupy a place of particular relevance in many chemical and biological processes. Carbohydrates, for example, are one of the fundamental building blocks of the biochemical activity and structure of the cell, including energy transport, cell recognition and signaling. Since cyclic molecules appear most of the time in non-planar, puckered conformations, knowing their structural and conformational properties is a task of primary interest. This thesis is devoted to the study of the conformation of monosacchrides in six-membered ring form. The main goal is to develop and apply new computational tools to investigate conformational properties and to improve the description of carbohydrates in the framework of molecular dynamics simulations.

My interest in computer simulations of carbohydrates developed within the framework of a scientific collaboration on bio-polymeric gels[1]. One of the subjects of the research project was indeed the study of a polysaccharide derived from hyaluronic acid (HYA), a biopolimer composed by two repeated units (glucuronic acid and N-acetil-glucoseamine). Experimentally, a suitable functionalization can mutate the native hyaluronic acid into an hydrogel (HYADD4, see Ref. [55]). This, together with the biocompatibility of the molecule, makes this system a valid candidate for applications in medicine, for example in the treatment of ostheoarticular pathologies as a substitute of the synovial fluid in unhealthy joints. For the aforementioned system, as for many others, computer simulation are extremely valuable, as they can complement the information on structure and dynamics of the system gathered in the course of experiments. However, for this specific problem modeling the system within the molecular dynamics framework presented troublesome aspects. The most important issue was that the force field chosen (the GROMOS 45a4 parameter set [112]) failed in

[1]Part of the present work was supported by a PRIN grant from the Italian Ministry of Public Education, University and Scientific Research (PRIN 2007 project: "Proprietà fisiche di biomatrici nanostrutturate a base polimerica", A.Deriu, F.Pederiva, G.Paradossi).

reproducing the conformational preferences of the sugar constituents, with the appearance of unphysical conformations [36].

Starting from this evidence, we soon realized that this problem turns to be a known, but often overlooked one, in computer simulations of carbohydrates. In the past, several approaches have been proposed, which were however circumventing the problem, rather than providing a solution to it [76, 31, 150]. Indeed, we found that until recent times, pragmatic approaches were somehow supported by a general lack of experimental evidences on conformational data for sugars. This lack stems from the fact that the conformational analysis of monosaccharides, and in particular of hexopyranose sugars like glucose, is a difficult task both from the experimental and from the theoretical point of view. On the experimental side, the characterization of the ring conformational preference is complicated by the extremely low occurrence of the second-most populated conformer, with only few remarkable exceptions (like altrose [10] and idose [163]). From the perspective of computer simulations, this conformational behavior, dominated by few structures, generates a severe bottleneck: the non-ergodicity of the system by any practical means. This aspect explains the interest in free energy calculations, since metastabilities are strictly linked to specific free energy features. When free energy differences between conformers (and the barriers in between) are rather large, as it appears to be for monosaccharides, standard computational approaches like molecular dynamics are ineffective. Methods exist, such as umbrella sampling [175, 17] or metadynamics [100, 99, 14], that allow to accelerate the sampling of different conformations by adding bias forces. In general, accelerated sampling methods are based on the choice of a (usually low) number of collective variables (CVs). The choice of CVs is of particular importance for the proper reconstruction of free energy landscapes: their number should be small in order to speed up the sampling consistently, and they should represent every slow degree of freedom of the system. Otherwise, the estimate of the profile could be severely biased [99]. In the field of conformational analysis, suitable CVs have to be considered to describe non-planar, puckered conformations of cyclic structures. The problem of an intuitive description of puckered ring conformers dates back to the work of Sachse [154], but it is only in 1975 that an exhaustive, mathematically correct description of a general N-membered ring puckering has been proposed [41]. The description of Cremer and Pople has from time to time been questioned because of some conflicts with the standard stereochemical interpretation of puckering [191] and for some difficulties in comparing with NMR experimental results [19, 64]. Nevertheless, the approach of Cremer and Pople has become the most widely employed way to describe puckered conformers. The calculation of puckering free energy landscapes, and the determination of conformer populations has been the subject of a series of recent computer simulation investigations [24, 159, 66, 166, 13], with the aid of the aforementioned accelerated dynamics methods.

The research work presented in this thesis touches several of the problems just mentioned. Indeed, one of the main goal that we are going to present is the enhancement of the GROMOS force field for carbohydrates, with respect to the ability of the G45a4 parameter set [112] to describe ring conformation (that is, puckering) of six-membered rings. To this end, the development of efficient computational tools for the investigation of the general puckering problem, which are not limited to monosaccharides but extend to all system presenting cyclic

structures with non-planar conformations, are presented. In particular, we indicate how to exploit the capabilities of the metadynamics algorithm applied to the investigation of puckered ring conformers. The present work is organized as follows:

Chapter 1: the general classification and the structural features of monosaccharides are discussed. The theoretical and experimental frameworks relevant for conformational analysis studies are also presented in this Chapter;

Chapter 2: simulation techniques for free energy calculation are presented, with particular emphasis on metadynamics;

Chapter 3: the puckering problem is addressed from the point of view of the reduced coordinate sets necessary to describe it (Staruss-Pickett dihedrals and Cremer-Pople representations);

Chapter 4: the calculation of the free energy of simple systems is employed to show the general features of metadynamics in the constext of puckering free energy landscape reconstruction for six-membered rings;

Chapter 5: different parametrizations of puckered structures are explored to assess their respective advantages as collective variables for metadynamics;

Chapter 6: a systematic evaluation of the accuracy of the GROMOS 45a4 Force Field in describing ring puckering is presented, and a modification of the parametrization is proposed to obtain agreement with theoretical and experimental data;

Chapter 7: the achievements of this work are summarized and future perspective are discussed.

Chapter 1

Monosaccharides

That's the time you must keep on trying
Smile, what's the use of crying?
You'll find that life is still worthwhile
If you just smile

Charlie Chaplin
Smile

In this Chapter a general overview on carbohydrates is given, and a detailed description of monosaccharide structural complexity is addressed. Conformational analysis of monosaccharides is presented with respect to theoretical and experimental developments in the field. Eventually, some basic notions of how computational techniques can help in exploring the subject are given.

Contents

1.1 Carbohydrates

Biological molecules, or "biomolecules", are compounds mainly produced in living organisms and specifically involved in life processes. Biomolecules composition, structure, properties, reactions, and preparation are the subject of different scientific disciplines, like biochemistry (the study of the chemistry of life processes), organic chemistry (the study of carbon-based compounds), molecular biophysics (the study of biological systems with methods from physical science), molecular biology (the study of molecular basis of biological activity), and bioinformatics (application of informatics to biology), to name a few.

Life processes are mainly realized in cells as a result of the interplay of two main classes of molecules: a) large polymeric molecules (proteins, protein complexes, nucleic acids, ...) referred to as *biological macromolecules*, and b) low-molecular-weight molecules (glucose, glycerol, ...) referred to as *metabolites*. Biomolecules are present in both these classes, and also inorganic molecules participate to this interplay. Nevertheless, evolution has set specific characteristics on macromolecules in order to allow simple chemical reactions on metabolites to realize complex life processes and functions like movement, reproduction, sensitivity and growth.

The most abundant component in biomolecules is *carbon*, that with its electronic configuration can make four covalent bonds. In this way carbon is able to form a very large variety of compounds, the greatest variety with respect to any other chemical element. Moreover, the possibility of stable C−C covalent bonds (a peculiar characteristic[1] in nature) permits biochemistry to benefit of the main foundations of organic chemistry: repetitive carbon-carbon stable covalent bonds allows the formation of complex macromolecules with linear, branched, cyclic or cage-like structures. Furthermore, in biomolecules a large variety of functional groups (like alcohols, carboxylic acids, carboxamides, thiols, thiolethers, ...) is present, increasing dramatically the chemical reactivity and thus broadening the spectrum of biomolecules functions.

Understanding biomolecules capabilities is more complex than the same issue with inorganic molecules. This is because there is not only the problem of detecting their composition, but it is also important to know their *three-dimensional structure*. In inorganic chemistry, once a formula for a generic substance is given, the structure can often be inferred immediately. This is because most of the time there is only one compound with a given formula (*e.g.* Na_2SO_4). However, this is not the case with organic chemistry, and thus also with biochemistry, even with small molecules. As an example, ethyl alcohol and dimethil ether have the same formula (C_2H_6O) but they are very different molecules: at room temperature the former is liquid while the latter is a gas. Another example is the chemical formula $C_6H_{12}O_6$ that refers simultaneously to glucose and to other 31 compounds. These 32 molecules have physico-chemical properties substantially different (for example, the sweetness of the compound), and the differences only relies in the specific structure arrangement (in Section 1.2 this stereoisomery problem will be extensively discussed for monosaccharides). Furthermore, with

[1] Beside carbon, silicon is the only other element that can form strong repetitive covalent bonds. However, Si−Si structures oxidize rapidly into silica (SiO_2) in an oxygen-rich atmosphere. Thus, complex organized structures, like the one realized by carbon, are scarcely probable with silicon.

very large objects, like hemoglobin or chromosomes, the number of monomers is so high that the three-dimensional structure is extremely complicated. This is because the final structure is the result of many interactions (covalent and hydrogen bonds, Van der Waals interaction, dipolar interaction with solvents, ...) that simultaneously occur. Also, the specific order in which interactions are established is important, because the more a molecule is large and specific the more is important its *pathway of formation*. This very high level of structural complexity can be considered responsible for biomolecules ability in perform specific actions in life processes.

The standard subdivision of macromolecules identifies four main classes: proteins, nucleic acids, lipids and carbohydrates. *Proteins* are extremely versatile macromolecules, and this provide them to be ubiquitous in living organisms: they are found as catalysts (enzymes), transport and store vectors (hemoglobin), mechanical support materials (histones), movement mediators (myosin), to give only few examples. Proteins are linear polymers built on monomeric unit called *amino acids*. There are only 20 different amino acids in natural proteins. The amino acid sequence of a protein determines its properties and its three-dimensional structure: from the primary structure (the bare amino acid sequence) to the secondary structure (locally folded segments) and eventually to the tertiary and quaternary structure (the global, compact, asymmetric structure). *Nucleic acids* (like DNA and RNA) are the main macromolecules involved in the storage and transmission of genetic information. They are very long linear polymers, made from a limited set of monomers called *nucleotides*. DNA has most of the time the form of a double helix, where two strands pairs each other by means of hydrogen bonds. The base pairing mechanism allows the transcription of genetic information from DNA to RNA for the subsequent translation into proteins by ribosomes. *Lipids* are water-insoluble biomolecules that are highly soluble in organic solvents. The key constituents of lipids are *fatty acids*, namely hydrocarbon chains of various lengths. In combination with other molecules fatty acids span the variety of lipids: fats, oils and waxes, sterols, phospholipid and glycolipid, to name a few. This wide range of molecules encompass various biological functions. They can be highly concentrated energy storage molecules and are involved in signal transmission and transduction. Moreover, they are the major component of biological membranes.

Carbohydrates costitute most of the organic matter on Earth. This is because of their simple composition (as the name suggests, they are made basically by carbon, oxygen and hydrogen) and their extensive role in all form of life. They are energy stores, fuels and metabolic intermediate; they are found in the backbone of nucleic acids and in cell walls of bacteria and plants (the polysaccharide *cellulose* is one of the most abundant organic compound in the biosphere); they can interact and link with other macromolecules to form a wide range of *glycoconjugates* which are determinant for cell-cell communication and interaction between cells and other elements in their environment. In addition, carbohydrates have several industrial applications: starch in manifacture of goods and pastas; gums in food processing; mono and oligo-saccharides as sweeteners (sugars); cotton and linen in clothing fabrics; wood in furniture and wood pulp in paper industry; fermentors to make alcoholic beverages; antibiotics in pharmaceutical products.

This huge variety of functions for carbohydrates is possible due to their

structural diversity. Carbohydrates basic unit are *monosaccharides*, and will be extensively described in Section 1.2. Here we want to stress that their general empirical formula[2] $(C{-}H_2O)_n$ collects a quite large number of different *isomers* for a given n, due to the presence of several chiral carbon atoms. Moreover, the presence of multiple functional groups (aldehyde and ketone groups, and multiple hydroxyl side chains) allows a large number of possible pair linkages between monosaccharides (by means of *glycosidic bonds*). Thus, the sheer number of possible *oligosaccharides* depends not only by an higher number of monomers (different monosaccharides are larger in number than amino acids), but also by a more versatile connectivity to linear or branched polymers (*e.g.* given 4 monomers, many more different oligosaccharides can be formed from 4 monosaccharides than polypeptides from 4 amino acids).

Finally, the most popular carbohydrates are *polysaccharides*, namely, long homopolymers (or eteropolymers with few different monomers) where the same linkage scheme is repeated along very long chains. The presence of hydroxyl groups permit at this level the formation of fibers (like in cellulose) or secondary structures as helices (like in starch), different three-dimensional structures related only to the different linkages scheme in polymerization and not to a different composition.

All this intricacy and diversity, making carbohydrates information-rich molecules [98], has the biological role of successfully performing many life processes. For example, the diverse carbohydrate structures displayed on cell surfaces are well suited to serve as sites of specific interaction between cells and their environment (*e.g.* in blood group recognition).

1.2 Monosaccharides

Monosaccharides are the building blocks of carbohydrates. Hence, the general characteristics of the latter naturally depends on the structural properties of the former. With the term "structure" we may refer simultaneously to three aspects (see Chapters 2 and 3 of Ref. [146]):

(i) *constitution*, namely the nature of atoms and their type of bonds in the compound;

(ii) *configuration*, that describes the spatial arrangement of atoms and cannot be changed without breaking and reforming a chemical bond. In carbohydrate chemistry this refers naturally to the specific configuration at chiral carbons;

(iii) *conformation*, that deals again with spatial arrangement, but can be solely changed by means of rotation around chemical bonds.

Each of these aspects contributes to the complexity of carbohydrates in various ways. In principle this three-level description is applicable to monomers of all biomolecules. However, we will show that for monosaccharides this distinction is much more important than for other biomolecules.

[2] This formula justifies somehow the name of this class of molecules, because we have each "carbon" atom "hydrated" with water.

In the following a brief summary of nomenclature and classification of carbohydrates will be given with respect to the three field of description indicated above (for a much more complete classification, including monosaccharides derivatives, oligo- and poly-saccharides, see Refs. [119, 146]).

Constitution

Monosaccharides are aldehydes and ketones with multiple hydroxyl groups, with the general formula $C_n(H_2O)_n$. These functional groups are attached to each carbon of the unbranched skeleton of the acyclic molecule. The position of the carbonyl group (C=O) determines their classification as *aldoses* (the carbonyl group is at one end oh the chain) or *ketoses* (the carbonyl group is in any other position than the ends of the chain).

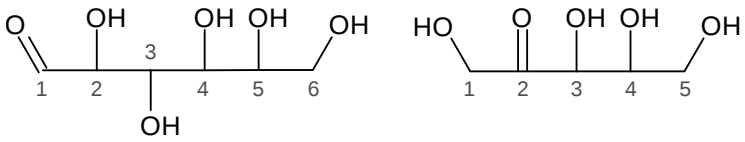


Figure 1.1: Example of constitution for simple monosaccharides. Rotated Fisher projection of glucose (an aldohexose, left) and ribulose (a ketopentose, right) are shown. Standard monosaccharides atom numbering scheme is presented, where the carbon of the carbonyl group gets the lowest possible number. Aliphatic hydrogens are omitted for simplicity.

Depending on the number of atoms, monosaccharides are further classified as *triose, tetrose, pentose, hexose*, etc. The combination of these two information (the number of carbon atoms and the position of the carbonyl group) simply classifies the constitution of monosaccharides. For example, in Fig. 1.1 two monosaccharides are shown along with their classification[3]: the aldo-hexose glucose (six carbon atoms and an aldehyde group) and the keto-pentose ribulose (five carbon atoms and a ketone group).

Configuration: the stereoisomery problem

As can be inferred from Fig. 1.1, monosaccharides have multiple asymmetric centers in carbon atoms with an hydroxyl group, due to the two possible absolute orientations of the $-$OH substituent. Different configurations at chiral carbons leads to different *stereoisomers*[4]. The actual number of stereocenters depends basically on three factors: (a) the total number of carbon atoms, (b) the position of the carbonyl group, and (c) the possible ring closure. In the *acyclic forms* of the $C_nH_{2n}O_n$ monosaccharide, chiral carbons are only non-terminal atoms with hydroxyl group ($n-2$ possible stereocenters). This automatically gives one stereocenter less for ketoses, which has a non-terminal carbonyl group.

[3] Hyphens, here used for clarity, could be omitted.

[4] Two molecules with the same chemical composition are *isomers*. There are two main forms for isomers: *structural isomers* with different order for atom bonds, *stereoisomers* (or spatial isomers, from Greek , solid, , equal, , part) with identical bond order but different spatial orientation of atoms.

Giving the atom numbering of Fig. 1.2, the highest-numbered asymmetric carbon is known as the *configurational carbon* and is useful for further classification. *Cyclic forms* appear when the carbonyl group reacts with an hydroxyl group to form an intramolecular hemiacetal or hemiketal for aldehydes or ketones, respectively. Due to ring closure, the carbon atom of the carbonyl group becomes a new asymmetric center, which is known as the *anomeric carbon*. Summarizing, there are 2^m stereoisomers with

$$m = n - 2 - j + k \ , \ j = \begin{cases} 0 & \text{for aldoses} \\ 1 & \text{for ketoses} \end{cases} , \ k = \begin{cases} 0 & \text{for acyclic form} \\ 1 & \text{for cyclic form} \end{cases} \tag{1.1}$$

(n is the total number of carbon atoms, see Fig. 1.2 for an explicit example).

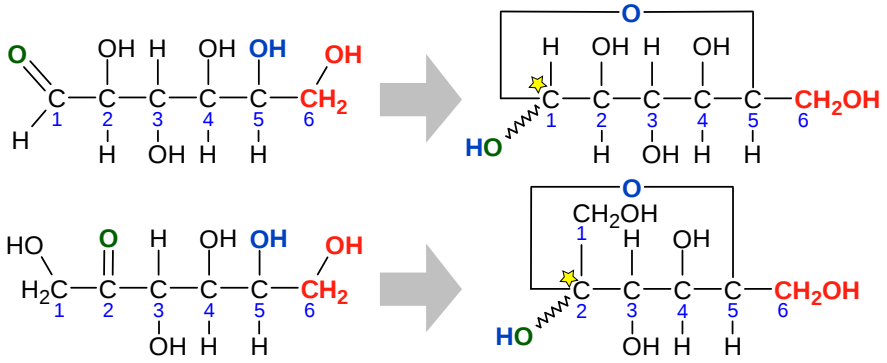


Figure 1.2: Chiral centers for monosaccharides. Rotated Fisher projection of glucose (an aldohexose, top) and fructose (a ketohexose, bottom) are presented to exemplify the counting of Eq. (1.1). In cyclic forms, the chirality at the anomeric carbons (starred) is left undefined (curly bond). Atoms involved in the intramolecular hemiacetal/hemiketal are colored (green and blue). The chirality at C5 (the configurational carbon) refers to the orientation of either the hydroxyl (blue) in acyclic chains or the hydroxymethil (red) in cyclic chains.

Besides, with the ring closure in principle an extra source of stereoisomery occurs, according to different ring closures. It is important because in fact the acyclic form is not the predominant one, especially in solution: monosaccharides with four or more carbon atoms are mostly present in cyclic forms. Two preferred ring structures are the *furanose* (five-membered, bottom of Fig. 1.2) and *pyranose* (six-membered, top of Fig. 1.2) ring[5]. Aldoses with four or more carbons and ketoses with five or more carbons can form furanoid rings, while aldoses with five or more carbons and ketoses with six or more carbons can form pyranoid rings. Therefore, in solution both anomeric and cyclic forms are expected: a small amount of acyclic forms in solution is the fingerprint of *mutarotation* (interconversion of anomers) and *tautomerization* (interconversion of ring forms), two equilibrium reactions that happen by means of opening and closing of rings. Although it is very difficult to separate in solution anomers and tautomers individually, their proportion can still be determined, for example with NMR spectroscopy [6, 146]. These measures, such as the extensive

[5] These names are related to the resemblance to furan and pyran molecules, respectively (see Fig. 1.3)

studies of Haworth and coworkers [70], show the preference in furanoyd rings for aldopentoses and in pyranoid rings for aldohexoses. In Table 1.1 and Fig. 1.3 some classification of mutual relationship between isomers, and relative group descriptors, are presented.

Table 1.1: Short glossary for isomers. The value m from Eq. (1.1) is the number of stereocenters. For a complete classification see Ref. [119].

Two isomers[§] are...	...if they are...	notes
enantiomers	mirror images	2^{m-1} possible pairs
diastereoisomers	not mirror images	D and L series[†] (two series with 2^{m-1} isomers)
epimers	only different at one stereocenter	indicated as CX epimers
anomers	only different at the anomeric carbon	α and β series[†] (two series with 2^{m-1} isomers)
tautomers	only different in ring form	*pyranose* and *furanose* sugars[‡] (see Fig. 1.3)

[§] the term "isomer(s)" here is used instead of stereoisomers for brevity
[†] descriptors D/L and α/β can be used as prefixes in combination with the constitution descriptors described in Section 1.2 (*e.g.* α-D-aldopentose, β-L-ketohexose)
[‡] suffixes "-furanose" and "-pyranose" are often used instead of number descriptors to lighten classifications when no ambiguity occurs (*e.g.* aldo-pyranose instead of aldo-hexo-pyranose)

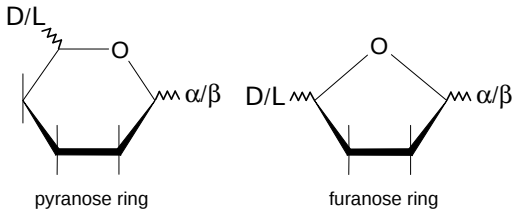


Figure 1.3: Pictorial view of stereoisomers. Simplified Haworth projection of the two main ring forms, with diastereoisomers and anomers descriptors. Absolute configuration at the configurational carbon (for D/L series) and relative configuration at the anomeric carbon (for α/β series) distinguish different sugar series.

Conformation: the puckering problem

Graphic representations like the Haworth projections of Fig. 1.3 are very helpful to show configuration at stereocenters, but somehow imply a planar ring shape. Conversely, as detected first by Sponsler and coworkers [167], cyclic forms are not stable as planar rings but assume *puckered conformations* (*i.e.* non-planar shapes). This happens for two main reasons: (a) atoms linked to carbons prefer a tetrahedral spatial distribution if possible, otherwise bond and angular strains occur; (b) ring substituents hinder each other if they are in close contact (*e.g.* $-$OH groups on the same side of the ring). Thus, steric and stereo-electronic

hindrance allows/imposes peculiar spatial arrangements to minimize, or if possible eliminate, strain effects:

- ring substituent orientations corresponding to (local) minima, with respect to steric hindrance, are preferred as (meta)stable *rotamers*;
- ring forms corresponding to (local) minima, with respect to chain strains, are preferred as (meta)stable *conformers*.

We want to stress again that interconversions between rotamers/conformers occur only by means of rotation around torsion angles, and this means by overcoming rotational/conformational free energy barriers, respectively.

The characterization of rotamers is quite simple, because steric hindrance prevents eclipsed orientation towards staggered one. Typically, three orientations are possible, the so-called *gauche+* , *gauche−* and *trans* (g^+, g^- and t, respectively) orientations, as shown in Fig. 1.4. It has to be mentioned, however,

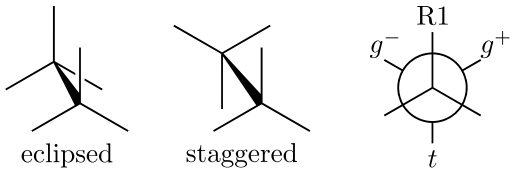
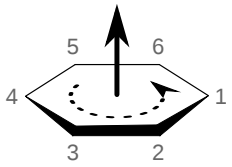


Figure 1.4: Rotamers definition. Eclipsed orientations (left) are local maxima of interaction energy, thus staggered orientation (center) occur. The Neumann projection (right) have exact tetrahedral substituents (apparent bond angles $\theta' = 120°$). Atom R1 is the reference for *gauche+*, *gauche−* and anti (or *trans*) orientations.

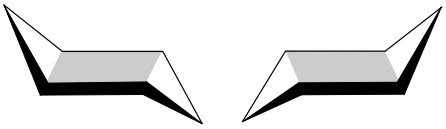
that strong rotameric preferences are generally present only for large exocyclic groups (*e.g.* the hydroxymethil group $-CH_2OH$ of glucopyranosides). Moreover, such torsional transitions are generally thermally activated, thus, very often, a distribution of rotamers is observed, namely the presence of more than one state in exchange equilirium.

Concerning conformers, it is possible to calculate theoretically a set of *ideal conformations* for a generic N-membered ring structure. In ideal structures, the spatial orientation for (almost) all substituent at every carbon atom follows an exact tetrahedral arrangement. Thus, bond and angular strains are theoretically minimized. For pyranose rings there are 38 ideal conformers (see Ref. [158, 47, 119]), divided into possible stable and strainless conformers (rigid *chairs*, flexible *boats* and *skew-boats*) and transition state conformers (*half-chairs*, *envelopes*). In Fig. 1.5 schematical representations and nomenclature of ideal conformers of aldohexoses in pyranose form (herafter referred as aldopyranoses) are shown. For five-membered rings there are 20 ideal conformers (see again Ref. [158, 47, 119]), shown in Fig. 1.6. There is less variability for furanoses because there are only two possible groups of ideal structures, divided into stable (*envelope*) and metastable (*twist*) forms. The interconversion between ring forms here is simpler, because both twist and envelope conformers are flexible.

We want to stress that ideal structures can represent only distinct families of pyranose ring conformations. Indeed, at a difference with the steroisomery

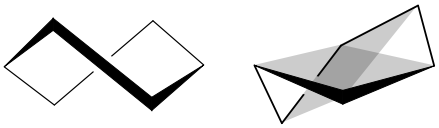


(a) Atom numbering for aldopyranoses. In the standard numeration atom 6 is the ring oxygen. The definition of the "up" side of the ring is showed.

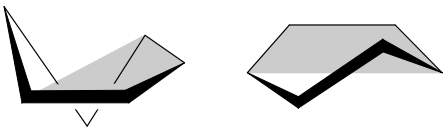


(b) Chair conformers. Atoms i and $i+3$ are at opposite side of ring reference plane. 2 conformers are the *chair* ${}^4C_1$ (right) and the *inverted chair* ${}^1C_4$ (left).

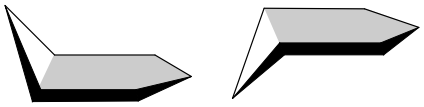
(c) Boat conformers. Atoms i and $i+3$ are on the same side of ring reference plane (6 conformers: ${}^{3O}B$, B_{14}, ${}^{25}B$, B_{3O}, ${}^{14}B$, B_{25}).



(d) Skew-boat conformers. Atoms i and $i+2$ are at opposite side of ring reference plane (6 conformers: ${}^3S_1$, ${}^5S_1$, ${}^2S_O$, ${}^1S_3$, ${}^1S_5$, ${}^OS_2$).

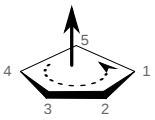


(e) Half-chair conformers. Atoms i and $i+1$ are at opposite side of ring reference plane (12 conformers: ${}^OH_1$, ${}^2H_1$, ${}^2H_3$, ${}^4H_3$, ${}^4H_5$, ${}^OH_5$; ${}^3H_4$, ${}^5H_4$, ${}^5H_O$, ${}^1H_O$, ${}^1H_2$, ${}^3H_2$).

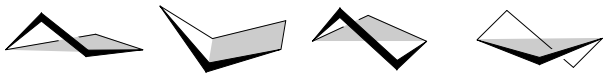


(f) Envelope conformers. Atom i is out-of-plane (12 conformers: ${}^OE$, ${}^1E$, ${}^2E$, ${}^3E$, ${}^4E$, ${}^5E$; ${}^3E$, E_4, ${}^5E$, E_O, ${}^1E$, E_2).

Figure 1.5: Schematic form for aldopyranose ideal conformers. Names of the 38 ideal structure are listed according to IUPAC recommendation [47]. Reference ring planes are showed as shaded regions. Alternative names could sometimes occur, due to different use of reference plane atoms and out-of-plane atom descriptors (*e.g.* ${}^2C_5 \equiv {}^4C_1$, ${}^3S_5 \equiv {}^OS_2$). Conformer descriptors can be used as suffixes to monosaccharide name (*e.g.* β-D-aldo-pyranose-${}^4C_1$).



(a) Atom numbering for aldofuranoses. In the standard numeration atom 5 is the ring oxygen. The definition of the "up" side of the ring is showed.



(b) Envelope conformers. Atom i is out of ring reference plane (10 conformers: ${}^OE$, ${}^1E$, ${}^2E$, ${}^3E$, ${}^4E$, E_O, ${}^1E$, E_2, ${}^3E$, E_4).

(c) Twist conformers. Atoms i and $i+1$ are at opposite side of ring reference plane (10 conformers: ${}^OT_1$, ${}^2T_1$, ${}^2T_3$, ${}^4T_3$, ${}^4T_O$, ${}^1T_O$, ${}^1T_2$, ${}^3T_2$, ${}^3T_4$, ${}^OT_4$).

Figure 1.6: Schematic form for aldofuranose ideal conformers. Names of the 20 ideal structure are listed according to IUPAC recommendation [47]. Reference ring planes are showed as shaded regions. Conformer descriptors can be used as suffixes to monosaccharide name (*e.g.* β-D-aldo-furanose-${}^OE$).

definitions already discussed, ring conformations are influenced by the dynamic equilibrium with the surrounding solvent. Thus, sugar rings oscillate between distorted forms around an ideal conformer, by means of small changes in molecular geometry. Transitions between conformer families need the crossing of free energy barriers, with transition state intermediates (like half-chairs and envelopes for six-membered rings).

The effective conformer stability depends not only in the ring shape but also in *ring substituents orientation*. Indeed, conformations like the one showed in Fig. 1.5 for aldopyranoses are suitable also for simpler six-membered ring without eteroatoms, like cyclohexane (C_6H_{12}). However, for cyclohexane all ring substituents are equivalent (all H atoms), while for monosaccharides this is not the case (they can be either $-H$, $-OH$ or $-CH_2OH$ groups for non-substituted monosaccharides). If exocyclic substituents are not geometrically or chemically equivalent, then their orientations play a role in the relative (in)stability of ring conformers. For ideal structures, two groups of orientation are possible: *axial* (nearly perpendicular to the average plane of the ring) and *equatorial* (nearly parallel to the average plane of the ring) substituents[6]. The actual position of an exocyclic group depends on the specific stereoisomer (for examples of this correspondence see Figs. 1.7 and 1.8), thus different stereoisomers exhibit different preferences in ring conformers. In Fig. 1.7 it can be seen clearly that the source of these conformational preferences are the possible sterical hindrance that occur if ring substituents emerge on the same side of the ring (*e.g.* the so-called 1,3-diaxial groups). However, also solvation effects are important at this stage, because hydroxyl group can form hydrogen bonds and so the weak interaction with the solvent can change the intrinsic stability of the conformer.

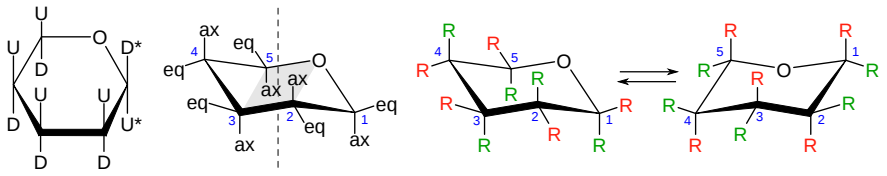


Figure 1.7: Ring substituents orientation. Haworth projection (configuration at chiral carbons) and Haworth conformational formulae (chair conformer) are shown. Except for the anomeric carbon, U is for up and D is for down residues with respect to the ring plane in Haworth projection. In the conformational view, the indicated mean ring axis gives the notion of axial (ax) and equatorial (eq) substituents in the chair form. Correspondences [U,D*] $\leftrightarrow$ ax, [D,U*] $\leftrightarrow$ eq between the representation hold (taken from Tab. 1.7 in [146]). When interchange of mirror image conformers occurs (like the showed chair and inverted chair), the position of substituents changes from axial to equatorial and vice versa.

The conformational variability described so far is quite different from the configurational one. The latter is in some sense a static property, related only on

[6] This axial/equatorial definition with parallel/perpendicular orientation of substituents is strictly speaking only valid for chairs conformer. For a detailed discussion on the definition of ring substituents see Ref. [39].

the specific characteristic of the composition and on stereoisomery. The former is much more a dynamic aspect, because concerns not only the structure itself but also the intra/inter-molecular interactions (substituent hyndrance, equilibrium with solvents, ...). Indeed, while we can keep the configuration fixed, conformations can only be regarded as average forms, namely percentages of preferred conformation.

From monosaccharides to carbohydrates

Monosaccharides derivatives, oligo- and poly-saccharides increase the complexity already explored within monosaccharides by means of a great variability in modification and polymerization schemes.

When one or more hydroxyl groups are substituted, monosaccharides derivatives are formed. The presence of different groups yields to different active molecules, that are important both in biological activities and in industrial applications. To give only few examples, deoxyribose (2-deoxy-D-ribose) is part of the backbone of DNA; fucose (6-deoxygalactose) occurs in some glycoproteins and glycolipids; ascorbic acid (vitamin C) is a biological oxidation-reduction agent; GlcN and GalN (2-deoxy-2-amino-glucoseamine and 2-deoxy-2amino-galactoseamine, respectively) are constituents of a class of antibiotics; GlcNAc, GalNAc (N-acetyl-glucoseamine and N-acetyl-galactoseamine, respectively) and sialic acids serves as recognition markers in glycoconjugates. As it can be seen, monosaccharide derivatives further expand the number of possible monomers for complex structures.

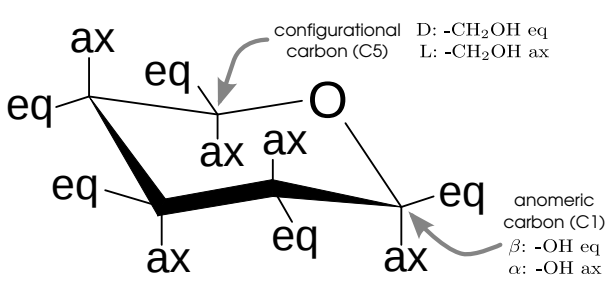
When the anomeric carbon reacts, glycosides and di- and ologi-saccharides occur with the formation of *glycosidic bonds*. The number of possible connection schemes is quite large due to the presence of a great number of hydroxyl groups and of their specific axial or equatorial orientations. When the degree of polymerization increases, the number of possible linkages grows dramatically, with linear or branched connections. To give an example, 20 different monosaccharides (like glucose, mannose, fructose,...) gives $\sim 9 \times 10^6$ trisaccharides, while with 20 amino acids only $\sim 8 \times 10^3$ tripeptides are in principle possible (for similar calculation see Ref. [98]).

However, unlike proteins, large polysaccharides may present a very simple primary structure: cellulose and starch are homopolymers made of glucose; heparin and hyaluronic acid (two example of glycoseamineglycans) are long polymers made by a repeated disaccharide. In the case of cellulose and starch, in addition, their difference is only in the polymerization scheme, that changes dramatically their macroscopic behavior, although their constitution is roughly the same. In order to find, as in proteins, highly specialized sequence made of different monosaccharide sub-unit, we have to seek for small oligosaccharides linked to proteins or to lipids. A famous example is the AB0 blood group system (based on antigens on cell surface that differ in few monosaccharides between A and B type), that is an example of the signaling attitude of saccharides in conjunction with carbohydrate-binding protein known as lectines (see Ref. [161] for an historical review on the importance of carbohydrate specificity in cell signaling).

1.3 Conformational analysis of aldopyranoses

The story of conformational analysis of six-membered rings dates back to 1890. In the work of Sachse and coworkers [154, 155] puckered forms like rigid chairs and flexible boats for cyclohexane structure were proposed. In 1926, Sponsler and Dore [167], recognized (for the first time) the sole chair structure in glucopyranose, using this puckered form to interpret X-ray data on cellulose. The work of Haworth [70] on the contrary, pointed out the possible presence of two kind of stable strainless structures in pyranoses (2 chairs and 6 boats, which will be from now denoted as "conformers"). The importance of conformations was stressed further on when major effects in reactivity, recognized in Isbell studies of oxidation of free sugars by bromine [79, 80, 81], were attributed to structural[7] differences. By these studies, the presence of two different anomers (denoted as α and β), and the nature of their difference in exocyclic group orientation, was proposed.

Starting from these findings, a number of theoretical and experimental investigations have been devised, both in solution and in solid state, to describe conformation of monosaccharides. Theoretical interpretation of conformer stability focused firstly on steric relations between exocyclic groups, and this is understandable if we consider, for example, the orientation of ring substituents of aldopyranoses, showed syntetically in Fig. 1.8. Since each stereoisomer has its



Stereoisomers	-OH group	C2	C3	C4
	Glc	eq	eq	eq
	Gal	eq	eq	ax
	Man	ax	eq	eq
	All	eq	ax	eq
	Tal	ax	eq	ax
	Gul	eq	ax	ax
	Alt	ax	ax	eq
	Ido	ax	ax	ax

Figure 1.8: Possible aldopyranoses. A total number of $2^5 = 32$ stereoisomers can be constructed for $C_6H_{12}O_6$, chosing the chirality at each stereocenter. The $^4C_1$ conformation (left) and common names for different stereoisomers (right) are shown.

own pattern of substituents orientation, the known differences in conformational preferences could be related to substituents interactions. To quantify this effect, the main idea was to assign, in the chair conformer for pyranosides, instability factors to mutual substituent specific orientations. Various authors proposed different contributions to instability:

- the Hassel-Ottar (H-O) effect [69] accounts the instability driven by hydroxyl and hydroxymethil groups in axial orientation (an example of 1,3 diaxial interaction);
- the $\Delta 2$ effect [147, 148] accounts the instability from an axial group at C2

[7] Hereafter, the term "structure" will be used as a synonym of the proper term "configuration", unless specifically indicated.

and an equatorial group at C1;

- generic axial hydroxyl groups (or axial hydroxymethil group not involved in the H-O effect) even without any interaction with other substituents accounts a slight instability [88];
- simple *syn*-diaxial (1,3-diaxial) effect accounts the instability due to the interaction of a pair of hydroxyl groups in axial orientation;
- the *anomeric effect* accounts, in the chair conformation, the axial preference of the hydroxyl group at the anomeric carbon C1 due to the interaction with the ring oxygen O5 and the substituent at vicinal carbon C2 (first discussed by Edward [48], it was named the anomeric effect by Lemieux and Chu [106] and can include partially the $\Delta 2$ effect).

All these effects were most of the time considered to be additive, and various combinations of instability factors were included in different theoretical schemes [69, 147, 148, 88]. However, these first attemps were designed to give only conformational preferences (*e.g.* ${}^4C_1$, ${}^1C_4$ or ${}^4C_1 \Leftrightarrow {}^1C_4$ situations), because arbitary "units of instability" were used for the aforementioned effects.

A significant advance in the field was the semi-empirical study of Angyal [5]. In contrast with previous studies, Angyal assigned energy values in kcal/mol (and not in arbitrary units) to instability factors in order to collect free energy estimates for all the hexo- and pento-pyranoses. He derived free energy values mainly from the equilibrium of cyclitols with their borate complexes (for generic instability effects) and from equilibrium data of D-glucopyranose and D-mannopyranose in acqueus solution (for the anomeric effect). In the same direction, a different approach explored for the first time by Rao and coworkers, was the use of classical potential functions. Their first estimates of non-bonded interaction energies was done using ideal geometries [173]. Further refinement were subsequently done by minimizing the conformational energy with slight tilting of ring substituents [145], and using a more general procedure of conformational search using the Strauss-Pickett spherical coordinates[8] to explore all possible pyranose conformations [84].

To the best of our knowledge, the theoretical free energy estimates for aldohexopyranoses of Angyal [5] and of Vijayalakshmi and Rao [183] (both reported in Table 1.2) are the only theoretical quantitative free energy estimates at disposal. Their data are in very good agreement with each other, even considering the approximations used in both schemes. Moreover, these proposed estimates also somehow agree with experimental findings that will be discussed in the next Section.

1.3.1 Experimental measurements of puckering properties

To measure ring puckering, experimental techniques sensible to ring shapes and/or ring substituents orientation are needed.

An example is *complex formation*, like the reaction of copper-ammonia complex (cuprammonium) with hydroxyl group. Complexes can be formed only if

[8] See Section 3.1 for further description of puckering coordinates.

Table 1.2: Calculated free energies for D-aldohexopyranoses. Data taken from Tab. 3.5 in [146] (Original data from Angyal [5] and Vijayalakshmi and Rao [183]). Energies are converted from the original units (kcal/mol) to kJ/mol , to simplify comparisons with data in this work.

	Glc	Gal	Man	All	Tal	Gul	Alt	Ido
				β-Anomer				
Angyal	24.89	21.97	19.66	12.97	16.74	10.04	8.37	5.44
Vijayalakshmi and Rao	24.98	15.48	16.19	15.48	13.64	11.97	8.87	3.77
				α-Anomer				
Angyal	17.36	14.43	12.76	6.07	9.83	3.14	0.84	−2.09
Vijayalakshmi and Rao	18.62	10.96	10.46	11.46	8.12	5.94	4.44	−0.08

† energy difference between $^1C_4$ and $^4C_1$ conformer for each D-aldohexopyranoside

hydroxyl groups are suitably located in the compound. Since the substituents orientation differs between conformers (in the case of chairs, the orientation differs drastically with ax↔eq exchange), this method was used by Reeves and coworkers to determine relative orientations of neighboring hydroxyls [148, 149]. It should be noted that this method is suitable for systems with a defined preferred conformation, but it is unable to give good results in cases of conformational equilibrium (*e.g.* $^4C_1 \Leftrightarrow {}^1C_4$ exchange equilibrium).

A powerful technique for conformational analysis is the *Nuclear Magnetic Resonance* Spectroscopy (NMR). NMR is generally known to be very useful for detecting the structure of biopolymers in solution. Its reliability for conformational analysis of carbohydrates was firstly indicated by Lemieux and coworkers [106]. They showed that differences in proton signals occur when specific orientations of ring substituents are present. In particular, vicinal coupling constants $^3J_{HH}$, that are a measure of the spin-spin coupling between ^{1}H atoms separated by three bonds, appears to be $2 \div 3$ time higher for ax–ax oriented residues than for ax–eq or eq–eq substituents. This peculiarity greatly helps the complete atom assignment for NMR spectra of monosaccharides. In fact, NMR spectroscopy was able to distinguish α/β and pyranose/furanose dynamic equilibrium (see Table 1.4 in Ref. [146]), stating the preference for pyranose form for aldohexoses. Moreover, with Karplus-type relationships [85, 86, 63, 4] it is possible to connect J-coupling data with the values ϕ of torsional angles around bonds connecting the atoms. These relationships permit on one hand (ϕ values from J ones) to have structural information on measured (mean) structures; on the other hand (J values from ϕ ones) expected coupling constants of specific ring arrangements can be calculated and compared with experimental data.

An illustrative example of NMR data and their interpretation is in Table 1.3. The observed coupling constants are sensible of the preferred conformational of the compounds. Thus, these data encodes the percentage of conformer populations at thermodynamic equilibrium. For example, with the assumption of a purely two-state system, two possible conformer X=A,B can occur. Thus the J-

coupling signal will be the mean between the signals J^{X} of a purely single-state system. The simple relation

$$ {}^3J_{HH} = P_{\mathsf{A}}\,{}^3J^{\mathsf{A}}_{HH} + P_{\mathsf{B}}\,{}^3J^{\mathsf{B}}_{HH} \quad , \quad P_{\mathsf{A}} + P_{\mathsf{B}} = 1 \tag{1.2} $$

estimates the molar fraction P_{X} (*i.e.*, the conformer population) from a single coupling constant[9], provided the limiting coupling constant J^{X} of a system with the single conformation X (population presented in Table 1.3 are calculated in this way). Eventually, by Boltzmann inversion we have

$$ P_{\mathsf{X}} = \mathrm{e}^{-\beta F_{\mathsf{X}}} \quad \Rightarrow \quad F_{\mathsf{A}} - F_{\mathsf{B}} = \Delta F_{\mathsf{AB}} = -kT \ln \frac{P_{\mathsf{A}}}{P_{\mathsf{B}}} \tag{1.4} $$

and free energy differences between two conformers A and B can be evaluated. In Table 1.3 this simple evaluation is reported. Eq. (1.3) was used against the

Table 1.3: NMR coupling constants. Vicinal coupling constants for the complete D-aldopyranoses are given. In addition, populations of $^4C_1$ conformer are compared to preferred conformation taken from Ref. [6]. Values $P[^4C_1]$ are calculated from Eq. (1.2), using $J^{^4C_1} = 9.8\,\mathrm{Hz}$ and $J^{^1C_4} = 3.6\,\mathrm{Hz}$ as limit values. $P[^1C_4] = 100\% - P[^4C_1]$. Population $P \sim 5\%$ are within the errors in NMR.

	$^3J_{H_1,H_2}$	$^3J_{H_2,H_3}$	$^3J_{H_3,H_4}$	$^3J_{H_4,H_5}$	$P[^4C_1]$	
β-Anomer						
Glc	7.8	9.5	9.5	9.5	95.2 %	$^4C_1$
Gal	8.0	10.0	3.8	1.0	∼100.0 %	$^4C_1$
Man	1.5	3.8	10.0	9.8	∼100.0 %	$^4C_1$
All	8.5	3.3	3.2	9.5	95.2 %	$^4C_1$
Tal	1.2	3.2	3.2	1.2	-	$^4C_1$
Gul	8.3	3.6	3.6	0.8	∼100.0 %	$^4C_1$
Alt	1.4	4.1	2.2	9.1	90.3 %	$^4C_1$
Ido	1.6	3.8	3.7	1.8	97.6 %	$^4C_1$
α-Anomer						
Glc	3.6	9.5	9.5	9.5	95.2 %	$^4C_1$
Gal	3.8	10.0	3.8	1.0	∼100.0 %	$^4C_1$
Man	1.8	3.8	10.0	9.8	∼100.0 %	$^4C_1$
All	4.0	-	-	-	-	$^4C_1$
Tal	1.9	3.2	3.2	1.3	-	$^4C_1$
Gul	≈4.0	-	-	-	-	$^4C_1$
Alt	3.4	5.6	3.7	7.6	66.1 %	$^4C_1 \Leftrightarrow {}^1C_4$
Ido	6.0	8.1	7.9	5.0	29.0 %	$^4C_1 \Leftrightarrow {}^1C_4$

Data are taken from Tab. 3.6 in [146] (see also reference within), except data for Altrose that are taken from Autieri et al. [10]. $^3J_{H,H}$ values are in Hz, $P[^4C_1]$ values are in %.

[9] For a three-state system, the relation between coupling constants and populations reads

$$ {}^3J_{HH} = P_{\mathsf{A}}\,{}^3J^{\mathsf{A}}_{HH} + P_{\mathsf{B}}\,{}^3J^{\mathsf{B}}_{HH} + P_{\mathsf{C}}\,{}^3J^{\mathsf{C}}_{HH} \quad , \quad P_{\mathsf{A}} + P_{\mathsf{B}} + P_{\mathsf{C}} = 1 \tag{1.3} $$

and thus two coupling constants are needed to obtain all the populations. For more states the number of required constants grows accordingly.

$^3J_{H,H}$ values that are ax–ax or eq–eq in the $^4C_1$ conformer according to Fig. 1.8. As it can be seen, the available estimated populations are in agreement with the theoretical estimation reported in Table 1.2. Thus, coupling constants proved to be sensitive of configuration, conformation and nature of substituents of the molecule (see for example Ref. [160]).

Since NMR gives important information on properties in solution, from this technique only average properties on conformations in dynamic equilibrium with solvent are available. On the contrary, *X-ray crystallography* can in principle give the atomic spatial arrangement of a single static structure. However, this technique needs samples with well-formed crystal structures. In general, given an aqueous solution only one of the four possible isomers (α/β anomer and pyranose/furanose form) crystallizes, but the final conformation is not necessarily the predominant one, especially for compounds which do not exhibit conformational preferences. For most of the aldopyranoses, the crystal structure data available confirm the $^4C_1$ preference, or a $^4C_1 \Leftrightarrow {}^1C_4$ equilibrium, in the same direction of data obtained with other experimental techniques. Crystal structure data are anyhow important, because they provide standard dimensions for pyranoses (bond lengths, bond angles, substituents angles) for modeling higher carbohydrates.

The techniques shown above seem to be able to distinguish only between the chair and the inverted chair structures. As discussed in Section 1.2, this is understandable because chairs are rigid structures while boats and skews are flexible ones. Thus, flexible structures very rapidly exchange in solution and very hardly crystallize. However, circumstantial experimental evidences (coupling constant that suggests three-state systems with intermediate skews [108, 78, 128], or high-resolution 3D crystallographic structures on monosaccharide derivative (GlcNAc) that present occurrences of flexible conformers [35, 77, 118, 188]) show that for some monosaccharides also flexible conformers can play a role. In this perspective, experimental techniques like *Atomic Force Microscope* spectroscopy (AFM), able to detect to some extent flexible conformers, are of great interest. Marszalek and co-workers used firstly AFM pulling technique to estimate glucose boats free energy on carboxymethil amilose [116, 107] and then to measure glucose skew-boats in dextran, cellulose and postulan [103, 115, 192]. Differently from NMR, here polymer structures were used, and the elongation process can produce or not non-chair conformers along the polymer in direct relation with the linkage scheme of the polymer itself. Thus, by subtracting free energy differences of different processes, the free energy contribution of the sole conformational transition can be evaluate (*e.g.*, skew-boat free energy was estimated to about 25 kJ/mol over the chair conformer [192]).

1.3.2 Towards a microscopic description: computer simulations

The main information that can be gathered from experiments and theoretical analysis concerns the stable, rigid chair forms. If one is interested in investigating in more details the role of other conformations, a computational approach is certainly necessary. Simulation methods have became very valuable tools to complement experimental findings in systems of biochemical relevance [87], and could be loosely classified in the two groups of *ab initio* and empirical (force

field based) methods.

Ab Initio methods

Ab Initio quantum chemical methods attempt to solve Schrödinger equation, taking into account electron densities with the aim of obtaining information on the electron structures of molecules, liquids, or solids, typically under the Born-Oppenheimer approximation. Molecular orbitals are considered as linear combination of a (limited set of) atomic orbitals. Since *ab initio* methods are based on first principles, they can be in general applied for any kind of system. However, for even relatively small systems like sugars, a very large number of integrals have to be evaluated, to take into account the interaction between electrons. Hence, the computational cost of such calculations is astronomical. For these reasons, *ab initio* predictions on puckered structures of monosaccharides are relatively scarce.

Some reports in the field [184, 185, 58, 59, 144] show simple calculations on small model systems for the purpose of geometry optimization, torsional barrier estimation and evaluation of the gauche and the (exo)anomeric effect. Even with the growing computational power of recent times, we are aware only of few works for a complete conformational analysis of monosaccharides: a Car-Parrinello metadynamics[10] of glucose in vacuum [24] (unfortunately, non-optimal collective variables were used in this study [159]); a Møller-Plesset perturbation approach with inclusion of solvation free energies contributions [16]; a density functional theory calculation at the B3LYP/6-311++G** level [7]; a systematic investigation at the B3LYP/6-311++G** level of all epimers of glucose, but only for potential energies [157].

Unfortunately, it is not clear how compare different *ab initio* predictions on monosaccharides (consider, for example, the case of the very different estimates of Appell (29.18 kJ/mol) and of Barrows (57 kJ/mol) for the free energy of the β-D-glucose-$^1C_4$ conformer). Since, to the best of our knowledge, no attempt was done in calculations on benchmark systems, like D-altrose and D-idose, in our opinion is difficult to assess the confidence of these estimates.

Empirical (force field based) methods

Empirical methods aim is to solve Newton equations of motion with a empirical potential energy function to represent molecular energy terms. The kind of potential energy terms, and their parameters, are globally called, the *force field* (FF). Force fields are generally assumed to be additive[11] and transferable (namely, the same atom types interact roughly in the same way for sufficiently similar molecules). Such approaches have a computational cost sensibly lower than *ab initio* calculation, giving access to larger systems (*e.g.* fully hydrated with explicit water simulations of monosaccharides are available, rather than calculation of a single molecule in gas phase).

A FF has intrinsic approximations, because functional forms of potential

[10] Metadynamics is an accelerated dynamics method that will be discussed extensively in Chapter 2.

[11] For the so-called Class I force field, at a difference with Class II force fields that contains in principle cross term interactions and polarizable electrostatics, see Ref. [113].

energy terms only effectively reproduce real interactions, and typically molecule topologies are fixed, thus no chemical reaction can be studied directly. In addition, the reliability of a FF calculation is limited by the way the FF itself is built. However, within these limitations the success of these methods is provided by their great accuracy, as they are able to reproduce several experimental accessible properties within measurement sensibility. This ability also benefits from constant algorithmic improvement and growing computational power. To push further the prediction accuracy, it is clear that the actual parametrization of each force field is subject to continuous refinement. Indeed, all new accessible data, including structural and thermodynamical properties, and also quantum chemical indication concerning electrostatic potentials of specific functional groups, permits the extension of FF range of reliability.

The fortune of force field approaches started with protein and nucleic acid systems. With growing interest for protein-glycan complexes, various extensions of FF parametrization to include also saccharide elements (for example in GROMOS [112, 133, 66], CHARMM [61, 62], AMBER-GLYCAM [91] and OPLS [43] force fields) were proposed. The presence of more than one force field reflects different choices in functional forms and parametrization strategies (comparisons between force field capabilities on carbohydrate description are present in literature, see for example [140, 170, 166, 57]). A complete overview of available force fields for carbohydrate simulations is beyond the scope of this work (a valuable review, even if not up-to-date, can be found in [113]). However, it is noteworthy to stress that computer simulation of carbohydrates with empirical force fields is currently a very rich research field. Indeed, during the last two decades, several authors [96, 36, 165, 112, 109] stressed that wrong percentage of puckered conformation can occur in molecular dynamics simulations. This possibility affects negatively the accuracy of force fields calculations, since structural and dynamics properties could be reproduced erroneously [96]. Thanks to these findings, the necessity of considering also the ability of force field to reproduce conformational features was claimed. This is why, in recent times, the need of accurate conformational description of monosaccharides is becoming an important aspect within the parametrization of force fields.

Chapter 2

Simulation Techniques

> But note that the computer as such offers us no understanding, only numbers.
>
> D.Frenkel & B.Smit
> Understanding Molecular Simulation

In this Chapter the connection between simulation and statistical mechanics is explored. The aim is to introduce free energy methods suitable for conformational analysis. Since most of the time systems of interest exhibit metastabilities, the recent metadynamics algorithm is presented as a tool to perform free energy reconstructions in the presence of rare events.

Contents

2.1 Simulations and statistical mechanics

Computer simulations are nowadays a very used and useful tool for scientific research. The advantages of the so-called "computer experiments" are numerous: it is possible to take "measurements" in extreme conditions (*e.g.* systems at very high temperature, pressure and/or density); properties of materials that have not yet been made can be predicted; simulations could be compared with experiments (and thus the computer model could be validated) and with theories (providing tools specifically designed to test theories). By and large, computer investigations are nowadays an essential link between experimental and theoretical work.

Computer simulations of many-particle systems have an intimate connection with *statistical mechanics*. In simulations, indeed, we have at disposal microscopical properties (*e.g.* atomic positions and velocities), and the equivalent of experimental measurements are represented by statistical averages over microscopical configurations. Two crucial points in computer simulations are:

1. the identification (and, then, the proper simulation) of the appropriate *statistical ensemble*;
2. the issue of *ergodicity*, *i.e.* whether a simulation is able to collect a suitable statistics for the desired observable.

Concerning statistical (thermodynamic) ensembles, a simple way to describe them is as prescriptions that induce specific "counting schemes" in the *phase space*

$$\left.\begin{array}{r} r \equiv (\mathbf{r}_1, \ldots, \mathbf{r}_N) \in \mathcal{D} \subset \mathbb{R}^{3N} \\ p \equiv (\mathbf{p}_1, \ldots, \mathbf{p}_N) \in \mathbb{R}^{3N} \end{array}\right\} \Rightarrow (r,p) \in \mathcal{D} \times \mathbb{R}^{3N} \tag{2.1}$$

of the N-particle system. In other words, physical characteristics of a system are translated in weighting prescriptions for the points (q,p) of the phase space $\mathcal{D} \times \mathbb{R}^{3N}$. Within a statistical ensemble, the calculation of the mean value $\langle \cdot \rangle_\mu$ of an observable $A(r,p)$ is espressed as the statistical average

$$\langle A(r,p) \rangle_\mu = \int\limits_{\mathcal{D}\times\mathbb{R}^{3N}} A(r,p)\mu(\mathrm{d}r\mathrm{d}p) \equiv \mathbb{E}_\mu[A] \tag{2.2}$$

(sometimes indicated as the expectation value $\mathbb{E}_\mu[\cdot]$ of A) where

$$\mu(\mathrm{d}r\mathrm{d}p) = \rho(r,p)\mathrm{d}r\mathrm{d}p \quad , \quad \text{with} \int\limits_{\mathcal{D}\times\mathbb{R}^{3N}} \mu(\mathrm{d}r\mathrm{d}p) = 1 \quad , \tag{2.3}$$

is a probability measure, sometimes indicated by its probability density $\rho(r,p)$. The integral showed in Eq. (2.3) indicates that $\mu(\mathrm{d}r\mathrm{d}p)$ is normalized to 1 on the whole phase space $\mathcal{D} \times \mathbb{R}^{3N}$.

Let's spend a few words on the physics behind this averaging procedure. The points (r,p) in the phase space $\mathcal{D} \times \mathbb{R}^{3N}$ represent the accessible *microstates* for the system. The measure $\mu(\mathrm{d}r\mathrm{d}p)$ represents relevant *macrostates*, the states in thermodynamic sense that the system could exhibit. The practical computation of ensemble averages requires to sample a set $\{(r^n, p^n)\}_{n=1,\ldots,M}$ of configurations

from the probability measure $\mu(\mathrm{d}r\mathrm{d}p)$, that is to sample configurations in the phase space with the correct probability provided by the physics of the system. After the sampling, the approximation

$$\langle A(r,p)\rangle_\mu \simeq \frac{1}{M}\sum_{n=1}^{M} A(r^n,p^n) \tag{2.4}$$

estimates the mean value of a given observable A.

Once the correct thermodynamic ensemble is identified, a simulation able to perform the requested sampling has to be addressed. The total duration[1] τ of the sampling, that is the number $M = \tau/\Delta t$ of sampled points at Δt distance, is connected to the second issue previously indicated: the *ergodicity* condition is fulfilled when almost all the phase space is sampled, according to the probability measure $\mu(\mathrm{d}r\mathrm{d}p)$, within the interval τ. Looking to Eq. (2.4) ergodicity is not obvious, because the only clear information is that the larger is the number M of sampled points, the longer is the "trajectory" $\{(r^n,p^n)\}_{n=1,\dots,M}$ in the phase space. The approximation of Eq. (2.4) resembles indeed a *time* average

$$\langle A(t)\rangle_\tau = \frac{1}{\tau}\int_0^\tau A(r(t),p(t))\mathrm{d}t \tag{2.5}$$

as if we are following the solution $\{(r(t),p(t))\}$ of some kind of equations of motion in the phase space. The *ergodic hypothesis*

$$\boxed{\lim_{\tau\to+\infty}\langle A(t)\rangle_\tau = \langle A(r,p)\rangle_\mu} \tag{2.6}$$

assumes that the time averages of Eq. (2.5) on very long trajectories and the ensemble averages of Eq. (2.2) converges (at least in the thermodynamic limit). To make use of this hypothesis, the interval τ has then to be sufficiently long to ensure that almost all the phase space is explored. In other words, to have a proper sampling we need to ensure that the system could explore in principle all the (relevant) regions of the phase space.

It should be stressed that, although in most cases a large number M of points is sufficient to make this hypothesis work, there are also situations in which ergodicity is questionable. We refer to systems that are non ergodic *in practice* (*e.g.* glasses, meta-stable phases, ...) or even *in principle* (*e.g.* nearly harmonic solids).

2.1.1 Molecular Dynamics

Molecular Dynamics (MD) is a quite old simulation technique. Its capabilities had evolved a lot, spanning a wide range of applications, from the first reported simulation on hard sphere packing (in 1956 by Alder and Wainwright [25] at Livermore) to a recent application on protein folding in explicit water (in 2010 by Shaw et al. [162]).

[1] With the word "duration" we do not necessarily refer to a real time variable in the simulation. The actual interpretation for distances Δt between sampled points and for the total duration $\tau = M\Delta t$ depends from the specific simulation choices.

MD simulations are based on calculating the time evolution of a physical system by integrating Newton's equations of motion

$$\begin{cases} m_i \ddot{\mathbf{r}}_i(t) = -\nabla_{\mathbf{r}_i} V(r(t)) \\ (\mathbf{r}_i(0), \dot{\mathbf{r}}_i(0)) = (\mathbf{r}_i^0, \dot{\mathbf{r}}_i^0) \end{cases} \quad , \quad i = 1, \ldots, N \tag{2.7}$$

providing a suitable integration scheme and an interaction potential function $V(r)$. The numerical solution of Eq. (2.7) can be summarized in a two step process: the calculation of forces and the update of positions and velocities (or momenta) according to these forces.

The definition of the potential $V(r)$ is crucial in the construction of a well-defined molecular simulation. In an atomistic framework where there are no orbitals at disposal, suitable functions has to be chosen to represent the interaction between atoms. The empirical potential in such a framework usually takes the form

$$V(r) = V_{\text{bonded}}(r) + V_{\text{non-bonded}}(r) + V_{\text{long-range}}(r) \tag{2.8}$$

where the specific form of these functions, together with the complete set of parameters for them, are globally called a *force field* (FF).

In the above formula, the three terms in the force field account for different effective interactions, listed below:

- V_{bonded} terms mimic the interaction between covalently bonded atoms. The behavior of covalent bonds is reproduced by a sum of contributions

$$V_{\text{bonded}}(r) = \sum_{ij}^{N} V_{\text{bond}}(r_{ij}) + \sum_{ijk}^{N} V_{\text{angle}}(\theta_{ijk}) + \\ + \sum_{ijkl}^{N} V_{\text{torsion}}(\phi_{ijkl}) + \sum_{ijkl}^{N} V_{\text{improper}}(\xi_{ijkl}) \tag{2.9}$$

  with 2-, 3- and 4-body interaction terms (for a pictorial view see Fig. 2.1):

  (i) 2-body terms are used for modeling *bond stretching* between atoms ij, typically with harmonic terms

$$V_{\text{bond}}(r_{ij}) = \frac{1}{2} k_{ij}^b (r_{ij} - r_{ij}^0)^2 \tag{2.10}$$

  (ii) 3-body terms are used for modeling *bond angle bending* between atoms ijk, typically with harmonic terms

$$V_{\text{angle}}(\theta_{ijk}) = \frac{1}{2} k_{ijk}^a (\theta_{ijk} - \theta_{ijk}^0)^2 \tag{2.11}$$

  (iii) 4-body terms are used for modeling *rotation around bonds*, typically with sinusoidal terms

$$V_{\text{torsion}}(\phi_{ijkl}) = \frac{1}{2} k_{ijkl}^d \left[1 + \cos\left(n\phi_{ijkl} - \phi_{ijkl}^0\right)\right] \tag{2.12}$$

(iv) 4-body terms are also used to *confine geometries* (*e.g.* to impose planar or tetrahedral configurations). Typically is done with harmonic terms

$$V_{\text{improper}}(\xi_{ijkl}) = \frac{1}{2}k^i_{ijkl}(\xi_{ijkl} - \xi^0_{ijkl})^2 \tag{2.13}$$

- $V_{\text{non-bonded}}$ terms mimic Van der Waals interactions, typically by means of the Lennard-Jones potential

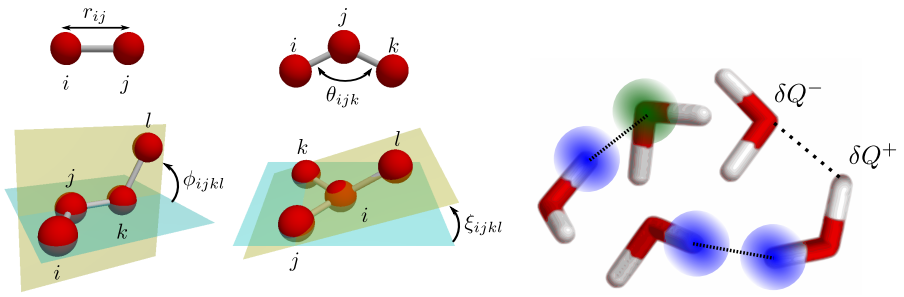
$$V_{\text{LJ}}(r_{ij}) = \varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \quad , \quad \begin{cases}\varepsilon_{ij} = \sqrt{\varepsilon_{ii}\varepsilon_{jj}} \\ \sigma_{ij} = \frac{1}{2}(\sigma_{ii} + \sigma_{jj})\end{cases} \tag{2.14}$$

where the indicated combination rules compute the interaction parameters between different species i and j from individual parameters;

- $V_{\text{long-range}}$ interactions are Coulomb interactions

$$V_{\text{C}}(r_{ij}) = \frac{1}{4\pi\varepsilon_0}\frac{\delta Q_i \delta Q_j}{\varepsilon_r r_{ij}} \tag{2.15}$$

between partial charges δQ_i and δQ_j (even if the system is globally neutral, partial charges are present to mimic displacement in electronic densities) at distance r_{ij}, with ε_r the relative dielectric constant.



(a) Bonded potentials (b) Non-bonded potentials

Figure 2.1: Interaction terms in force fields

At each dynamical step the MD engine should perform the calculation of forces

$$\mathbf{F}_i = -\nabla_{\mathbf{r}_i} V(r) \tag{2.16}$$

for each particle $i = 1, \ldots, N$, to permit the successive update of position and velocities. The whole FF is evaluated, using the proper parameters and the necessary prescriptions to handle interactions (*e.g.* lists of neighbors, cutoff distances, proper algorithm for long range electrostatics,...). For further informations on the subject the reader is reminded to the literature (like in Ref. [56, 1], or for an example of explicit implementation see [20, 111]).

Direct integration and the microcanonical ensemble

An integration scheme for Eq. (2.7) could start from the Taylor expansion

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \dot{\mathbf{r}}_i(t)\Delta t + \frac{1}{2}\ddot{\mathbf{r}}_i(t)(\Delta t)^2 + \ldots \tag{2.17}$$

of the new position $\mathbf{r}_i(t+\Delta t)$ about the current position $\mathbf{r}_i(t)$. With appropriate algebraic manipulations it is possible to calculate updated positions and velocities at the time-step $t+\Delta t$ with the desired level of approximation. As an example, we will show briefly the so-called *leap-frog* scheme [73]. First, from the forces $\mathbf{F}_i = -\nabla_{\mathbf{r}_i} V(r)$ new velocities

$$\dot{\mathbf{r}}_i(t+\tfrac{\Delta t}{2}) = \dot{\mathbf{r}}_i(t-\tfrac{\Delta t}{2}) + \Delta t \frac{\mathbf{F}_i}{m_i} + \mathcal{O}\left((\Delta t)^3\right) \tag{2.18}$$

are calculated at (half) time-step position. Second, from the velocities new positions

$$\mathbf{r}_i(t+\Delta t) = \mathbf{r}_i(t) + \dot{\mathbf{r}}_i(t+\tfrac{\Delta t}{2})\Delta t + \mathcal{O}\left((\Delta t)^3\right) \tag{2.19}$$

at (full) time-step are calculated (for more details on this algorithm and for comparison with respect to other implementation see [22]).

To circumvent the natural finite size of the simulation box, besides this integration scheme *periodic boundary conditions* (PBC) are often used. With PBC, when a dynamic step pushes a particle out through one side of the simulation box, then the particle is re-positioned on the opposite side. In this way the spurious finite-size effect from the boundaries of the simulation box can be avoided, and properties of bulk systems can be described. Connected to the use of PBC, distances in force calculation have to be computed using the so-called the *minimum image convention*. Moreover, in order to avoid self-interactions with periodic copies, the simulation box and the interaction cutoff distance has to be commensurable.

Letting a simulation routine to integrate Eq. (2.7), the points of the trajectory $\{(r(t_i), p(t_i))\}_{i=0,\dots,M}$ are naturally eligible as regular sampled configuration from the *microcanonical ensemble*: trajectory points, indeed, are sampled within a system with constant energy up to machine precision, if the integration scheme is simplectic (see Refs. [56, 105]). The microcanonical probability measure is presented in Panel 2.1

The microcanonical probability measure is defined as

$$\mu_{\text{NVE}}(\mathrm{d}r\mathrm{d}p) = \frac{\delta_{H(r,p)-E}(\mathrm{d}r\mathrm{d}p)}{\mathrm{Z}_{\text{NVE}}} \quad , \quad \mathrm{Z}_{\text{NVE}} = \int\limits_{S(E)} \delta_{H(r,p)-E}(\mathrm{d}r\mathrm{d}p) \tag{2.20}$$

where $S(E) = \left\{(r,p) \in \mathcal{D} \times \mathbb{R}^{3N} \;\middle|\; H(r,p) = E\right\}$ is the hypersurface at constant energy E. The normalization Z_{NVE} is the microcanonical partition function. The measure $\delta_{H(r,p)-E}(\mathrm{d}r\mathrm{d}p)$ is intuitively the standard measure $\mathrm{d}r\mathrm{d}p$ on $\mathcal{D} \times \mathbb{R}^{3N}$ "projected" onto the hypersurface $S(E)$. Formally, $\delta_{H(r,p)-E}(\mathrm{d}r\mathrm{d}p)$ is constructed as the Lesbegue measure for the set $\mathcal{N}_{\Delta E}(E) = \left\{(r,p) \in \mathcal{D} \times \mathbb{R}^{3N} \middle| E \le H(r,p) \le E + \Delta E\right\}$, in the limit $\Delta E \to 0$, by means of

$$\int\limits_{S(E)} A(r,p)\delta_{H(r,p)-E}(\mathrm{d}r\mathrm{d}p) = \lim_{\Delta E \to 0} \frac{1}{\Delta E} \int\limits_{\mathcal{N}_{\Delta E}(E)} A(r,p)\mathrm{d}r\mathrm{d}p \tag{2.21}$$

with A a test observable.

Panel 2.1: Microcanonical probability measure

It is worth spending a few words on algorithm stability. It is known that integration schemes, like the leap-frog algorithm, could in principle suffer of numerical instabilities or dependences on the initial conditions. It is for example the case of the so-called *Lyapunov instability*, for which two trajectories with very close initial conditions could undergo in an exponential divergence in short times. If the full trajectory is needed from the simulation, then this kind of instability is not desirable because a very different dynamic description could emerge even from very close initial conditions. However, this problem affects the dynamical description, while it does not affect the single points of an arbitrary trajectory to be regularly sampled from the microcanonical distribution. Thus, if the focus is on statistical averages or on the (ergodic) exploration of the phase space, then this numerical instability does not affect simulation results.

Modified dynamics: canonical and isothermal-isobaric ensembles

The direct solution of Eq. (2.7) is by construction limited to the microcanonical ensemble. This means that without any modification only the microcanonical ensemble can be sampled in a MD simulation. In our case we are interested in doing simulations of monosaccharides, in vacuum or in explicit water, at room temperature T_0 and atmosferic pressure P_0. To escape the microcanonical sampling a modification of the equation of motion is then required.

For vacuum simulations we are only interested in keeping the system at constant temperature T. One way to do so is to integrate the Langevin's equations of motion

$$\mathrm{d}\dot{\mathbf{q}}_i(t) = -\nabla_{\mathbf{q}_i} V(q(t))\mathrm{d}t - \gamma\dot{\mathbf{q}}_i(t)\mathrm{d}t + \sigma\mathrm{d}W(t) \quad (\sigma^2 = 2k_B T\gamma) \tag{2.22}$$

instead of Newton's equations of motion. In Eq. (2.22) is also indicated the fluctuation-dissipation relation: it connects the dissipative term $-\gamma\dot{\mathbf{q}}_i(t)\mathrm{d}t$ and the random Brownian term $\sigma\mathrm{d}W(t)$, and assures that this stochastic dynamics samples the *canonical ensemble* (see Ref. [105]). The system has then constant temperature T and fluctuating energy U. A trajectory $\{(r(t), p(t))\}$ from Eq. (2.22) is then ergodic with respect to the probability measure shown in Panel 2.2.

When simulations with explicit water are needed, a finite simulation box is necessary. Moreover, a pressure control on the system is also possible, besides the temperature control. In these cases, the use of *extended Lagrangian* methods could be useful not only for the temperature T but also for pressure P control. The main idea is to introduce in the Lagrangian some extra degrees of freedom that couple with the standard one. The resulting equations of motion turns out to represent properly the effect of a heat and/or pressure reservoir acting on the system. To give a practical example, the *Nosé-Hoover thermostat* [131, 129, 130, 74, 75] is briefly presented. Consider the Lagrangian

$$\mathcal{L}_{\text{Nosé}}(r, \dot{r}, s, \dot{s}) = \sum_{i=1}^{N} \frac{1}{2} m_i s^2 \dot{r}_i^2 - V(r) + \frac{Q}{2}\dot{s}^2 - g k_B T_0 \ln s \tag{2.26}$$

where a rescaling factor s for the velocities is introduced (the standard Lagrangian is recovered when $s = 1$). At a difference with simple rescaling meth-

The canonical probability measure is defined as

$$\mu_{\text{NVT}}(\mathrm{d}r\mathrm{d}p) = \frac{\mathrm{e}^{-\beta\mathcal{H}(r,p)}\,\mathrm{d}r\mathrm{d}p}{\mathrm{Z}_{\text{NVT}}} \quad , \quad \mathrm{Z}_{\text{NVT}} = \int\limits_{\mathcal{D}\times\mathbb{R}^{3N}} \mathrm{e}^{-\beta\mathcal{H}(r,p)}\,\mathrm{d}r\mathrm{d}p \tag{2.23}$$

($\beta = 1/k_BT$, k_B being the Boltzmann's constant). The normalization Z_{NVT} is the canonical partition function.
For separable Hamiltonians ($\mathcal{H}(r,p) = K(p) + V(r)$) the form

$$\mu_{\text{NVT}}(\mathrm{d}r\mathrm{d}p) = \nu(\mathrm{d}r)\kappa(\mathrm{d}p) \quad , \quad \begin{cases} \nu(\mathrm{d}r) = {\mathrm{Z}_\nu}^{-1}\,\mathrm{e}^{-\beta V(r)}\,\mathrm{d}r \\ \kappa(\mathrm{d}p) = {\mathrm{Z}_\kappa}^{-1}\,\mathrm{e}^{-\beta K(p)}\,\mathrm{d}p \end{cases} \tag{2.24}$$

for the probability measure, and the factorization

$$\mathrm{Z}_{\text{NVT}} = \mathrm{Z}_\nu\mathrm{Z}_\kappa \quad , \quad \begin{cases} \mathrm{Z}_\nu = \int\limits_{\mathcal{D}} \mathrm{e}^{-\beta V(r)}\,\mathrm{d}r \\ \mathrm{Z}_\kappa = \int\limits_{\mathbb{R}^{3N}} \mathrm{e}^{-\beta K(p)}\,\mathrm{d}p = \left(\frac{2\pi}{\beta}\right)^{3N/2} \prod\limits_{i=1}^{N} m_i^{3/2} \end{cases} \tag{2.25}$$

for the partition function, are possible.

Panel 2.2: Canonical probability measure

ods, like the one of Berendsen [23][2], the parameter s is introduced as a real extra dynamical variable (a degree of freedom of "mass" Q), and g is a number related to the total number of degrees of freedom. Solving the equations of motion generated from $\mathcal{L}_{\text{Nosé}}$ produces a MD simulation that samples the microcanonical partition function

$$\mathrm{Z}_{\text{Nosé}}(N,V,\mathcal{E}) = \int\limits_{\mathcal{D}\times\mathbb{R}^{3N}\times[0,+\infty)\times\mathbb{R}} \delta(\mathcal{H}_{\text{Nosé}}(r,p,s,p_s) - \mathcal{E})\mathrm{d}s\mathrm{d}p_s\mathrm{d}r\mathrm{d}p \tag{2.27}$$

(and the associated microcanonical measure) in the extended phase space. Under suitably chosen conditions this microcanonical sampling of trajectory points $\{(r(t_i), p(t_i), s(t_i), p_s(t_i))\}_{i=1,\ldots,M}$ in the extended system (r,p,s,p_s) performs a canonical sampling for two "real" variables r' and p' of a temperature controlled system, namely

$$\frac{1}{M}\sum_{i=1}^{N} A(r(t_i), p(t_i), s(t_i), p_s(t_i)) \simeq \langle A(r,p,s,p_s)\rangle_{\text{Nosé}} = \langle A(r',p')\rangle_{\text{NVT}} \quad . \tag{2.28}$$

The effective behavior of a thermostat could be shown considering the equation of motion of the friction coefficient $\xi = \dot{s} = s'p'_s/Q$, that is

$$\dot{\xi} \propto (T - T_0) \quad , \quad T = \frac{1}{3Nk_{\text{B}}}\sum_{i=1}^{N}\frac{\mathbf{p}_i'^2}{m} \tag{2.29}$$

[2] The description of such methods is beyond the scope of this thesis work, for further reading see [56].

This means that the kinetic temperature T oscillates around the chosen reference temperature T_0, giving the effect of a thermostat.

As we stated before, in a similar way it is possible to extend the Lagrange's equation of motion to include also a pressure coupling, in order to simulate systems at constant pressure. A famous realization of this kind of barostat is the *Parrinello-Rahman* scheme [138], in which a scaling factor acting on the simulation box length is treated as an extra dynamical variable. Combined with the Nosé-Hoover scheme, the aim is again to sample a microcanonical measure on an extended system that reproduces in real variables the sampling from the isothermal-isobaric ensemble (for the corresponding probability measure see Panel 2.3).

The isothermal-isobaric probability measure is defined as

$$\mu_{\text{NPT}}(\mathrm{d}r\mathrm{d}p\mathrm{d}V) = \frac{\mathrm{e}^{-\beta PV}\,\mathrm{e}^{-\beta\mathcal{H}(r,p)}\,\mathrm{d}r\mathrm{d}p\mathrm{d}V}{\mathrm{Z}_{\text{NPT}}} \quad ,$$
$$\mathrm{Z}_{\text{NPT}} = \int_{\mathcal{D}_V\times\mathbb{R}^{3N}\times(0,+\infty)} \mathrm{e}^{-\beta PV}\,\mathrm{e}^{-\beta\mathcal{H}(r,p)}\,\mathrm{d}r\mathrm{d}p\mathrm{d}V \qquad (2.30)$$

where $\mathcal{D}_V$ is the configurational space of given volume V, with admissible volume values $V \in (0,+\infty)$. The normalization Z_{NPT} is the isothermal-isobaric partition function.

Panel 2.3: Isothermal-isobaric probability measure

2.2 Free energy calculations and accelerated MD

The concept of free energy is central in thermodynamics and in all recent biochemical studies. Indeed, the representation of chemical reactions in terms of free energy, naturally emerging from a dimensional reduction of the problem, focuses directly on the small set of relevant parameters. In the following we will show an overview on some aspects related to free energy definition and usage. We will address the meaning of known formulas on free energy

$$F = -\frac{1}{\beta}\ln\int \mathrm{e}^{-\beta V(r)}\,\mathrm{d}r \qquad (2.31a)$$

$$F(z) = -\frac{1}{\beta}\ln\int \mathrm{e}^{-\beta V(r)}\,\delta(z-\xi(r))\mathrm{d}r \qquad (2.31b)$$

and we will introduce the computational methods to perform free energy computations.

Consider a system which probability distribution is canonical[3] (see Panel 2.2, following again the notation of Ref. [105]). From statistical physics, the *absolute free energy* F of a system is

$$F = -\frac{1}{\beta}\ln \mathrm{Z}_{\text{NVT}} \quad , \quad \mathrm{Z}_{\text{NVT}} = \int_{\mathcal{D}\times\mathbb{R}^{3N}} \mathrm{e}^{-\beta\mathcal{H}(r,p)}\,\mathrm{d}r\mathrm{d}p \qquad (2.32)$$

[3] The extension to other thermodynamical ensemble is straightforward.

namely the logarithm of the canonical partition function. The free energy of Eq. (2.32) is the so-called *Helmholtz free energy*[4]

$$F = U - TS \tag{2.33}$$

known from thermodynamics: F is the thermodynamical potential associated to the canonical ensemble, where U the internal energy and S the entropy. Similarly, given the partition function Z of other statistical ensembles, the proper thermodynamic potential $\mathcal{A} = -k_B T \ln \mathrm{Z}$ is recovered (*e.g.* in the microcanonical ensemble gives the entropy S, in the isothermal-isobaric ensemble gives the *Gibbs free energy* $G = F + PV$). If it is possible to sample the partition function $\mathrm{Z}_{\mathrm{NVT}}$, then the value of F could be calculated. In the case of separable Hamiltonian (see Panel 2.2), the definition of Eq. (2.32) simplifies to

$$F = -\frac{1}{\beta} \ln \mathrm{Z}_\nu - \frac{1}{\beta} \ln \mathrm{Z}_\kappa \tag{2.34}$$

and the difficult part in the sampling is only the configurational part. With this simplification, the structure of Eq. (2.34) now resembles the commonly indicated form of Eq. (2.31a).

The free energy inherits from the potential energy $V(r)$ the property of being defined up to an immaterial additive constant. This explains the interest in *free energy differences* ΔF rather than in absolute values F. The calculation of such differences implies the definition of at least two different thermodynamic states. Often, different thermodynamic states, like reactants and products of chemical reactions, can be distinguished by means of few parameters, called *reaction coordinates*. These coordinates are most of the time collective variables (CVs) of the configurational coordinates:

$$\xi : \mathcal{D} \to \mathbb{R}^m \quad , \quad m \leq 3N \quad . \tag{2.35}$$

Different values of the reaction coordinate $\xi(r) = z$ define a foliation of the space of configurations as a collection of the subsets (submanifolds) $\Sigma(z)$, namely

$$\Sigma(z) = \left\{ r \in \mathcal{D} \;\middle|\; \xi(r) = z \right\} \quad \Rightarrow \quad \mathcal{D} = \bigcup_{z \in \mathbb{R}^m} \Sigma(z) \quad . \tag{2.36}$$

By integrating the canonical measure Eq. (2.23) on the subspaces $\Sigma(z) \times \mathbb{R}^{3N}$, we perform for each $z \in \mathbb{R}^m$ the dimensional reduction, induced by the function $\xi(r)$, of the probability measure. In other words, starting from the canonical probability measure on the reduced space $\mathbb{R}^m$ the marginal probability distribution

$$\mu^\xi(\mathrm{d}z) = \underbrace{\left(\int_{\Sigma(z) \times \mathbb{R}^{3N}} \frac{\mathrm{e}^{-\beta \mathcal{H}(r,p)}}{\mathrm{Z}_{\mathrm{NVT}}} \delta_{\xi(r)-z}(\mathrm{d}r)\mathrm{d}p \right)}_{\rho(z)} \mathrm{d}z \quad , \quad \int_{\mathbb{R}^m} \mu^\xi(\mathrm{d}z) = 1 \tag{2.37}$$

[4] For a detailed discussion on the connection between the microscopic definition of Eq. (2.32) and the macroscopic one of Eq. (2.33) see [105].

is induced. The measure $\mu^{\xi}(\mathrm{d}z)$, that represents the states of the system with respect to the given $\xi(r)$ coordinate, is normalized automatically from the normalization of the original measure $\mu_{\mathrm{NVT}}(\mathrm{d}r\mathrm{d}p)$. The definition of the measure $\delta_{\xi(r)-z}(\mathrm{d}r)$ through the relation[5]

$$\delta_{\xi(r)-z}(\mathrm{d}r)\mathrm{d}z = \mathrm{d}r \tag{2.38}$$

leads to the physical insight that this mathematical definition resembles somehow a *coarse graining* procedure. Indeed, the reaction coordinate collects all the relevant degrees of freedom for a transition represented as the evolution between different regions $\Sigma(z)$ of the phase space. The measure $\delta_{\xi(r)-z}(\mathrm{d}r)$ represents all degrees of freedom that have to be "integrated out" (in each subset $\Sigma(z)\times\mathbb{R}^{3N}$) from the whole coordinates r to describe the system in terms of the sole z variable. The weight function $\mathrm{e}^{-\beta\mathcal{H}(r,p)}/\mathrm{Z}_{\mathrm{NVT}}$ in Eq. (2.37) ensures that this coarse graining procedure is performed in the canonical ensemble, namely that the marginal probability distribution $\mu^{\xi}(\mathrm{d}z)$ is derived from the canonical measure $\mu_{\mathrm{NVT}}(\mathrm{d}r\mathrm{d}p)$.

The absolute free energy is defined as the log-density of the marginal distribution of Eq. (2.37)

$$\boxed{\rho(z) = \mathrm{e}^{-\beta F(z)} \quad \Leftrightarrow \quad F(z) = -\frac{1}{\beta}\ln\left(\int\limits_{\Sigma(z)\times\mathbb{R}^{3N}} \frac{\mathrm{e}^{-\beta\mathcal{H}(r,p)}}{\mathrm{Z}_{\mathrm{NVT}}}\delta_{\xi(r)-z}(\mathrm{d}r)\mathrm{d}p\right)} \quad . \tag{2.39}$$

This $F(z)$ is somehow a free energy "density", because the integration within the logarithm is not extended on the whole space $\mathcal{D}$ but only on a submanifold $\Sigma(z)$. This definition has now a formulation closer to Eq. (2.31b). The free energy difference between two states (for example z_0 and z_1) is then

$$\Delta F = F(z_1) - F(z_0) = -\frac{1}{\beta}\ln\left(\frac{\displaystyle\int_{\Sigma(z_1)\times\mathbb{R}^{3N}} \mathrm{e}^{-\beta\mathcal{H}(r,p)}\,\delta_{\xi(r)-z_1}(\mathrm{d}r)\mathrm{d}p}{\displaystyle\int_{\Sigma(z_0)\times\mathbb{R}^{3N}} \mathrm{e}^{-\beta\mathcal{H}(r,p)}\,\delta_{\xi(r)-z_0}(\mathrm{d}r)\mathrm{d}p}\right) \tag{2.40}$$

$$= -\frac{1}{\beta}\ln\left(\frac{\displaystyle\int_{\Sigma(z_1)} \mathrm{e}^{-\beta V(r)}\,\delta_{\xi(r)-z_1}(\mathrm{d}r)}{\displaystyle\int_{\Sigma(z_0)} \mathrm{e}^{-\beta V(r)}\,\delta_{\xi(r)-z_0}(\mathrm{d}r)}\right) \tag{2.41}$$

where the last equality holds for separable Hamiltonians.

It has to be mentioned that $z\in\mathbb{R}^m$ values may not represent directly thermodynamic states. In fact, this depends on the actual definition of the reaction coordinate ξ, because it could happen in a region $S\subset\mathbb{R}^m$ that free energy values $F(z)|_{z\in S}$ are within an energy window ΔF comparable to the thermal energy k_BT. In this case, in thermodynamic sense all the points $z\in S(z)$ belong to the same thermodynamic basin, because their free energies $F(z)$ are within the thermal accessible threshold[6]. For these sets S it is useful to calculate the

[5] The formal definition is the same described for the microcanonical measure of Eq. (2.20).

[6] The $z\in S\subset\mathbb{R}^m$ values correspond somehow to "mesostates", between the original microstates $(r,p)\in\mathcal{D}\times\mathbb{R}^{3N}$ and the thermodynamics macrostates S.

population probability

$$P[S] = \int_S \mu^\xi(\mathrm{d}z) = \int_S \mathrm{e}^{-\beta F(z)}\,\mathrm{d}z \tag{2.42}$$

(where by construction[7] $P[\mathbb{R}^m] = 1$) and the *"mean" Free Energy*

$$\bar{F}[S] = -\frac{1}{\beta}\ln P[S] \tag{2.43}$$

by means of direct Boltzmann inversion. These extra summations over the reduced space $\mathbb{R}^m$ in Eqs. (2.42) and (2.43) complete somehow the summation over the whole phase space that the dimentional reduction $\mathcal{D} \xrightarrow{\xi} \mathbb{R}^m$ does not perform totally. The aforementioned quantities result then to be much closer to measured observables in terms of thermodynamics states.

2.2.1 Rare events and the need for acceleration

Most natural phenomena exhibit a wide range of characteristic time scales. To give an example, for proteins there are very different time scales, from the vibration of covalent bonds, to the rotation around backbone dihedral angles and of side chains, and to the dynamic oscillation of secondary structure, to name a few. In a MD simulation the integration time-step has to be commensurate with the fastest internal dynamics (typically $\Delta t \sim 1\,\mathrm{fs}$). This obviously sets a limit to the available total simulated time and, consequently, to the ability of a simulation to reproduce some classes of events. However, this limitation (nowadays around hundreds of ns) is in principle not only a matter of lack of computational power, and cannot be simply solved with larger computers. There are, indeed, phenomena that happen at time scales several order of magnitude larger than the integration time scale. A simple example are chemical processes, that could happen in the range µs ÷ s, very far from accessible simulation times. Similarly, systems which are characterized by *rare events* (*e.g.* protein folding processes, structural phase transition, ...) or systems that exhibits *metastabilities* (in which thermodynamical basins are separated by very high, non-thermal free energy barriers) highly suffer for the time scale separation.

Systems which exhibit the above features are non-ergodic in practice from the simulation point of view. For example, in the case of metastable systems an evaluation of the free energy $F(z) = -k_B T \ln \rho(z)$ by a direct estimate of the probability density

$$\rho(z) = \int_{\mathbb{R}^m} \rho(z')\delta(z'-z)\mathrm{d}z' \xrightarrow[z'=\xi(r)]{\text{ergodic hyp.}} \lim_{\mathcal{T}\to+\infty} \frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} \delta(\xi(r(\tau)) - z)\mathrm{d}\tau \tag{2.44}$$

(that is, the direct free energy reconstruction with histogram methods) will be unfeasible: the ergodic hypothesis is not be fulfilled in practice because the trajectory $z(t) = \xi(r(t))$ remains trapped in free energy basins for an unpredictably long time (see Fig. 2.2). For these classes of phenomena increasing

[7] Here the proper normalization factor Z^{-1} in $\mu^\xi(\mathrm{d}z)$ is present by construction. In computational calculations the normalization has to be imposed properly.

Figure 2.2: Pictorial view of metastable systems. The barrier crossing is a stochastic event because the order of magnitude of the free energy barrier is much higher than the thermal energy k_BT. The indicated basins are then metastable states.

computational power is not the solution, as the transition probabilities decrease exponentially with the free energy barrier heigth.

To overcome the issue of non-ergodicity, two main theoretical approaches are possible. On one hand, it is possible to switch from a full atomistic description to *coarse grained* models. The timescale problem is then circumvented by a dimensionality reduction. The computational gain is substantial by means of a huge reduction in the number of particles that has to be simulated. On the other hand, it is possible to modify the dynamical description of the system in order to *accelerate rare events*. In this way available computer resources are used with enhanced efficiency. The timescale problem is avoided and the sampling ability is improved substantially.

In the last few years an impressive number of methods have been developed in order to accelerate the dynamics of rare events/metastabilities. This large variety of different solutions underlines once more the great interest in this kind of problems. In an extremely simplified view of the subject, one could identify two general purposes of these methods:

(i) provide *sampling methods* to enhance the sampling of the phase space. Slight modifications in the Hamiltonian, for example with simple time-independent terms, are made to improve the statistical sampling in user defined regions. The sampling could be re-weighted in order to recover the unbiased probability distribution (*e.g.* thermodynamics integration [33, 168], umbrella sampling [139], weighted-histogram techniques [53, 97, 152], Jarzynski's identity-based methods [82, 42], adaptive force bias [44, 151]);

(ii) provide *searching methods* to enhance the exploration of the phase space. In some cases, the Hamiltonian of the system is modified (to give non-Newtonian equation of motion, or to consider multiple copies, ...) to provide the necessary acceleration. With this respect, the most efficient modifications benefit of the progressive build up of time-dependent bias potentials that disfavor the system to return in previously visited con-

figurations. The exploration of the phase space is very fast, but some methods suffer of scaling problem with the dimensionality of the system. (*e.g.* transition path sampling [46, 45, 141], parallel tempering [120] and replica exchange [172], local elevation [44, 151], metadynamics [100, 99]).

We want to stress that the proposed division, inspired by Ref. [99], is rather arbitrary: most of the accelerated MD methods present in literature are techniques somehow in between these two extremes. Moreover, the rapid evolution and extension of these techniques constantly enhance the specific ability of simulation methods, producing much more versatile tools.

To have an overview on the subject of accelerated MD techniques, two very interesting reviews (with some indication of historical development and tentative general classifications) are presented in the work of Laio and Gervasio on Metadynamics and its variants [99], and in the work of Hansen and Hünenberger on Local Elevation Umbrella Sampling method [66]. The techniques described in these works are also two examples of searching methods that could benefit also of sampling ideas. In this work the chosen method to accelerate rare events and to overcome metastabilities is the metadynamics algorithm described in the following.

2.2.2 Metadynamics: basic concepts

Consider a N-particles system, interacting with the potential $V(r)$, and evolving under the action of a dynamics (*e.g.* Molecular Dynamics, Langevin Dynamics, Monte Carlo, ...) having a canonical equilibrium distribution at a temperature T. In the spirit of Section 2.2, it is possible to define a reaction coordinate $\xi(r)$, and its values $z \in \mathbb{R}^m$ label thermodynamic states of interest. For system with metastability the dynamics will be stuck in some local minimum of $V(r)$, escaping from it with a very low probability (see Fig. 2.2). The idea is to add a history-dependent potential to bias the system not to return to previously visited points in the configurational space. In basic implementations of metadynamics [100, 101, 14] the bias potential reads

$$\boxed{V_{\text{B}}(\xi(r),t) = \sum_{\substack{t_k=k\tau_G \\ k\in\mathbb{N}}}^{t_k<t} w \prod_{a=1}^{m} \mathrm{e}^{-\frac{[\xi^a(r)-z_k^a]^2}{2\sigma_a^2}}} \tag{2.45}$$

which is the sum of Gaussian functions deposited every τ_G time interval at positions $\xi^a(r_{\text{B}}(t_k)) \equiv z_k^a$ along the trajectory $r_{\text{B}}(t)$. A pictorial view of this deposition process is given in Fig. 2.3. The parameters in Eq. (2.45), the Gaussian height w, the Gaussian widths σ_a, and the deposition time interval τ_G, are the user-defined control parameters.

The system (time-dependent) Hamiltonian now reads

$$\mathcal{H}_{\text{B}}(r,p,t) = K(p) + V(r) + V_{\text{B}}(\xi(r),t) \tag{2.46}$$

and the biased dynamics is driven by the force

$$\mathbf{F}_i^{\text{B}} = -\nabla_{\mathbf{r}_i} V(r) - \nabla_{\mathbf{r}_i} V_{\text{B}}(\xi(r),t) = \mathbf{F}_i + \mathbf{F}_i^{\text{META}} \tag{2.47}$$

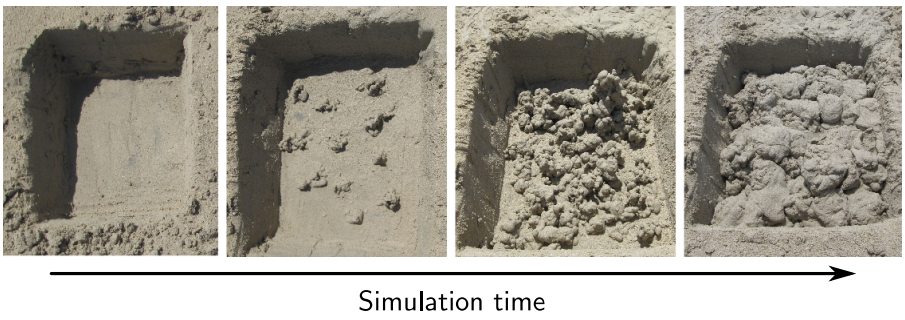


Figure 2.3: History dependent potential concept. Images, based on the example of [99], are the representation of changes in the potential $V(r) + V_{\text{B}}(\xi(r), t)$. A single particle in a basin (the empty sand pool, first image) is trapped with $V(r) = 0$ on the floor and $V(r) = V_0$ at walls. The particle follows its dynamical trajectory, and during the time evolution Gaussian functions (the sand spots, second image) are deposited every time interval τ_G, increasing the term $V_{\text{B}}(\xi(r), t)$. The basin is progressively filled (third image), and eventually the height of the wall potential barrier has been lowered enough for the particle to escape (forth image). The deposited sand is the negative image of the basin, and thus it is an estimate of the shape of the basin itself.

instead of the sole standard force term $\mathbf{F}_i$ of Eq. (2.7). The extra "meta-forces" $\mathbf{F}_i^{\text{META}}$ are explicitly

$$\begin{aligned}
\mathbf{F}_i^{\text{META}} &= -\nabla_{\mathbf{r}_i} V_{\text{B}}(\xi(r), t) \\
&= -\sum_{\substack{t_k = k\tau_G \\ k \in \mathbb{N}}}^{t_k < t} w \nabla_{\mathbf{r}_i} \prod_{a=1}^{m} \mathrm{e}^{-\frac{[\xi^a(r) - z_k^a]^2}{2\sigma_a^2}} \\
&= \sum_{\substack{t_k = k\tau_G \\ k \in \mathbb{N}}}^{t_k < t} w \left(\sum_{b=1}^{m} \frac{\xi^b(r) - z_k^b}{\sigma_b^2} \nabla_{\mathbf{r}_i} \xi^b(r) \right) \prod_{a=1}^{m} \mathrm{e}^{-\frac{[\xi^a(r) - z_k^a]^2}{2\sigma_a^2}}
\end{aligned} \tag{2.48}$$

by direct calculation from Eq. (2.45). As can be noticed, these "meta-forces" are re-summations of the deposited Gaussians, weighted by a factor that contains the derivative $\nabla_{\mathbf{r}_i} \xi(r) = (\nabla_{\mathbf{r}_i} \xi^1(r), \ldots, \nabla_{\mathbf{r}_i} \xi^m(r))$ of the collective variable $\xi(r)$ with respect to the original coordinates $\mathbf{r}_i$. Once the reaction coordinate $\xi(r)$ is chosen, the gradients $\nabla_{\mathbf{r}_i} \xi(r)$ could be calculated analytically.

If now we compute the equivalent of Eq. (2.37) for the biased Hamiltonian

of Eq. (2.46) we obtain

$$\begin{aligned}\mu_{\text{B}}^{\xi}(\mathrm{d}z,t) &= \left(\int\limits_{\Sigma(z)\times\mathbb{R}^{3N}} \mathrm{e}^{-\beta\mathcal{H}_{\text{B}}(r,p,t)}\,\frac{\delta_{\xi(r)-z}(\mathrm{d}r)}{\mathrm{Z}(t)}\mathrm{d}p\right)\mathrm{d}z \\ &= \frac{1}{\mathrm{Z}(t)}\left(\int\limits_{\Sigma(z)\times\mathbb{R}^{3N}} \mathrm{e}^{-\beta[\mathcal{H}(r,p)+V_{\text{B}}(\xi(r),t)]}\,\delta_{\xi(r)-z}(\mathrm{d}r)\mathrm{d}p\right)\mathrm{d}z \\ &= \mathrm{e}^{-\beta V_{\text{B}}(z,t)}\,\frac{\mathrm{Z}_{\text{NVT}}}{\mathrm{Z}(t)}\mu^{\xi}(\mathrm{d}z) \end{aligned} \tag{2.49}$$

where $\mathrm{Z}(t) = \int\limits_{\mathcal{D}\times\mathbb{R}^{3N}} \mathrm{e}^{-\beta\mathcal{H}_{\text{B}}(r,p,t)}\,\mathrm{d}r\mathrm{d}p$ is the normalization factor for $\mu_{\text{B}}^{\xi}(\mathrm{d}z,t)$ at the given time t, while Z_{NVT} is the canonical partition function of the unbiased system (see Panel 2.2). We want to stress that, because we have an out-of-equilibrium dynamics, strictly speaking the measure $\mu_{\text{B}}^{\xi}(\mathrm{d}z,t)$ is not a probability measure in the sense of statistical mechanics, and $\mathrm{Z}(t)$ consequently is not a partition function. However, in the limit of very large values of τ_G the function $V_{\text{B}}(\xi(r),t)$ slowly varies. Thus we can consider that between two Gaussian deposition the indicated quantities are "equilibrium" properties. With this respect, from Eq. (2.49) we evaluate the "equilibrium" probability density as

$$\rho_{\text{B}}(z,t) = \mathrm{e}^{-\beta[F(z)+V_{\text{B}}(z,t)]}\,\frac{\mathrm{Z}_{\text{NVT}}}{\mathrm{Z}(t)} \tag{2.50}$$

which shows that at the time t the free energy of the biased system is

$$F_{\text{B}}(z,t) = -\frac{1}{\beta}\ln\rho_{\text{B}}(z,t) = F(z) + V_{\text{B}}(z,t) \tag{2.51}$$

namely the original function $F(z)$ modified by the bias potential $V_{\text{B}}(z,t)$.

The core assumption of standard metadynamics (the so-called *"direct" metadynamics*, see Ref. [99]) is that the bias potential is eventually an unbiased estimator for the free energy, namely

$$\boxed{\lim_{\mathcal{T}\to+\infty} V_{\text{B}}(z,\mathcal{T}) \sim -F(z)} \quad . \tag{2.52}$$

In other words, up to an immaterial additive constant, the out-of-equilibrium quantity $V_{\text{B}}(z,\mathcal{T})$, with a sufficient high $\mathcal{T}$, is an unbiased estimator of the equilibrium quantity $F(z)$. This statement could be understood using again the slow-deposition argument. This regime is reached with slow Gaussian deposition (Gaussian height $w \to 0$ and/or deposition time $\tau_G \to +\infty$, or combined $w/\tau_G \to 0$), and when V_{B} varies slowly during the adding process the system is in a quasi-free dynamics. Nevertheless, if the system is in a local minimum z_0 for $F(z)$, then the trajectory $\xi(r(t))$ will spend a long time exploring that region, and consequently more and more Gaussians are added. When the local minimum is almost completely filled, a surrounding region $\Omega(z)$ of the CVs space has $F(z) \sim -V_{\text{B}}(z,t)$ and thus the probability ρ_{B} will be nearly constant, except for fluctuation of the order w. It has to be stressed that this argument is only qualitative and that, to the best of our knowledge, there is not a rigorous convergence proof or a thermodynamic argument that gives Eq. (2.52).

2.2.3 The choice of Collective Variables

The relevant dynamics of the system, as stated before, will be encoded in a suitable formed reaction coordinate $\xi(r)$. Equivalently, the CV retains all the information about the relevant changes in the system (the slow or rare event) without taking into account all the microscopic details of the dynamics of the system (*i.e.*, "integrating out" the fast degrees of freedom).

To this extent, a good definition for a CV set should satisfy three properties:

(i) to clearly distinguish between different states of the system (initial, final, intermediate);

(ii) to describe *all* the slow degrees of freedom related to the process of interest;

(iii) to be not too much in number (small value for m).

Point (iii), in the context of metadynamics calculation, is a necessary operative requirement. Indeed, the time needed for a metadynamics exploration of the reduced space $\mathbb{R}^m$ typically grows exponentially with the dimensionality m. If the number m is to high, the free energy reconstruction could be extremely time consuming. Concerning Points (i) and (ii), if the wrong reaction coordinate is chosen, or if an important collective variable is neglected, then a severe bias in the reconstruction is introduced. In particular, the biased dynamics could be affected by hysteresis effects, as described in [99] for a toy-model system, leading to wrong evaluation of the free energy landscape. A possible check to evaluate the goodness of a chosen CV is to look if the metadynamics along that coordinate is able to sample ergodically the reduced space. This aspect in relation with monosaccharides will be extensively described in Chapter 4, with the help of some toy-model free energy calculations.

Given these indications, every possible function $\xi(r)$ of the spatial coordinates is eligible to be a reaction coordinate. For example, *geometry related variables* (distances, angles, dihedrals, gyration radius, ...) or *"interaction" variables* (potential energy, coordination number, hydrogen bonds, ...) could be addressed for metadynamics. In Chapter 3 specific variables for monosaccharide puckering will be described.

2.2.4 Example of a metadynamics simulation

In practice, a metadynamics implementation could be inserted in every MD code in a simple way. The algorithm requires the call of a proper subroutine/function, during the main loop of the dynamics code, that performs these operations:

- every $\tau_G/\Delta t$ time-steps updates the bias potential: a new Gaussian function is added at the actual position $\xi(r) = z$ (that becomes the centroid of a new Gaussian in $V_{\text{B}}(z,t)$);
- every time-step computes the derivative $-\nabla_{\mathbf{r}_i} V_{\text{B}}(\xi(r),t)$ with respect to the original coordinates r; then, add these meta-forces $\mathbf{F}_i^{\text{META}}$ to the usual forces $\mathbf{F}_i$ on atoms (cfr. Eq. (2.47), this gives the $\mathbf{F}_i^{\text{B}}$ to bias the dynamics).

An operative scheme for a simulation may be the subsequent:

1. run a standard MD simulation to monitor the evolution of the selected reaction coordinate $\xi(r)$. This is useful to establish the characteristic size of variation of the variables in unbiased condition, that is the size of the smallest feature of interest in the free energy landscape. It is convenient to take the Gaussian widths σ^a smaller than the smallest observed fluctuation. Metadynamics is unable to reconstruct features in $F(z)$ on a scale smaller than σ^a;

2. guess the parameter (σ^a, τ_G, w) for the algorithm to work. Metadynamics bias potential $V_{\text{B}}(z,t)$ is unable to reconstruct free energy barrier lower than $w = \omega\tau_G$;

3. start metadynamics: during the run, collect the deposited Gaussians (centroids z_k^a and widths σ^a at time-steps t_k) for further free energy estimate;

4. monitor the explored positions in order to evaluate the diffusivity in the CV space and stop the reconstruction at time t^*.

During the run, the bias potential will in time "fill" all the minima of the free energy landscape. The bias itself (the sum of all deposited Gaussians) will be the estimate of the free energy, *i.e.*

$$\boxed{F_{\text{META}}(z,t^*) = -V_{\text{B}}(z,t^*)} \simeq F(z) \tag{2.53}$$

for a sufficient long simulation time t^*. After this simulation time, the dynamics of the reaction coordinates will be nearly diffusive. Equivalently, the probability density $\rho_{\text{B}}(z,t^*)$ of the biased system becomes nearly uniform (see Fig. 1 of Ref. [99]).

The slow deposition argument does not give a real proof of convergence of the bias potential to the free energy density function. However, it is clear that the filling process is able at least to give an estimator $F_{\text{META}}(z,t^*)$ that oscillates around the real free energy function $F(z)$. An estimate of the error of metadynamics is possible by means of averaging processes that benefit from this oscillation around the real value. On one hand, it can be shown (see [30, 99], where an overdamped Langevin model is implemented to study this problem of accuracy) that the average value of $V_{\text{B}}(z,t^*)$ over several independent metadynamics runs is equal to $-F(z)$. Denoting $\langle . \rangle_{\text{M}}$ the average over several independent metadynamics realizations, then we have

$$F_{\text{META}}(z,t^*) = -\left\langle V_{\text{B}}(z,t^*_{\text{M}}) \right\rangle_{\text{M}} \tag{2.54a}$$

$$\varepsilon^2(z) = \left\langle (V_{\text{B}}(z,t^*_{\text{M}}) - \left\langle V_{\text{B}}(z,t^*_{\text{M}}) \right\rangle_{\text{M}})^2 \right\rangle_{\text{M}} \tag{2.54b}$$

for the free energy function estimator $F_{\text{META}}(z)$ and its standard deviation[8] $\varepsilon(z)$.On the other hand, if a simulation is continued for a long time $t_{\text{F}} \gg t^*$ after diffusivity is reached at time t^*, then the best possible estimate of the free

[8] The error tipically is only marginally larger at the boundary $\partial\Omega$ of the z domain Ω, thus it can be conveniently characterized by its average $\bar{\varepsilon}^2 = \frac{1}{\text{vol}(\Omega)} \int_\Omega \varepsilon^2(z) \text{d}z$.

energy $F(z)$ is the arithmetic average (the chronological mean)[9]

$$F_{\text{META}}(z,,t^*) = -\frac{1}{t_{\text{F}} - t^*} \int_{t^*}^{t_{\text{F}}} V_{\text{B}}(z,t)\text{d}t \quad \longrightarrow \quad - \sum_{\substack{t'=t^*+k\mathcal{T} \\ k\in\mathbb{N}}}^{t'<t_{\text{F}}} V_{\text{B}}(z,t') \quad , \qquad (2.55)$$

with standard deviation decaying to zero with the law that is determined by the autocorrelation time of the profile. The extrapolation as a sum is used to avoid profiles correlation in the integral (taking different profiles $V_{\text{B}}(z,t')$ at a time distance $\mathcal{T}$ suitable chosen to lower correlations).

[9] For technical reason, it is necessary that the dynamics of the CVs after t^* is bound in a finite region of the CVs space (see [99]).

Chapter 3

Puckering Coordinates and Free Energy reconstruction

> Eccoci giunti al capitolo terzo,
> capitolo che tagliamo perché
> tutta la storia dell'Azzeccagarbugli
> è lunga e non serve a un gran che.
>
> Oblivion
> *"I Promessi Sposi" in 10 minuti*

In this Chapter the quantitative description for puckering conformational analysis of monosaccharides is introduced. Some different definitions for puckering coordinates are briefly shown, and between these the Cremer-Pople parametrization scheme is discussed in details. Finally, the connection of the Cremer-Pople coordinates to metadynamics is indicated.

Contents

3.1 Puckering: historical development and the Strauss-Pickett coordinates

Since monosaccharides have a closed ring structure, we are led easily to the notion of a reference plane for the ring itself. A natural issue will be then whether the ring is planar or not. For this reason, the conformational analysis of monosaccharides requires some effort in the description of their intrinsically non-planar structure, *i.e.* their *puckered* structure.

Given a non-planar structure, a quantitative measure of ring puckering is a mathematical procedure that reduces the total information of the ring (*e.g.* the total coordinates $\{\mathbf{r}_j\}_{j=1,\dots,N}$ of ring atoms) to a small set of parameters. This procedure could have two main goals:

a) quantitative measure of ring puckering: the idea is to construct the minimal parameter set that encodes the non-planarity in a suitable compact way from the atomic coordinates or torsional angles;

b) quantitative comparison of puckered structure: give a quantitative estimate of how a compound is puckered, most of the times in terms of linear combination of ideal structures (like the standard IUPAC structures, see [47]).

Due to the freedom in the choice of the mathematical reduction procedure, various quantitative descriptions/parametrizations for conformational analysis of ring puckering are present in literature.

The description of non-planar, non-aromatic ring puckering has a long history. The starting point are the conformational analysis works of Sachse and coworkers[154, 155], but the first *quantitative* definition of ring puckering came with the work of Kilpatrick, Pitzer and Spitzer[89]. Their analysis of the cyclopenthane structure led to the description of puckered structures in terms of two parameters (q, F_{KPS}) (a puckering amplitude and a puckering phase angle, respectively), which are related to the out-of-plane position

$$z_j = \sqrt{2/5}\, q \cos\left(2F_{\text{KPS}} + 4\pi(j-1)/5\right) \quad , \quad j = 1, \dots, 5 \tag{3.1}$$

of the ring atoms. The (q, F_{KPS}) parametric representation allows for the first time, differently from the description with the atomic elevations z_j, the introduction of the concept of *pseudorotation* in conformational transition. Indeed, the transformation between two ring conformers is described by the change of the phase angle F_{KPS}. This change is an internal "rotation" for the system that does not give an overall angular momentum. For this reason, the term pseudorotation has been chosen.

A theoretical improvement in the subject was given by Strauss and Pickett, with their analysis of cyclohexane and related oxanes [171]. Concerning six-membered rings, their quasi-planar conformations can be parametrized with three coordinates (r, Θ, Φ), by means of which the six out-of-plane positions

$$Z_j = r\left[\frac{1}{2}(-1)^j \cos\Theta + \sin\Theta \cos\left(\frac{2\pi j}{3} + \Phi\right)\right] \quad , \quad j = 1, \dots, 6 \tag{3.2}$$

of ring atoms are uniquely defined. In other words, the only three independent ways to perform infinitesimal deformation of a six-membered ring can be encoded in the parameters (r, Θ, Φ), while the elevations $\{Z_j\}_{j=1,\dots,6}$ represent the

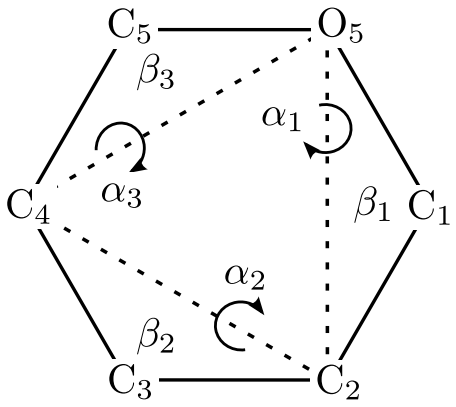


Figure 3.1: Definition of the internal (virtual) dihedral angles
$\alpha_1 \equiv C_4 - O_5 - C_2 - C_1$,
$\alpha_2 \equiv O_5 - C_2 - C_4 - C_3$,
$\alpha_3 \equiv C_2 - C_4 - O_5 - C_5$
and of the (real) bond angles
$\beta_1 \equiv O_5 - C_1 - C_2$,
$\beta_2 \equiv C_2 - C_3 - C_4$,
$\beta_3 \equiv C_4 - C_5 - O_5$.

vertical displacement of the i-th atom from the planar structure. The authors of Ref. [143] showed that this special $N = 6$ case is in agreement with the generic case of $N-3$ coordinates that could be constructed for N-membered rings from group theory analysis on ring deformations. The main ideas of group theory analysis on closed rings are the following. The infinitesimal displacement of an atom is only an elevation Z_j from the planar ring. Thus, N out-of-plane coordinates are sufficient to describe the resulting conformation. Group theory allows then to rewrite these out-of-plane elevations $\{Z_j\}_{j=1,\ldots,N}$ to new coordinates, by means of the symmetries of the ring itself. A set of new "symmetry-adapted" coordinates is then produced from irreducible representation of of the symmetry group. Overall translations and rotations of the system, that are not relevant at this stage, can be eliminated and, eventually, we can reduce the description from the N elevations $\{Z_j\}_{j=1,\ldots,N}$ to $N-3$ "symmetry-adapted" puckering parameters for an N membered ring.

Since ring conformations are most of the time far from being planar, the description of Eq. (3.2) has to be slightly changed to take properly into account the structural constraints given by bond lengths and bond angles. The authors of Ref. [171] proposed to start from the internal coordinates $\{\alpha_i, \beta_i\}_{i=1,2,3}$ (see Fig. 3.1 for their definitions) rather that from $\{Z_j\}_{j=1,\ldots,6}$ elevations. This is because changing dihedral angles $\{\alpha_i\}_{i=1,2,3}$ does not produce artificial stretching in bond length, as can be done with non-infinitesimal changes in atom vertical displacement. Furthermore, changes in $\{\beta_i\}_{i=1,2,3}$ angles can be generally omitted, or treated as constraints. Thus, from these angles it is possible to produce again a set (r', Θ', Φ') similar to the previous one, using the definitions

$$\alpha_i = r' \left[\cos\Theta' + 2\sin\Theta' \cos\left(\frac{4}{3}\pi i - \frac{2\pi}{3} + \Phi' \right) \right] \quad , \quad i = 1, 2, 3 \quad . \qquad (3.3)$$

By means of Eq. (3.3), both the sets $\{\alpha_i\}_{i=1,2,3}$ and $\{r, \Theta, \Phi\}$ can be used as the so-called Strauss-Pickett (SP) puckering coordinates. Due to the similar characteristics in structure, these coordinates can be intended to be valid not only for cyclohexane but also for generic six-membered rings.

3.2 The general Cremer-Pople coordinate set

The first general and systematic definition of puckering parameters for a generic N-membered ring was the work of Cremer and Pople [41]. The authors firstly stated a set of equations that uniquely define a reference plane, and then the definition of the reduced set of $N-3$ puckering coordinates from the $3N$ total Cartesian coordinates. Further works [26, 51, 50, 40] completed the theoretical framework and its application to conformational analysis.

In the following, some notions will be given to introduce the Cremer-Pople (CP) formulae with two main purposes: present the framework of this thesis work, and settle some of the formalism that will be used further.

3.2.1 Mean plane and molecular axes definitions

Consider a generic N-membered ring (with non necessarly equal atoms), where $\left\{\mathbf{r}_j = (r_x^j, r_y^j, r_z^j)\right\}_{j=1,\dots,N}$ are the atomic coordinates in an arbitrary reference frame. All the coordinates can be simply translated to the *geometrical center* of the ring

$$\mathbf{R} = \frac{1}{N}\sum_{j=1}^{N}\mathbf{r}_j \tag{3.4}$$

with respect of which the atomic positions now reads

$$\mathbf{R}_j = \mathbf{r}_j - \mathbf{R} = \sum_{i=1}^{N}\Delta_{ij}\mathbf{r}_i \quad , \quad \Delta_{ij} = \delta_{ij} - \frac{1}{N} = \begin{cases} (N-1)/N & \text{if } i = j \\ -1/N & \text{if } i \neq j \end{cases} \tag{3.5}$$

where δ_{ij} is the usual Kronecker symbol. For these positions the equation

$$\sum_{j=1}^{N}\mathbf{R}_j = \mathbf{0} \tag{3.6}$$

holds automatically. A *mean plane* of the ring will be now chosen to pass through this geometric center, and the $\hat{\mathrm{n}}$ direction orthogonal to this plane will be the *molecular axis* of the system. Eventually, the N projections z_j of the ring atoms along the $\hat{\mathrm{n}}$ direction can be calculated.

In order to do so, a transformation (a global rotation $\mathfrak{R}$ of the Cartesian coordinates) of the $\mathbf{R}_j$ vector coordinates (X_j, Y_j, Z_j) into new coordinates (x_j, y_j, z_j) has to be properly defined. After the transformation, for the projections z_j the equivalence

$$\sum_{j=1}^{N} z_j = 0 \tag{3.7}$$

holds automatically[1]. Together with Eq. (3.7), we impose two additional con-

[1] If a global rotation $\mathfrak{R}$ is applied on the vectors $\mathbf{R}_j$, from Eq. (3.6) we have $\sum_{j=1}^{N}\mathfrak{R}\mathbf{R}_j = \mathfrak{R}\sum_{j=1}^{N}\mathbf{R}_j = \mathfrak{R}\mathbf{0} = \mathbf{0}$. This lead to the equivalence of Eq. (3.7) and to the same equivalence for the summation over the x_j and the y_j coordinates

ditions to the elevation z_j

$$\sum_{j=1}^{N} z_j \sin\frac{2\pi(j-1)}{N} = \sum_{j=1}^{N} z_j w_{1,j} = 0 \quad , \tag{3.8a}$$

$$\sum_{j=1}^{N} z_j \cos\frac{2\pi(j-1)}{N} = \sum_{j=1}^{N} z_j v_{1,j} = 0 \quad , \tag{3.8b}$$

where the symbols

$$w_{m,j} = \sin\frac{2m\pi(j-1)}{N} \quad , \quad v_{m,j} = \cos\frac{2m\pi(j-1)}{N} \tag{3.9}$$

are defined here for further use. Imposing Eqs. (3.8) it is assured that elevations z_j are not related each other by global rotations of the system around the geometric center of the ring. Similarly, the condition of Eq. (3.7) assures that z_j displacement are not related to global translations of the ring with respect to the mean plane. As a whole, the conditions of Eqs. (3.7) and (3.8) are sufficient to define the mean plane of the system, with respect of which atomic displacements z_j will represent genuine non-planar structures. This definition of the mean plane is general, unique and invariant against conformational transformations.

To define the molecular axis, we can start from the vectors

$$\mathbf{R}' = \sum_{j=1}^{N} \mathbf{R}_j w_{1,j} \quad , \quad \mathbf{R}'' = \sum_{j=1}^{N} \mathbf{R}_j v_{1,j} \tag{3.10}$$

that by construction belong to the mean plane[2]. In order to have the direction orthogonal to the mean plane, the cross product of $\mathbf{R}'$ and $\mathbf{R}''$ is taken, obtaining the unit vector

$$\hat{\mathrm{n}} = \frac{\mathbf{R}' \times \mathbf{R}''}{|\mathbf{R}' \times \mathbf{R}''|} \tag{3.11}$$

that defines the direction of the molecular axis. With this definitions we can calculate the projections

$$\boxed{z_j = \mathbf{R}_j \cdot \hat{\mathrm{n}}} \tag{3.12}$$

of ring atoms.

The Cremer-Pople reference frame

Concerning the coordinate transformation from the starting reference frame to the Cremer-Pople one, two orthogonal axes on the mean plane have still to be defined. In Ref. [41] the following convention is proposed. The direction of the new y coordinates will be the direction of $\mathbf{R}_1$ (the ring atom numbered as one) once projected onto the mean plane, namely

$$\hat{\mathrm{m}} = \frac{\mathbf{R}_1 - (\mathbf{R}_1 \cdot \hat{\mathrm{n}})\hat{\mathrm{n}}}{|\mathbf{R}_1 - (\mathbf{R}_1 \cdot \hat{\mathrm{n}})\hat{\mathrm{n}}|} \quad . \tag{3.13}$$

[2] Both $\mathbf{R}'$ and $\mathbf{R}''$ fulfill Eqs. (3.8) by construction

The new x direction will be the third possible orthogonal unit vector

$$\hat{\mathrm{l}} = \hat{\mathrm{m}} \times \hat{\mathrm{n}} \quad . \tag{3.14}$$

The complete coordinates of $\mathbf{R}_j$ vectors in the Cremer-Pople reference frame reads

$$x_j = \mathbf{R}_j \cdot \hat{\mathrm{l}} \quad , \quad y_j = \mathbf{R}_j \cdot \hat{\mathrm{m}} \quad , \quad z_j = \mathbf{R}_j \cdot \hat{\mathrm{n}} \quad . \tag{3.15}$$

It has to be mentioned that the definition of the $\hat{\mathrm{n}}$, $\hat{\mathrm{m}}$ and $\hat{\mathrm{l}}$ directions depends naturally on the atom numbering. Conversely, the mean plane definition from Eqs. (3.7) and (3.8) is independent on the choice of the numbering scheme. For more details on the dependences of the Cremer-Pople formulae on atom numbering see Appendix B.

3.2.2 Generalized puckering coordinates

Once transformed in the Cremer-Pople reference frame, the "planar" geometry of N-membered rings is encoded in the $2N-3$ suitable chosen (x_j, y_j) coordinates. The genuine "puckered" geometry (that means, intrinsically non-planar) is instead encoded in the sole N elevations z_j. Starting from these data, in Ref. [41] a scheme was proposed to recombine $\{z_j\}_{j=1,\ldots,N}$ elevations in such a way that only $N-3$ free parameters are left. There are two possible situations, depending on N being even or odd.

N **odd** : from z_j elevations define $\frac{N-3}{2}$ pairs (q_m, ϕ_m) with $m = 2, 3, \ldots, \frac{N-1}{2}$ by means of

$$q_m \cos\phi_m = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} z_j v_{m,j} \tag{3.16a}$$

$$q_m \sin\phi_m = -\sqrt{\frac{2}{N}} \sum_{j=1}^{N} z_j w_{m,j} \tag{3.16b}$$

where the symbols $w_{m,j}$ and $v_{m,j}$ are defined in Eq. (3.9).

N **even** : from z_j elevations define $\frac{N-4}{2}$ pairs (q_m, ϕ_m) with $m = 2, 3, \ldots, \frac{N}{2}-1$ like in Eqs. (3.16), and add the single coordinate

$$q_{N/2} = \sqrt{\frac{1}{N}} \sum_{j=1}^{N} z_j \cos[\pi(j-1)] = \sqrt{\frac{1}{N}} \sum_{j=1}^{N} z_j (-1)^{j-1} \tag{3.17}$$

For further details on the choice of m ranges in the above definitions see Appendix A. With these definitions the concept of pseudorotation, discussed in Section 3.1, is naturally extended to generic N-membered rings by means of the phase angle ϕ_m of each pair (q_m, ϕ_m). Thus, the conformational space is divided in $(N-1)/2$ *pseudorotational subspaces* (described by the (q_m, ϕ_m) pairs). When N is even also the so-called *inversion subspace* appears (described by $q_{N/2}$).

Explicitly, from Eqs. (3.16) and Eq. (3.17) the free *puckering coordinates* and their ranges are

$$q_m = \sqrt{\frac{2}{N}}\sqrt{\mathcal{A}_m^2 + \mathcal{B}_m^2} \qquad q_m \geq 0 \tag{3.18a}$$

$$\tan\phi_m = -\frac{\mathcal{A}_m}{\mathcal{B}_m} \qquad \phi_m \in [0, 2\pi) \tag{3.18b}$$

$$q_{N/2} = \sqrt{\frac{1}{N}}\mathcal{C} \qquad q_{N/2} \in \mathbb{R} \qquad (N \text{ even}) \tag{3.18c}$$

with

$$m \in I_N \quad , \quad I_N = \begin{cases} \{2, 3, \ldots, \frac{N-1}{2}\} & N \text{ odd} \\ \{2, 3, \ldots, \frac{N}{2} - 1\} & N \text{ even} \end{cases} \tag{3.19}$$

and where the dependence on the z_j projections is indicated in the symbols

$$\mathcal{A}_m = \sum_{j=1}^{N} z_j w_{m,j} \quad , \quad \mathcal{B}_m = \sum_{j=1}^{N} z_j v_{m,j} \quad , \quad \mathcal{C} = \sum_{j=1}^{N} z_j (-1)^{j-1} \tag{3.20}$$

defined for further use[3]. Besides, the so-called *total puckering amplitude*

$$Q = \sqrt{\sum_{j=1}^{N} z_j^2} = \sqrt{\sum_{m \in I_N} q_m^2 + q_{N/2}^2} \qquad Q \geq 0 \tag{3.21}$$

(where $q_{N/2}^2$ term is present only for N being even) could be defined. It is somehow a measure of the dispersion of the elevation z_j, similar to a standard deviation. The normalization factors of Eqs. (3.16) and Eq. (3.17) were chosen in such a way that the second equivalence in Eq. (3.21) holds.

As a final remark, it is worth mentioning that all Cremer-Pople definitions given above produce puckering coordinates that are always well-defined functions of the projections z_j. This is an important feature for our further manipulation.

Connection with the Strauss-Pickett scheme

There is a 1:1 correspondence between the Cremer-Pople coordinates and the "symmetry-adapted" coordinates of Strauss and Pickett. This is because both approaches are in agreement with a group theory analysis of ring deformation. As shown in some technical works [26, 51], on one hand Strauss and Pickett [143] described the possible "symmetry-allowed" displacement of planar structures; on the other hand, Cremer and Pople [41] aim was to reduce the non-planar structure to elevation from a planar ring, *i.e.* the inverse procedure. At a difference with the SP approach, for the CP formulae there are no changes between the infinitesimal and finite description of displacements. This fact, as will be shown in Section 3.3, led to some criticism even if it is mathematically rigorous.

[3] it can be seen that the re-summation $\mathcal{C} = \mathcal{B}_{N/2}$ and, in addition, that the analog term $\mathcal{A}_{N/2} = 0$ because $\sin[\pi(j-1)] = 0\ \forall j$. Thus, the definitions of Eqs. (3.16) seems to collect both the N odd and even case. However, a difference still holds in the normalization factors ($\sqrt{2/N}$ for the odd case, $\sqrt{1/N}$ for the even case), due to the condition $\mathcal{A}_{N/2} = 0$ that holds automatically in the even case.

3.2.3 Cremer-Pople coordinate inversion

Although the reduction from $3N$ Cartesian coordinates to $N-3$ puckering parameters is quite simple, the inverse procedure is less straightforward. Due to the relations of Eqs. (3.7) and (3.8), given the $N-3$ parameters (q_m, ϕ_m) and $q_{N/2}$ the elevations of the ring atoms can be calculated as

$$z_j = \sqrt{\frac{2}{N}} \sum_{m \in I_N} q_m \cos\left[\phi_m + \frac{2m\pi(j-1)}{N}\right] + \sqrt{\frac{1}{N}} q_{N/2}(-1)^{j-1} \qquad (3.22)$$

(the $q_{N/2}$ term appears only for N being even, I_N sets are defined in Eq. (3.19)).

The other $2N$ coordinates (x_j, y_j) have to be calculated using extra information other than puckering parameters. For example, the set of N bond lengths $\{b_{ij}\}_{i,j=1,\ldots,N}$ and N bond angles $\{\beta_{ijk}\}_{i,j,k=1,\ldots,N}$ have to be supplemented. Cremer [40] proposed a five step procedure to perform this reconstruction:

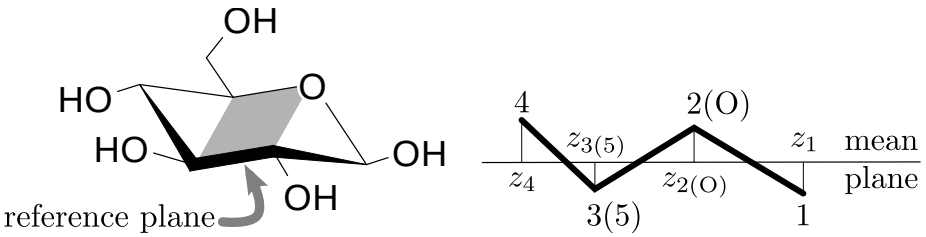
1. calculate the elevations z_j from the puckering coordinates (q_m, ϕ_m) and $q_{N/2}$ (using Eq. (3.22));
2. calculate the projections of bond length b_{ij} and bond angles β_{ijk} on the mean plane;
3. perform a geometrical partition of the planar ring;
4. calculate the atomic planar positions on the ring partition;
5. perform the full ring reconstruction.

With this procedure the full Cartesian coordinates in the Cremer-Pople reference frame are calculated. For more details about six-membered ring reconstruction, see Appendix C.

3.3 Other definitions beyond Cremer-Pople

Although the Cremer-Pople approach to conformational analysis became a *de-facto* standard for many years, the description of Cremer and Pople has been questioned from time to time [191, 64, 19, 72]. The reasons could be found basically in the relationship between the CP scheme and the puckering interpretation from stereochemistry. On one hand, given all the Cartesian coordinates of ring atoms, the CP approach builds a natural, simple framework in which the minimal number of quantitative parameters are obtained. On the other hand, this framework is intrinsically different from the stereochemical concept of ring puckering, that is intuitively based on angles between planar subdomain of the molecule itself, or on endocyclic torsional angles.

The simplest way to understand this conceptual difference is to compare in both frameworks the "reference plane" with respect to which a structure is considered puckered. In the standard stereochemical interpretation, as it stated in IUPAC recommendation [158, 47, 119], the reference plane for a basic structure, like the chair conformer of Fig. 3.2(a), always contain some ring atoms. This plane changes from one standard structure to another because is qualitatively



(a) standard stereochemical (IUPAC) reference plane for chair conformer

(b) schematic representation of the mean plane of Cremer-Pople, given for chair conformer

Figure 3.2: Different notion of ring reference plane. For the sake of simplicity, the reference plane is given in both cases for a six-membered ring in the chair-like conformation.

determined for all standard structures. Conversely, the Cremer-Pople mean plane (schematically represented in Fig. 3.2(b)) does not in principle contain any ring atom and does not change by means of conformational transitions. In other words, the Cremer-Pople scheme gives in some sense its own re-definition of the notion of puckering generally accepted in stereochemistry. This is why some authors wanted to revisit the subject in a way closer to the standard stereochemistry.

Besides this conceptual difference, some other open points remain with the use of Cremer-Pople scheme:

- the quantitative values of the CP parameters (like the total puckering amplitude Q or the elevations z_j) may not be comparable when calculated on structures with the same number of ring member but with different atom types (cfr. Ref. [191]). Indeed, bond lengths and bond angles can vary a lot when the composition changes. The absolute values of puckering parameters change correspondingly, even for structures that are qualitatively similar;

- sometimes all the Cartesian coordinates are unreachable, and then there is the necessity of puckering coordinate sets defined on dihedral angles (like in Refs. [191, 64, 19]). For example, dihedral (torsional) angles could be extracted by experimental data almost directly, while to recover all the Cartesian coordinates from data or from puckering coordinates may need much more effort and assumptions;

- computing (q_m, ϕ_m) and $(q_{N/2})$ values gives a purely quantitative measure of puckering. In some sense, an "absolute" measure of puckering is provided, without any natural notion of the distance of a given conformation with respect to the others. In other words, a comparison between a given structure and some standard conformers (somehow a "relative" measure of puckering, in terms of quantitative coefficients which measure how two structures are "close" or "far") is not straightforward. Conversely, other approaches were constructed directly for this latter purpose (like the one of Berces et al. [19]).

In our context of computer simulations, the choice of the description scheme for

puckering properties is simplified by the direct acces to the full Cartesian coordinates. Thus, in principle every parametrization scheme is accessible. Nonetheless, the advantages of the CP scheme are still striking:

- ✓ it is a framework with simple, general and rigorous formulae
- ✓ it is widely used and quite simple to implement in computer programs;
- ✓ from computer simulations all the Cartesian coordinates are directly obtained. Moreover, every missing information to make *a posteriori* ring reconstruction from the puckering parameters can be simply collected, if needed;
- ✓ in the context of accelerated sampling methods (like metadynamics) the Cremer-Pople parameters are relatively simple in handling both calculations of the collective variables and in their gradients;
- ✓ monosaccharides transitions are well represented in CP coordinates, particularly the pseudorotation and inversion transformations (as will be shown in Section 3.4.2);
- ✓ in our work the comparison between different molecules is simplified by means of a general uniformity in ring structures (most of the time are C_5O rings). In this context there are no problems related to the comparison of absolute values of the puckering parameters (like those discussed by Zefirov et al. [191]).

3.4 Derivation of Cremer-Pople coordinates for six-membered rings

The Cremer-Pople description for six-membered rings is simply derived from the general formulae. However, some manipulations on the native CP formulae are presented, in order to provide much more insight for conformational analysis of monosaccharides.

3.4.1 Original Cremer-Pople coordinate representation

With $N = 6$ there are 3 puckering coordinates that spans the phase space of conformational structures of six-membered rings. Thus, the conformational space will be an open set $\Omega \subset \mathbb{R}^3$ and, according to the formulae of Section 3.2.2, we are in the even case with

- a pseudorotational coordinate pair (q_2, ϕ_2);
- an inversion coordinate q_3;

to describe it. From Eqs. (3.18) we have

$$q_2 = \sqrt{(\mathcal{A}_2^2 + \mathcal{B}_2^2)/3} \qquad q_2 \geq 0 \quad , \tag{3.23a}$$

$$\tan\phi_2 = -\mathcal{A}_2/\mathcal{B}_2 \qquad \phi_2 \in [0, 2\pi) \quad , \tag{3.23b}$$

$$q_3 = \mathcal{C}/\sqrt{6} \qquad q_3 \in \mathbb{R} \quad , \tag{3.23c}$$

as functions of the re-summation $\mathcal{A}_2$, $\mathcal{B}_2$ and $\mathcal{C}$ of the z_j elevations from the mean plane (see Eq. (3.20) for their definitions). Looking at the structure of the coordinates above, the Cremer-Pople coordinates are a *cylindrical* coordinate set (q_2, ϕ_2, q_3) for $\Omega \subset \mathbb{R}^3$, where q_2 is the planar radius, ϕ_2 is the phase angle and q_3 is the elevation from the $q_3 = 0$ plane.

3.4.2 Alternative representation: the "puckering sphere"

Given the cylindrical set (q_2, ϕ_2, q_3), a transformation to a *spherical* (Q, θ, ϕ) representation is not only possible but also straightforward. Indeed, the total puckering amplitude

$$Q = \sqrt{\sum_{j=1}^{6} z_j^2} = \sqrt{q_2^2 + q_3^2} \qquad , \qquad Q \geq 0 \tag{3.24}$$

will be naturally the radius of the spherical set. The colatitude θ is defined by

$$\begin{cases} q_2 = Q \sin\theta \\ q_3 = Q \cos\theta \end{cases} \tag{3.25}$$

where the original puckering amplitudes $q_2 \geq 0$ and $q_3 \in \mathbb{R}$ are naturally associated with the sine and the cosine of an angle $\theta \in [0, \pi]$, respectively, while the Q factor preserves the normalization of Eq. (3.24). The last spherical coordinate, the longitude ϕ, will be simply the original phase angle ϕ_2. Explicitly, the new representation reads

$$Q = \sqrt{\frac{2(\mathcal{A}_2^2 + \mathcal{B}_2^2) + \mathcal{C}^2}{6}} \qquad Q \geq 0 \tag{3.26a}$$

$$\theta = \arctan\left[\frac{\sqrt{2(\mathcal{A}_2^2 + \mathcal{B}_2^2)}}{\mathcal{C}}\right] \qquad \theta \in [0, \pi] \tag{3.26b}$$

$$\phi = \arctan\left[-\mathcal{A}_2/\mathcal{B}_2\right] \qquad \phi \in [0, 2\pi) \tag{3.26c}$$

again as functions of the original re-summation $\mathcal{A}_2$, $\mathcal{B}_2$ and $\mathcal{C}$ of the z_j elevations from the mean plane.

Eventually, also a *Cartesian* (q_x, q_y, q_z) representation could be obtained, with the definitions

$$q_x = \mathcal{B}_2/\sqrt{3} \qquad q_x \geq 0 \tag{3.27a}$$

$$q_y = -\mathcal{A}_2/\sqrt{3} \qquad q_y \geq 0 \tag{3.27b}$$

$$q_z = \mathcal{C}/\sqrt{6} \qquad q_z \in \mathbb{R} \tag{3.27c}$$

again as functions of the original re-summation $\mathcal{A}_2$, $\mathcal{B}_2$ and $\mathcal{C}$ of the z_j elevations from the mean plane. It is interesting to observ that the original formulae of Eq. (3.16) give directly the Cartesian coordinates (q_x, q_y, q_z). Thus, also the Cartesian representation of puckering coordinates is naturally present in the original CP definitions.

In Table 3.1 the three derived coordinate sets are summarized. Clearly, every other rewriting of the Cremer-Pople formulae gives a well-defined parameter set.

Table 3.1: Standard Cremer-Pople puckering coordinate set (summary). The coordinates from the original definition are the *cylindrical* ones.

Cremer-Pople coordinates ($N = 6$) interconversion		
cylindrical (q_2, ϕ_2, q_3)	spherical (Q, θ, ϕ)	Cartesian (q_x, q_y, q_z)
$q_2 \cos \phi_2$	$Q \sin \theta \cos \phi$	q_x
$q_2 \sin \phi_2$	$Q \sin \theta \sin \phi$	q_y
q_3	$Q \cos \theta$	q_z

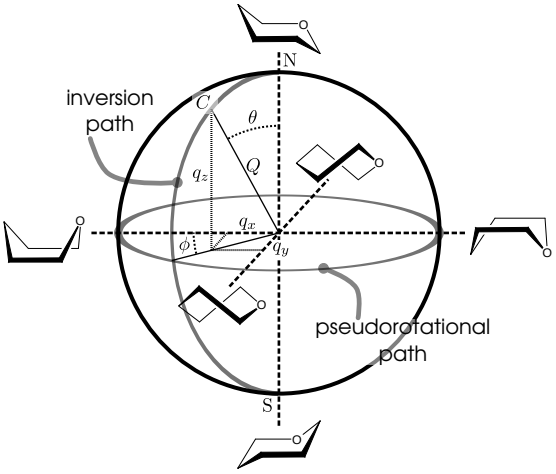


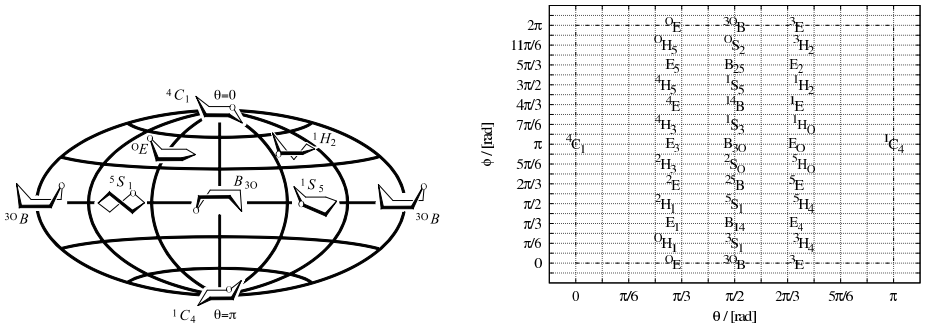
Figure 3.3: The Cremer-Pople coordinates and the puckering sphere. A conformation C can be localized on the surface of a sphere of (quasi) constant radius Q. The main transition pathway between conformation are indicated, along the pseudorotational and the inversion subspaces.

However, the spherical set is the one that matches more clearly the so-called *puckering sphere* concept. Indeed, the total puckering amplitude Q is a quantity that does not vary too much for different puckered structures. This is because bond lengths and bond angles are generally almost fixed, or will show only very small variation around their equilibrium values. Thus, the conformational transitions are well represented on the surface of a sphere with (quasi) constant radius[4] and this naturally leads to the use of the (Q, θ, ϕ) set. Moreover, in this way not only the variations of longitude ϕ follow naturally the pseudorotational path of flexible ring conformers, but also the variations of the colatitude θ are closer to the real inversion paths than the original inversion coordinate q_3 (see Fig. 3.3).

As a conclusive remark, we want to show the localization of ideal ring conformations defined by the IUPAC standards (see Fig. 1.5) on the puckering sphere. Maps like the one proposed in Fig. 3.4 could be used to complete somehow the quantitative conformational description with qualitative indications. For example, it is possible to compare the actual structure with vicinal ideal structures, or to follow a conformational transition path and describe it in terms of standard intermediate conformers. It is worth underlining that in the Cremer-Pople

[4]More precisely, it can be shown that a six-membered ring with fixed bond lengths and bond angles with harmonic constraints covers a spherical shell, see Section 4.1.1 for details

scheme this kind of localization of standard structures is quite simpler than in other puckering scheme, like the Strauss-Pickett approach (a detailed discussion of this aspect will be given in Section 5.2).



(a) Hammer-Aitoff (equal area) representation. Examples of ring conformers are approximately located at their position on the representation. Parallels represent pseudorotational paths, meridians represent inversion paths.

(b) Plate Carrée (equirectangular) representation. The 38 ideal conformers (stable chairs C, flexible boats B and skews S, transition halfchairs H and envelopes E) are indicated, with (Q, θ, ϕ) values taken from Table 1 in Ref. [72]). Chair conformers $^4C_1$ and $^1C_4$ are located along the whole segments at $\theta = 0$ and $\theta = \pi$, respectively.

Figure 3.4: Ideal conformers on the puckering sphere. The positions of ideal IUPAC conformer are indicated on a Hammer-Aitoff projection (left) and on a *Plate Carrée* (from French, for "square plate") projection of the puckering sphere (right).

3.5 Puckering properties with metadynamics

We now want to use the Cremer-Pople description within metadynamics in order to provide a free energy reconstruction method for conformational transition of monosaccharides. Indeed, if the free energy $F(z)$ is known, thermodynamic properties (like free energy differences, free energy barriers, conformers populations, ...) could be calculated and also reaction pathways could be investigated. Cremer-Pople parameters have in principle the characteristic to be "good" collective variables in the sense of Section 2.2.3: they are able to clearly distinguish between different states, and only a small number of parameters are neccessary to do it. In addition, the considerations given in Section 3.4.2 indicate that the Cremer-Pople coordinates, expecially in the form of spherical coordinates, are "close" to the transition pathways, namely that a single coordinate change spans a single conformational transition pathway.

For the metadynamics algorithm, Eqs. (3.26) defines the variables (Q, θ, ϕ). Then, we are interested in the expressions for the gradients in the (Q, θ, ϕ) representation, that means

$$\nabla_i Q \quad , \quad \nabla_i \theta \quad , \quad \nabla_i \phi \quad , \tag{3.28}$$

where $\nabla_i = \left(\dfrac{\partial}{\partial X_i}, \dfrac{\partial}{\partial Y_i}, \dfrac{\partial}{\partial Z_i} \right)$ is the gradient with respect to the i-th ring atom

coordinates ($i = 1, \ldots, 6$) given in the reference frame[5] in which the atomic positions $\mathbf{R}_j = (X_j, Y_j, Z_j)$ were given. By direct calculation, we have:

$$\nabla_i Q = \frac{1}{Q} \sum_{j=1}^{6} z_j \nabla_i z_j \quad , \tag{3.29}$$

$$\nabla_i \theta = \frac{1}{3\sqrt{2}Q^2} \frac{1}{\sqrt{\mathcal{A}_2^2 + \mathcal{B}_2^2}} \Big[\mathcal{C} \left(\mathcal{A}_2 \nabla_i \mathcal{A}_2 + \mathcal{B}_2 \nabla_i \mathcal{B}_2 \right) - (\mathcal{A}_2^2 + \mathcal{B}_2^2) \nabla_i \mathcal{C} \Big] \quad , \tag{3.30}$$

$$\nabla_i \phi = \frac{\mathcal{A}_2 \nabla_i \mathcal{B}_2 - \mathcal{B}_2 \nabla_i \mathcal{A}_2}{\mathcal{A}_2^2 + \mathcal{B}_2^2} \tag{3.31}$$

(see Appendix A for details). As the symbols $\mathcal{A}_2$, $\mathcal{B}_2$ and $\mathcal{C}$ are functions of the atomic elevations z_j, their gradients are explicitly

$$\nabla_i \mathcal{A}_2 = \sum_{j=1}^{6} w_{2,j} \nabla_i z_j \quad , \quad \nabla_i \mathcal{B}_2 = \sum_{j=1}^{6} v_{2,j} \nabla_i z_j \quad , \quad \nabla_i \mathcal{C} = \sum_{j=1}^{6} (-1)^{j-1} \nabla_i z_j \quad . \tag{3.32}$$

This shows that in all the the derivatives of the (Q, θ, ϕ) coordinates we have a re-summation of the terms $\nabla_i z_j$. This is true also for gradients of the puckering coordinates in the other representation of Table 3.1 (see Appendix A for details).

Thus, the terms $\nabla_i z_j$ are needed for analytical gradients of Cremer-Pople coordinates. Going back to the definition of Eq. (3.12), this means to calculate

$$\begin{aligned} \nabla_i z_j &= \nabla_i \left(\mathbf{R}_j \cdot \frac{\mathbf{R}' \times \mathbf{R}''}{|\mathbf{R}' \times \mathbf{R}''|} \right) \\ &= -\frac{z_j}{2|\mathbf{R}' \times \mathbf{R}''|^2} \nabla_i \big[(\mathbf{R}' \times \mathbf{R}'') \cdot (\mathbf{R}' \times \mathbf{R}'') \big] + \frac{\nabla_i \big[\mathbf{R}_j \cdot \left(\mathbf{R}' \times \mathbf{R}'' \right) \big]}{|\mathbf{R}' \times \mathbf{R}''|} \end{aligned} \tag{3.33}$$

that is explicitly given as

$$\begin{aligned} &\nabla_i \big[(\mathbf{R}' \times \mathbf{R}'') \cdot (\mathbf{R}' \times \mathbf{R}'') \big] = \\ &\quad = 2\mathcal{E}_i \left[\mathbf{R}' \left| \mathbf{R}'' \right|^2 - \mathbf{R}''(\mathbf{R}'' \cdot \mathbf{R}') \right] + 2\mathcal{F}_i \left[\mathbf{R}'' \left| \mathbf{R}' \right|^2 - \mathbf{R}'(\mathbf{R}'' \cdot \mathbf{R}') \right] \quad \text{(3.34a)} \\ &\nabla_i \big[\mathbf{R}_j \cdot \left(\mathbf{R}' \times \mathbf{R}'' \right) \big] = \Delta_{ij} \mathbf{R}' \times \mathbf{R}'' + \mathcal{E}_i \mathbf{R}'' \times \mathbf{R}_j + \mathcal{F}_i \mathbf{R}_j \times \mathbf{R}' \quad \text{(3.34b)} \end{aligned}$$

(see Appendix A for details) where the symbols

$$\Delta_{ij} = \delta_{ij} - \frac{1}{6} = \begin{cases} 5/6 & i = j \\ -1/6 & i \neq j \end{cases} \quad , \quad \mathcal{E}_i = \sum_{l=1}^{6} \Delta_{il} w_{1,l} \quad , \quad \mathcal{F}_i = \sum_{k=1}^{6} \Delta_{ik} v_{1,k} \tag{3.35}$$

are defined for simplicity[6].

With all these building blocks it is possible to calculate analytical gradients for the set (Q, θ, ϕ). All the gradients presented here are in agreement with

[5] he general translation to the geometrical center does not affect the derivatives.

[6] The symbols Δ_{ij}, $\mathcal{E}_i$ and $\mathcal{F}_i$ have the special $N = 6$ value because here we are interested in six-membered rings. However, the calculation reported are valid for generic N, and thus also the expression for $\nabla_i z_j$ is general (see Appendix A).

the analytical gradients for the generic set (q_m, ϕ_m) and $q_{N/2}$ given by Cremer [40]. Nevertheless, it is worth mentioning that the actual calculations are in some sense new. Indeed, the authors of Ref. [40] were only interested in gradients of the puckering coordinates with respect to the Cartesian coordinates in the Cremer-Pople reference frame. This means that in Ref. [40] the gradient $\nabla_i^{\text{CP}} = \left(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial y_i}, \frac{\partial}{\partial z_i} \right)$ for the i-th atom were calculated. Since the puckering parameters depends only on the set $\{z_j\}_{j=1,\dots,N}$ of elevations, Cremer analytical gradients reduces only to the derivative $\frac{\partial}{\partial z_i}$ of puckering coordinates. On the contrary, in our actual calculation we need the gradients ∇_i with respect to an arbitrary reference frame. This is necessary because all metadynamics calculation will be performed in an arbitrary reference frame, even if the Cremer-Pople parameters are calculated in their own reference frame. This lead to the explicit calculation of the gradient $\nabla_i z_j$ shown above. The relationship between the Cremer gradients and the calculation proposed here is simply the chain rule on derivation, namely:

$$\nabla_i \mathcal{X} = \frac{\partial \mathcal{X}}{\partial z_j} \nabla_i z_j \quad , \quad \mathcal{X} = q_m, \phi_m, q_{N/2} \quad . \tag{3.36}$$

Software implementation of puckering coordinates

All the simulations of this thesis were performed with the software packages GROMACS [179] and NAMD [142] for molecular dynamics[7], and with the software packages GROMETA [32, 20, 111] and PLUMED [27] for metadynamics calculations[8]. Two different MD engines were used for united-atom simulations (GROMETA and GROMACS) and for all-atom simulation (NAMD). As indicated in Section 2.2.4, to practically perform metadynamics the implementation of suitable routines in an MD engine is needed. This means the calculation of $\xi(r)$ and $\nabla_{\mathbf{r}_i} \xi(r)$, where $\xi(r)$ will be for our purposes the Cremer-Pople coordinates already described. The specific code for puckering coordinates (in particular, for the spherical and Cartesian Cremer-Pople coordinates) was written by the authors of Ref. [159] in the GROMETA/PLUMED source code. The code for Cremer-Pople spherical representation is now part of the official version of the PLUMED plugin since the 1.1.0 version of the software package.

[7] Available at `http://www.gromacs.org/` and `http://www.ks.uiuc.edu/Research/namd/` .

[8] Available at `http://www.mi.infn.it/~provasi/grometa/Site/Welcome.html` and `https://sites.google.com/site/plumedweb/` .

Chapter 4

Metadynamics for six-membered rings

> Quando il gioco si fa duro, io vorrei essere da un'altra parte
>
> Paperinik
> PKNA#0 - Evroniani

In this Chapter the application of Metadynamics to free energy reconstruction of six-membered ring is described. Before applying it to hexopyranoses, we test our algorithm which uses puckering coordinates as collective variables. Following these tests, some refinements to the basic algorithm and some additional collective variables are proposed.

Contents

4.1 Toy-models for metadynamics and puckering coordinates

The application of metadynamics to six-membered ring conformations is a relatively new subject in the field of carbohydrate simulations. In order to calibrate the usage of metadynamics with puckering collective variables for hexopyranoses, some tests could (and should) be done. In particular, we need to verify the ability of the puckering variables to represent the relevant states of the system (*i.e.* to distinguish between states and to capture all the slow degrees of freedom), but also the ability of the algorithm itself to work properly with the chosen variable (*i.e.* to reach "convergence" in the reconstruction). For test purposes, the use of simplified structures is helpful. As far as conformational

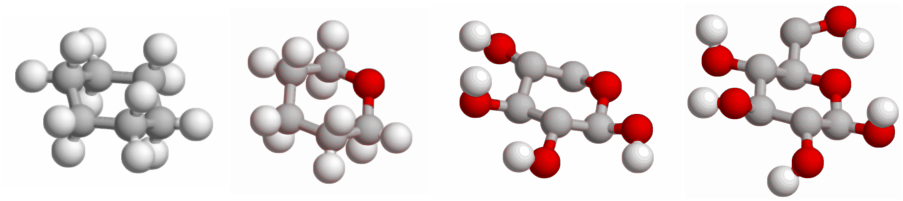


Figure 4.1: Hierarchical "construction" of hexopyranoses. From left to right: cyclohexane (C_6H_{12}), tetrahydropyran ($C_5H_{10}O$), β-D-Xylopyranose ($C_5H_{10}O_5$) and β-D-Glucopyranose ($C_6H_{12}O_6$).

properties of six-membered rings are concerned, the simplest system to study is cyclohexane (C_6H_{12}). The backbone of the ring is made by carbon atoms, and the ring substituents are all simple hydrogen atoms. Another similar structure is tetrahydropyran ($C_5H_{10}O$), where a ring carbon atom is substituted by an oxygen atom, as in hemiacetal/hemiketal ring closure of monosaccharides (see Section 1.2). Adding one $-OH$ group on this backbone to ring carbons one obtains poly-hydroxyl structures. One example of them is the pentopyranose β-D-Xylopyranose (a stereoisomer of $C_5H_{10}O_5$), where carbon atoms from 1 to 4 are substituted with an hydroxyl group. Eventually, hexopyranoses are somehow obtained from pentopyranose with an extra $-CH_2OH$ group at carbon atom C5. On the previous example, this extra substitution leads to β-D-Glucopyranose (a stereoisomer of $C_6H_{12}O_6$). To some extent, the presented hierarchical construction of hexopyranoses (for a pictorial view see Fig. 4.1) could guide in testing features and issues of metadynamics algorithm for hexopyranoses: at the simplest level the proper recognition of ring backbone structure, with and without etheroatoms in the ring; then, the influence of simple side group ($-OH$) on the algorithm could be investigated; eventually, the effect of a side chain ($-CH_2OH$) with its proper rotameric structures could be included.

4.1.1 Accessible regions in puckering space

We start with a simplified united-atom description of cyclohexane (see Fig. 4.2). For a toy-model we could in principle use an arbitrary, even non-physical parametrization to describe its topology. Obviously, there is no need to select completely unphysical parameters for Lennard-Jones, bond stretching (2-body)

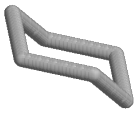


Figure 4.2: Toy-model N.1 . United-atom flexible cyclohexane (bond length $l_{CC} = 0.154\,\text{nm}$, bond angle $\alpha_{CCC} \simeq 109.4°$).

and bond angle bending (3-body) energy terms. This is because at the atomic level bond length and bond angles are almost fixed[1] around their equilibrium values, then we are free to omit the torsional (4-body) interactions. Torsional terms provide the physical description of the flexibility of the ring structure with respect to the rotation around the covalent bonds of the ring skeleton, and we could avoid it in a toy-model.

MD parameters	
time-step	0.2 fs
simulation time	~ 100 ns
integrator	sd (Langevin)
friction coeff.	$1\,\text{ps}^{-1}$
ref-T	300 K
non-bonded cutoff	0.7 nm
ε_r	80

no PBC, COM motion remotion (roto-translation), constrained bonds (SHAKE) [153]

Panel 4.1: Simulation Parameters for united-atoms toy-models. Parameters for the reported MD simulations of a single molecule in vacuum, performed with the GROMACS [179] and PLUMED [27] software (see Section 3.5).

In Fig. 4.3 the points of a free dynamic trajectory of such a flexible compound are represented in the puckering space, while details of the computational protocol are in Panel 4.1. As it can be seen, the available conformations cover only a thin spherical shell of the puckering space, with a limited range of values for the radius (the total puckering amplitude Q). Only for the poles and the equatorial region a slight increment in the Q parameter is registered. Moreover, the center of the puckering sphere is never visited, as it is a six-membered ring with fixed bond and slightly variable bond angles. With these constraints, the structure could not reach the planar conformation (located at the origin $(0, 0, 0)$ in the puckering phase space) by geometrical restrictions.

This confirms the considerations reported in Section 3.4.2, and the idea that the only relevant degrees of freedom for puckering transformation are the Cremer-Pople angles (θ, ϕ). The last statement is also supported by a simple extra test: in Fig. 4.4 the same simulation as before is showed, with the only difference that all harmonic potentials on bond angles have been removed. Now the planar conformation is clearly accessible and the exploration of the radial direction of the puckering space is possible. However, the energy terms on bond angles are unavoidable for realistic structures. Thus, geometrical constraints on carbon linkages and on the spatial direction of covalent bonds give the indication that the "real" conformational space is an accessible subset $\Omega \subset \mathbb{R}^3$ similar to the one presented in Fig. 4.3.

[1] In simulations bond lengths are constrained using the SHAKE algorithm [153], while bond angles varies with an harmonic potential.

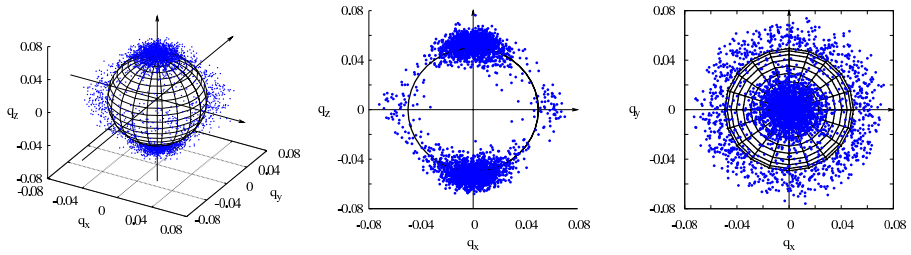


Figure 4.3: Toy-model N.1: visited conformations. Projection on the puckering sphere of the visited puckering conformations for flexible cyclohexane. Left: three dimensional view. Center: side view ($(q_x q_z)$-plane) of a thin slice around $q_y = 0\,\mathrm{nm}$. Right: top view ($(q_x q_y)$-plane) of the norther emisphere ($q_z > 0$, the southern emisphere is similar). Solid lines are a regular graticule on a sphere of radius $Q = 0.05\,\mathrm{nm}$ used as a guide for the eye. The conformation where sampled as (Q, θ, ϕ) data and converted to (q_x, q_y, q_z) coordinates.

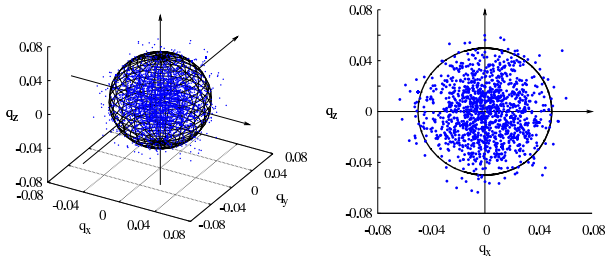


Figure 4.4: Toy-model N.1 revisited: visited conformations. Projection on the puckering sphere of the visited puckering conformation for flexible cyclohexane without any harmonic potential on bond angles. Left panel: three dimensional view. Right panel: side view ($(q_x q_z)$-plane) of a thin slice around $q_y = 0\,\mathrm{nm}$. The solid lines are a regular graticule on a sphere of radius $Q = 0.05\,\mathrm{nm}$ used as a guide for the eye. The sampling conditions were the same to the one used in Fig. 4.3

These tests suggest in some sense the following interpretation for the contribution to the total ring puckering. On one hand, we have the contribution from the intrinsic backbone properties, related to the sole ring closure and ruled by geometrical[2] constraint. This could be seen as an "entropic" contribution to puckering. On the other hand, there are features related to the stereochemistry of the compound, that are dictated by the real interactions. This could be seen as an "energetic" contribution to puckering, and the presented toy-model had removed explicitly this contribution.

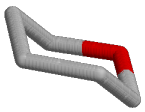


Figure 4.5: Toy-model N.2 . United-atom, flexible and uncharged tetrahydropyran (bond length $l_{CC} = 0.154\,\text{nm}$ and $l_{CO} = 0.143\,\text{nm}$, bond angle $\alpha_{CCC} \simeq 109.4°$ and $\alpha_{OCC} = \alpha_{COC} \simeq 109.5°$).

4.1.2 Ring simmetries and Cremer-Pople representation

Another possible toy-model is a simplified united-atom tetrahydropyran (see Fig. 4.5). The same considerations given for the previous model hold here, thus Lennard-Jones, bond lenght and bond angle interactions are chosen from realistic structure. At the same time, as long as the simulation is performed in vacuum an uncharged (partial charges on ring oxigen and on related atoms set to zero) and flexible (no torsional interaction) model united-atom tetrahydropyran is considered (MD details are listed in Panel 4.1).

In Fig. 4.6 the free dynamics is "projected" on the puckering space as before. The presence of an etheroatom does not change the thin shell aspect of the accessible conformations. However, there is a clear modification in the pattern of visited conformations. The most evident change is that the exact symmetry between the population around the puckering sphere is lost. Instead, a displacement in opposite direction is registered at poles, followed by a consistent change in visited conformations towards the equator. This is in agreement with the presence of the ring oxygen: the ring skeleton is no more completely symmetric, and the accessible puckering space is a spherical shell with an asymmetric distribution of sampled points, accordingly.

These symmetry issues are worth spending a few words. For its high symmetry, cyclohexane and tetrahydropyran puckering structures are much less than the 38 ideal structures presented before for hexopyranoses (cfr. Fig. 1.5). Indeed, in cyclohexane all carbons are in principle chemically indistinguishable. In tetrahydropyran the situation is almost the same, except that a ring orientation is selected by the ring oxygen (the O atom could be above or below the mean plane without ambiguity given by symmetry transformations). Thus from stereochemistry the possible stable structures are 6 conformers, namely

- two chairs: ${}^{i}C_{i+3}$ and ${}^{i+3}C_{i}$ (for cyclohexane are connected by a simple rotation around the molecular axes);
- two boats: ${}^{i(i+3)}B$ and $B_{i(i+3)}$ (for cyclohexane are connected by a simple spatial reflection);
- two skew-boats: ${}^{i}S_{i+2}$ and ${}^{i+2}S_{i}$ (for cyclohexane are connected by a spatial reflection);

As long as single molecules are concerned, real (stereochemical) differences between the 38 ideal conformers appear when the ring carbon atoms become chiral, namely when ring substituents different than simple $-$H are present. Conversely, for the Cremer-Pople coordinates point of view, the atoms in the ring structure are all different, and somehow chiral. The labelling from 1 to

[2]Strikly speaking these effects are reproduced by both purely geometric constraints (fixed bond lenght) and energetic terms (almost fixed bond angles, Lennard-Jones repulsion of atoms). However, the global effect is basically a constraint in the ring geometry.

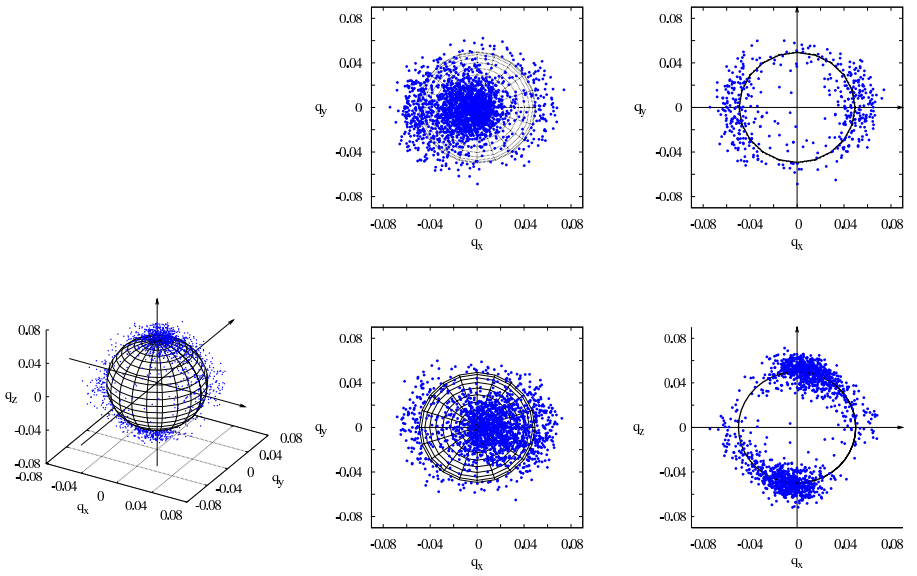


Figure 4.6: Toy-model N.2: visited conformations. Projection on the puckering sphere of the visited puckering conformation for flexible, uncharged tetrahydropyran. Left: three dimensional view. Center: north pole (down) and south pole (up). Right: sections around $q_y = 0\,\text{nm}$ ($(q_x q_z)$-plane view, down) and around $q_z = 0\,\text{nm}$ ($(q_x q_y)$-plane view, up). The simulation protocol is the same used for the test of Fig. 4.3

N serves for distinguishing the atoms in order to define the weights in the definition of the puckering parameters (as shown in Chapter 3). We remind the reader to Appendix B for a detailed description of the behavior of Cremer-Pople coordinates against the numbering scheme. Here we want to stress that with highly symmetric structures, the Cremer-Pople description artificially distinguishes between structure that are identical for their stereochemistry. In this sense, all the 38 ideal conformers turn to be separately sampled for cyclohexane and tetrahydropyran. It is clear, for example, from Fig. 4.3, where the unique chair conformer is sampled in two regions as a chair (north pole) and an inverted chair (south pole). This artificial redundancy of puckering structure is a good test for the proper behavior of our code implementation. The sampled structures for cyclohexane in Fig. 4.3 fulfill both the ring symmetries ShN_n, that shift the atom numbering by n position in clockwise direction, and InN, that invert the direction of numbering from clockwise to counterclockwise (their description is given in details in Appendix B). On the contrary, the sampled structures for tetrahydropyran in Fig. 4.6 fulfill only the symmetry InN, because the ring oxygen suppresses the ShN_n symmetry.

4.1.3 Reconstruction patterns and free energy profiles

The following results are from metadynamics calculations on all-atom models of cyclohexane and tetrahydropyran (the structures on the left and on the right of Fig. 4.7, respectively). The parametrization of the compounds was taken from

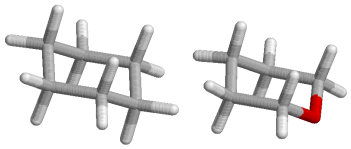


Figure 4.7: All-atoms models. Cyclohexane (left) and Tetrahydropyran (right)

the CHARMM force field [61], with 330 TIP3P [83] water molecules (all the other details are listed in Panel 4.2).

MD parameters	
time-step	0.2 fs
box size	~ 2 nm
integrator	md (leap-frog)
VdW cutoff	13.5 Å
Coulomb	PME [110, 164]
T-coupling	Langevin
T-ref, τ_T	300 K , 10 ps^{-1}
P-coupling	Nosé-Hoover Langevin [117, 52]
p-ref, τ_p	1.013 bar , 0.5 ps

PBC, COM motion remotion (translation), constrained bonds (SHAKE [153]), LJ correction

Metadynamics parameters	
CV (ξ^1, ξ^2)	(θ, ϕ)
σ_1, σ_2	0.05 rad , 0.05 rad
τ_G	200 step
w	0.15 kcal/mol
#Gaussians	~ 120000

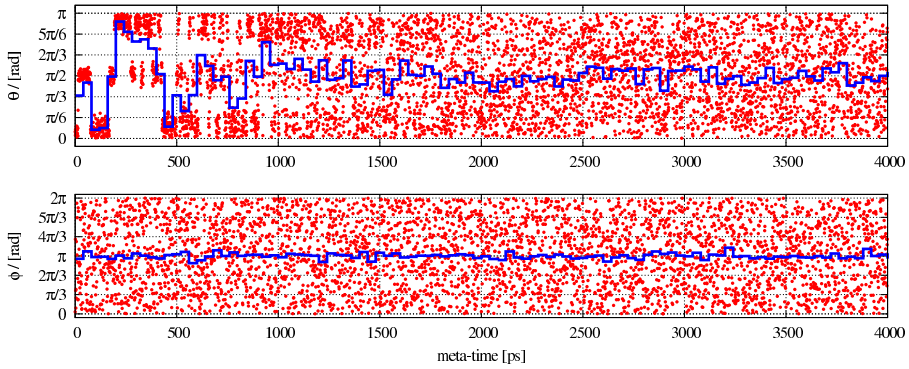
Q variable monitored ($\tau_C = 50$ step)

Panel 4.2: Simulation Parameters for all-atoms models (cyclohexane and tetrahydropyran). Parameters for the reported metadynamics simulations of a single molecule in TIP3P [83] water, performed with the NAMD [142] and PLUMED [27] software (see Section 3.5).

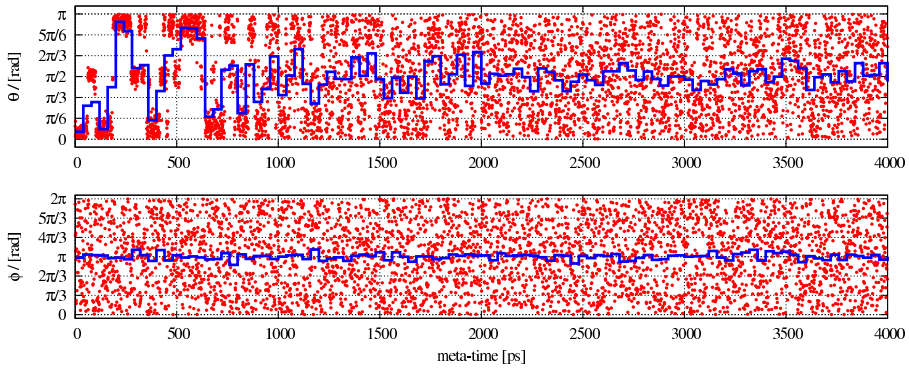
The *convergence pattern* for all atom models is shown in Fig. 4.8. It represents the position of the centroid of the Gaussians functions deposited during the metadynamics simulation. The simulation gave a total number of Gaussians that is quite high for such systems. Indeed, the behavior of the collective variables is clearly diffusive in the sense of Ref. [99] in a very short time. The θ value starts to explore a limited region of the conformational space, as a slow degree of freedom does in standard dynamics; as long as the bias potential is accumulated the system is able to move between free energy minima that are progressively filled. Eventually, almost a free exploration of the whole range $[0, \pi]$ is reached. The variable ϕ, on the contrary, seems to be from the beginning a fast degree of freedom, and this is comprehensible by means of the symmetry of the compound and the simple structure of the side chains. The simulations were anyway continued for a long time after the diffusivity was reached to explore possible long meta-time correlations along the biased variables. In Fig. 4.8 it is visible that, for such simple systems, the diffusivity reached in short meta-time and is stable.

Once the diffusivity of the CV is recognized, the "convergence" of metadynamics is established[3]. The free energy profile from this estimate are presented in Fig. 4.9. Free energy basins are clearly visible, and are located around the ideal conformers. We can evaluate qualitatively the goodness of the reconstruction following the symmetry consideration of Section 4.1.2. Since Cremer-Pople coordinates artificially distinguish between states that are identical for the stere-

[3]For standard metadynamics it is not a real convergence, but rather the beginning of the oscillation of the bias potential around the real free energy profile.



(a) Cyclohexane model compound



(b) Tetrahydropyran model compound

Figure 4.8: All-atoms models: convergence pattern. The centroids (θ_i, ϕ_i) of the Gaussians in the bias potential are presented as functions of the meta-time t_i of the simulation. To clarify the diffusivity, the mediated position of the centroids is indicated (blue line) with the mean taken for blocks of 1000 successive deposited Gaussians. The asymptotic values of the block means converge to the approximate center of the θ and ϕ ranges.

ochemistry, multiple estimates on the free energy profile are produced. In particular, we have two chairs (the chair and the inverted chair at poles, $\theta = 0, \pi$), and six skews (on the equator, $\theta = \pi/2$) estimates. The map representations of Fig. 4.9 exhibit the ring symmetry correctly (see Appendix B for further details):

- in cyclohexane profile, the region $(\theta, \phi) \in [0, 2\pi/3] \times [0, 2\pi]$ is symmetric by simultaneous θ inversion and ϕ shifting by a $2\pi/3$ angle, as in the ShN_n symmetry; the whole profile is also symmetric by ϕ inversion, as in the InN symmetry;
- when the first atom in the numeration is the ring oxygen, as in the presented profile, tetrahydropyran profile is symmetric by ϕ inversion, as in the InN symmetry.

To give a more quantitative indication of this statement, in Table 4.1 free

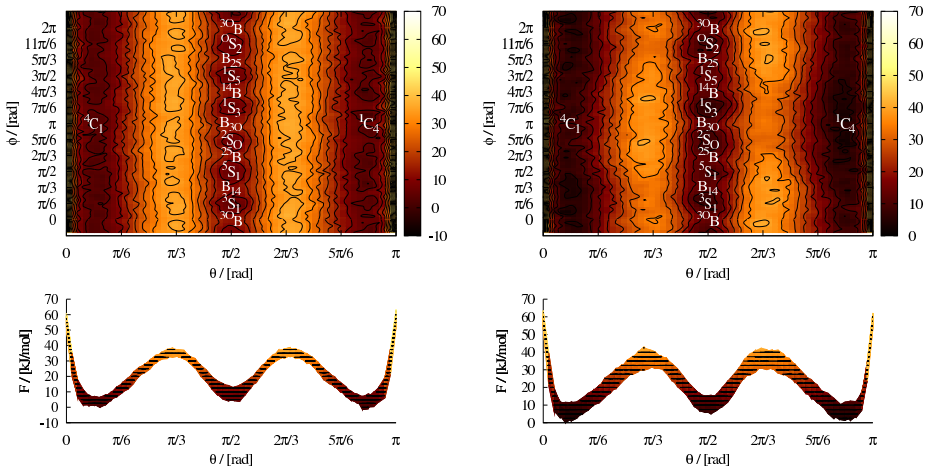


Figure 4.9: All-atoms models: free energy landscapes. Puckering free energy $F_{\text{META}}(\theta, \phi)$ (see Eq. (2.53)) of (a) cyclohexane and (b) tetrahydropyran model. Each profile is set to zero at the position of the minimum in the $^4C_1$ basin ($\theta < \pi/3$). Top: Plate Carrée projections, with isolines drawn every k_BT ($T = 300\,\text{K}$). The profile is replicated in two thin stripes at $\phi < 0$ and $\phi > 2\pi$ to stress the ϕ periodicity. Stable ideal conformers are also indicated. Bottom: projections of the free energy profile onto the $\phi = 0$ plane, with isolines drawn every k_BT. Darker colors corresponds to lower values of energy.

energy differences and conformer populations are reported. Data are shown both in the redundant form given by the simulation and for the real stereochemical states. As it is shown, the values $\Delta F = F(z_i) - F(z_0)$ calculated from local differences (where z_i is the local minimum in the considered basin while z_0 is the absolute minimum in the area $[0, \pi/3] \times [0, 2\pi)$ of the $^4C_1$ conformer) are close for the artificially different states, and they are also close to the corresponding synthetic value $\Delta F[S]$.

By and large, the previous tests confirm that the implementation of spherical Cremer-Pople coordinates is effective in handling conformational transition if the free energy reconstruction is performed with metadynamics.

4.1.4 Side chain effect on free energy landscapes reconstruction

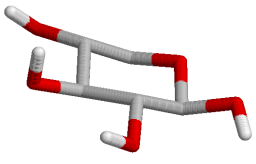


Figure 4.10: Test model: xylose (β-D-Xylopyranose).

We move now to a test-model system closer to hexopyranoses: a united-atom Xylopyranose (β-D-Xyl$p$). The parametrization for xylose was taken by

Table 4.1: All-atoms model: free energies and conformer populations. Values of ΔF and $P[S]$ are calculated from the free energy surface of Fig. 4.9 (only values for recognizable basins are reported). In differences $\Delta F = F(z_i) - F(z_0)$, z_i is the local minimum in the considered basin while z_0 is the absolute minimum in the area $[0, \pi/3] \times [0, 2\pi)$ (the area of the $^4C_1$ conformer). Probabilities $P[S]$ are calculated within rectangular domain centered on the selected conformer (following the grid on the map of Fig. 3.4). Differences $\Delta F[S]$ are calculated from population values, summing the population of stereochemical identical conformers, and giving ΔF with respect to $F_{^4C_1}$.

CP state	C_6H_{12}			$C_5H_{10}O$		
	ΔF	$P[S]$	$\Delta F[S]$	ΔF	$P[S]$	$\Delta F[S]$
$^4C_1$	0.00	50.5	0.00	0.00	43.7	0.00
$^1C_4$	−0.34	42.4		0.31	48.5	−0.62
$^5S_1$	1.14	1.0	1.53	–	–	–
^2S_O	1.28	0.6		1.77	1.0	1.59
$^1S_3$	1.13	0.9		1.20	2.0	
$^1S_5$	0.98	1.1		–	–	–
$^OS_2$	1.41	0.7		1.88	0.8	1.81
$^3S_1$	0.91	1.4		1.40	1.3	

Energies are in kcal/mol (with an estimated error $\delta F = 0.21$ kcal/mol), Populations are in %

direct simplification of the topology of Glucopyranose (*i.e.* β-D-Glc*p* without the $-CH_2OH$ side chain) within the G45a4-ASPG FF [10]. Such simplification permits to evaluate the influence of simple side chains. Simulations were performed with a single sugar molecule solvated by $\sim$ 500 SPC water molecules [21] (all the other details are listed in Panel 4.3).

MD parameters	
time-step	0.2 fs
box size	$\sim$ 2 nm
integrator	md (leap-frog)
non-bonded cutoff	1.0 nm
T-coupling	Nosé-Hover [129, 74]
T-ref, τ_T	300 K , 1 ps
P-coupling	Parrinello-Rahman [138]
p-ref, τ_p	1.013 25 bar , 1 ps
compressibility	4.5×10^{-5} bar^{-1}

PBC, COM motion remotion (translation), constrained bonds (SHAKE), Dispersion correction (energy and pressure)

Metadynamics parameters	
CV (ξ^1, ξ^2)	(θ, ϕ)
σ_1, σ_2	0.05 rad , 0.05 rad
τ_G	200 step
w	0.15 kcal/mol
#Gaussians	1×10^5 to 4×10^5

Q variable monitored (τ_C = 50 step), `INVERT` algorithm for sharp boundaries, the actual duration depends on metadynamics diffusivity

Panel 4.3: Simulation Parameters for united-atoms models (xylose and glucose). Parameters for the reported metadynamics simulations of a single molecule in SPC [21] water, performed with the GROMETA [32, 20, 111] software or the GROMACS [179] and PLUMED [27] software (see Section 3.5).

In Fig. 4.11 the time evolution of the position of the Gaussians centroids is presented. The presence of hydroxyl groups suggests in principle that the "con-

vergence" has to be slower than in toy-models. However, the registered behavior also evidences the presence of persistent undesired oscillations (in particular for the θ variable), even for very long meta-times. This pattern could in principle be interpreted both as a diffusivity regime not reached yet, or as a hysteresis in Gaussians deposition. If the second case holds, as pointed out in Ref. [99], other slow degrees of freedom could have been missed by metadynamics.

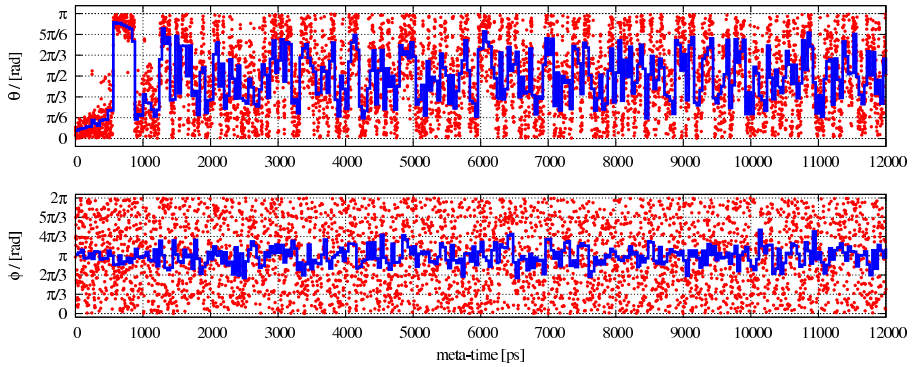
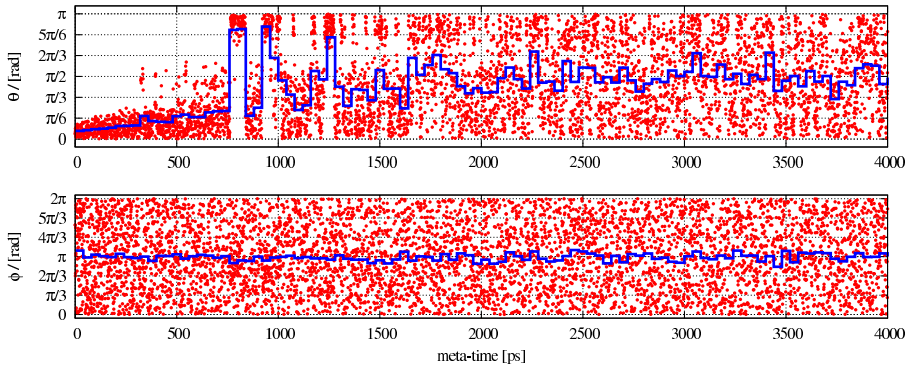


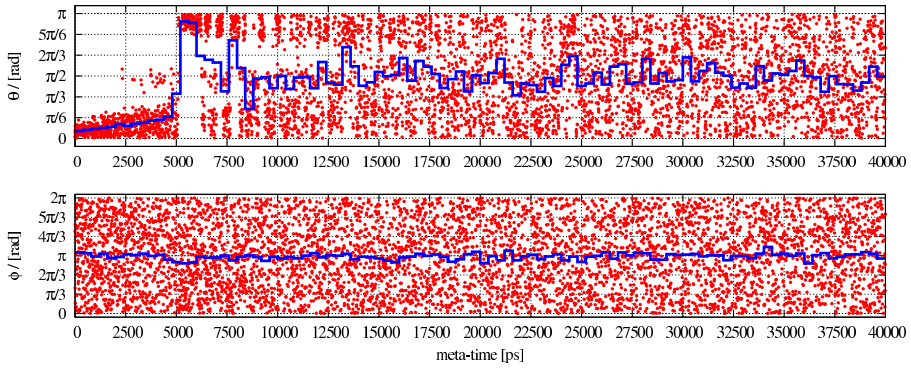
Figure 4.11: Test model: convergence pattern. The centroids (θ_i, ϕ_i) of the Gaussians in the bias potential are presented as functions of the meta-time t_i of the simulation. To clarify the diffusivity, the mediated position of the centroids is indicated (blue line) with the mean taken for blocks of 1000 successive deposited Gaussians. The asymptotic behavior of the block means oscillates sensibly around the centers of the θ and ϕ ranges.

To understand the underlying problem in our case, two other similar simulations were performed in different conditions: first all the side chains were kept uncharged (results are in Fig. 4.12(a)), and secondly the topology is unchanged but a Gaussians deposition rate ten times higher than the previous one is used ($\tau_G = 2000$ step, results are in Fig. 4.12(b)). In both cases the diffusivity is clearly reached, even if ring substituents are present. The amplitude of the oscillation around a purely diffusive behavior is still noticeably larger than in the cyclohexane or tetrahydropyran cases, but this is completely understandable because of the higher complexity of the compound interactions (steric interaction between ring substituents, possible hydrogen bonds with water model molecules, ...). In particular, the uncharged test indicates that an important source of hysteresis could occur due to electrostatic interaction, for hydrogen bonds formation between $-$OH groups and water molecule can form metastable structures. It is simple to understand that the larger is the number of relevant collective motions, the more difficult it is for metadynamics to perform a reliable free energy reconstruction. In other words, these interactions could be considered as the source of extra collective degrees of freedom, whose time-scale could affect metadynamics. Fortunately, the use of larger deposition times shows that electrostatic interactions seem to generate collective degrees of freedom on an intermediate time-scale between the "fast" atomic motion and the "slow" reaction coordinates. Indeed, with a lower deposition rate (the inverse of the deposition time), a longer equilibration time between two Gaussians deposition is provided. In this condition not only "fast" but also "not-so-slow" degrees

of freedom could properly be in thermal equilibrium during the simulation.



(a) Uncharged model



(b) Slower deposition time

Figure 4.12: Test model revisited: convergence pattern. The centroids (θ_i, ϕ_i) of the Gaussians in the bias potential are presented as functions of the meta-time t_i of the simulation. To clarify the diffusivity, the mediated position of the centroids is indicated (blue line) with the mean taken for blocks of 1000 successive deposited Gaussians. Since the test of Fig. 4.12(b) has a deposition rate ten times higher than the uncharged system of Fig. 4.12(a), the meta-time is ten times higher but the number of deposed Gaussians is similar. The asymptotic behavior of the block means now has in both cases a nearly convergent behavior to the centers of the θ and ϕ ranges.

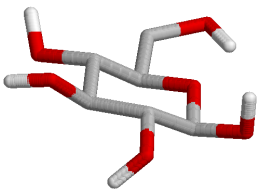


Figure 4.13: "Production" model: glucose (β-D-Glucopyranose).

The last model we present here is our basic "production" system: a united-atom Glucopyranose compound (β-D-Glc$p$). The parametrization for Glucopyranose was taken from Ref. [10]. In parallel with the previous description

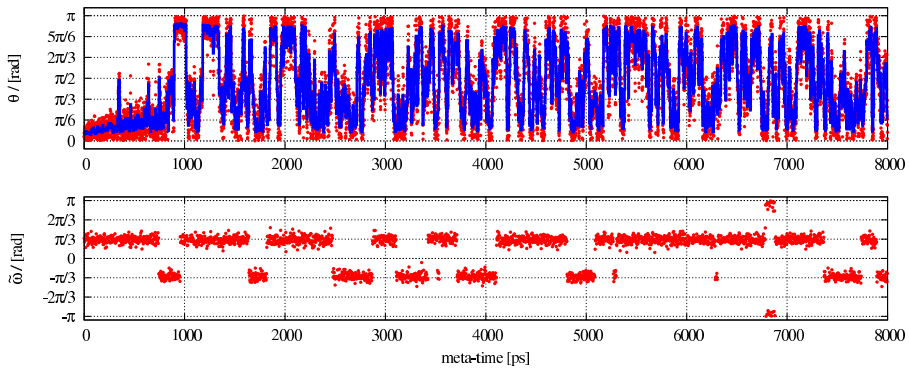
of side chain effects on metadynamics, in the following we will consider only the behavior of the hydroxymethil group[4]. However, we considered this case separately from the other side chains because, since the $-CH_2OH$ group is a structured side chain, we expect a proper rotameric distribution, namely a specific preferences for the torsional angle O5–C5–C6–O6 ($\widetilde{\omega}$ angle). This is indeed the case, as confirmed by experimental data (see for example the data of Refs. [127, 132, 126, 29, 174] reported in Tab.6 of Ref. [112]). Thus, its effect on metadynamics is potentially more important than the one from hydroxyl groups.

In Fig. 4.14 the time evolution of the $\widetilde{\omega}$ with respect to the evolution of the θ puckering angle is shown. The torsional angle $\widetilde{\omega}$ was only monitored while the θ puckering one and the puckering angle ϕ omitted in Fig. 4.14 are the CVs along which the bias potential is activated. Two deposition protocols, that differ only in the meta-time distance between two Gaussians deposition, were used. From these plots it is clear that the angle $\widetilde{\omega}$ is a slow degree of freedom too, as it was expected. It presents indeed metastable states at $\widetilde{\omega}$ values of $\pi/3$ (gauche+), 0 (trans) and $5\pi/3$ (gauche-), and the second of them results to be very poorly visited. Thus, in principle a metadynamics reconstruction that omits the $\widetilde{\omega}$ could produce a biased reconstruction because of not taking into account a slow degree of freedom. However, it has to be noted that the transition of the $\widetilde{\omega}$ torsion angle are uncorrelated to the transition of the θ and ϕ puckering angles. Thus, the $-CH_2OH$ group seems to be in thermal equilibrium like in an unbiased dynamics. This behavior is more clear when a slow deposition protocol is employed (see Fig. 4.14(b)).

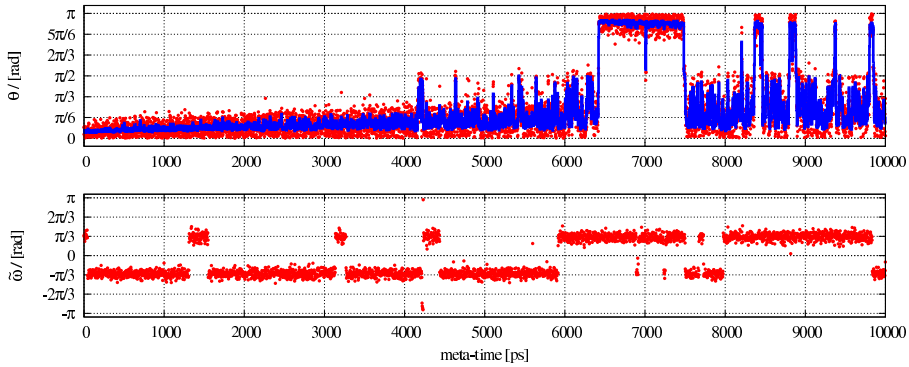
Only for the sake of qualitative comparison, in Fig. 4.15 the free energy surfaces of β-D-Xylopyranose and of β-D-Glucopyranose are presented. As it can be seen, the presence of the $-OH$ group at carbon atoms C1 to C4 and of $-CH_2OH$ at carbon atom C5 removes all the symmetry from the structure. Clear specific preferences and obstacles for ring conformers are now present, due to the different type and orientation of ring substituents. In both cases, for example, the conformer with equatorial orientation of ring substituents (the $^4C_1$ chair) results to be preferred with respect to a similar conformer but with axial substituents (the $^1C_4$ inverted chair). The presence of the $-CH_2OH$ group, also, changes significantly the stability pattern of flexible conformers on the equator of the presented plots. These are the kind of free energy profiles that we can expect for hexopyranoses.

Thus, the presence of ring substituents has clearly some effect in the ability to perform a metadynamics reconstruction. In particular, obstacles against a diffusive regime for the puckering variable emerge from the presence of other (collective) degrees of freedom. However, here there is clear evidence that it is possible to decouple properly the time scales of different collective degrees of freedom, in order to have a proper metadynamics reconstruction in the direction of the puckering coordinates only.

[4] A more complete description of free energy reconstruction of glucose and Aldopyranosides will be given in Chapter 6.

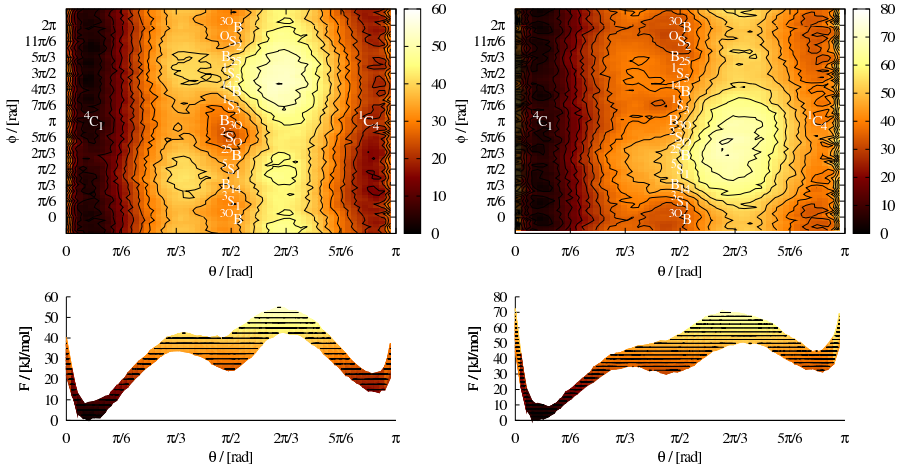


(a) Faster deposition time



(b) Slower deposition time

Figure 4.14: "Production" model: convergence pattern. The centroids θ_i of the Gaussians in the bias potential are presented as functions of the meta-time t_i of the simulation. The mediated position of the centroids is indicated (blue line) with the mean taken for blocks of 1000 successive deposited Gaussians. The meta-time evolution of the $\widetilde{\omega}$ torsion angle is presented to evaluate the possible correlation between the slow degrees of freedom θ and $\widetilde{\omega}$. In both cases (a) and (b) there is no correlations between the two variables, and the behavior of the $\widetilde{\omega}$ is compatible with an unbiased dynamics for the $-CH_2OH$ group.



(a) β-D-Xyl$p$ (b) β-D-Glc$p$

Figure 4.15: Test-models: free energy landscapes. Puckering free energy $F_{\text{META}}(\theta, \phi)$ (see Eq. (2.53)) of (a) xylose and (b) glucose model. Each profile is set to zero at the position of the minimum in the ${}^4C_1$ basin ($\theta < \pi/3$). Top: Plate Carrée projections, with isolines drawn every $2k_BT$ ($T = 300\,\text{K}$). The profile is replicated in two thin stripes at $\phi < 0$ and $\phi > 2\pi$ to stress the ϕ periodicity. Stable ideal conformers are also indicated. Bottom: projections of the free energy profile onto the $\phi = 0$ plane, with isolines drawn every k_BT. Darker colors corresponds to lower values of energy.

4.2 Improving standard metadynamics

Some aspects on metadynamics with puckering coordinates were discussed in the previous Section. We discussed how it is necessary to handle all relevant slow degrees of freedom (the aforementioned hysteresis problem) and to assure that a "convergence" has established at the end of the run. Besides, two other main issues arise for the plain metadynamics scheme: first, the free energy reconstruction could suffer from systematic dependence on the protocol parameters (Gaussians height w and widths σ_a, and deposition time τ_G); second, the estimation of the error on the reconstruction is still not straightforward, because, as stated in Section 2.2.4, it requires a set of independent free energy reconstructions.

To address these particular aspects, in the present Section we will describe two general variants of the standard metadynamics: the *Well-tempered metadynamics* scheme [15] and the *Umbrella Sampling refinement* scheme [12, 9].

4.2.1 Well-tempered Metadynamics

The structure of the history-dependent potential of Eq. (2.45) is arbitrary, thus in principle other choices are possible. For example, following Ref. [15], the function

$$V_{\text{B}}(\xi(r),t) = k_B\Delta T \ln\left[1+\frac{\omega}{k_B\Delta T}N(\xi(r),t)\right] \quad , \tag{4.1}$$

where ω has the dimension of an energy rate, ΔT denotes a temperature window, is a history-dependent bias because the function

$$N(\xi(r),t) = \int_0^t \delta(\xi(r)-\xi(r_{\text{B}}(t)))\mathrm{d}t \tag{4.2}$$

follows the evolution of variable $\xi(r)=z$ along the biased trajectory $r_{\text{B}}(t)$. The function $N(\xi(r),t)$ is proportional to the density $\rho_{\text{B}}(\xi(r),t) = N(\xi(r),t)/t$, thus the use of the logarithm of $N(\xi(r),t)$ produces a bias potential that resembles a free energy term. The prefactor $\omega/k_B\Delta T$ has the dimension of $[t]^{-1}$ and keeps the proper unit. Taking now the time derivative of Eq. (4.1)

$$\frac{\mathrm{d}}{\mathrm{d}t}V_{\text{B}}(z,t) = \frac{\omega}{1+\frac{\omega}{k_B\Delta T}N(z,t)}\dot{N}(z,t) = \omega\,\mathrm{e}^{-V_{\text{B}}(z,t)/k_B\Delta T}\,\delta(z-z_{\text{B}}(t)) \quad , \tag{4.3}$$

and integrating again the expression we obtain:

$$\begin{aligned} V_{\text{B}}(z,t) &= \int_0^t \omega\,\mathrm{e}^{-V_{\text{B}}(z,t)/k_B\Delta T}\,\delta(z-z_{\text{B}}(t))\mathrm{d}t \\ &= \lim_{\tau_G\to\mathrm{d}t}\sum_{\substack{t_k=k\tau_G \\ k\in\mathbb{N}}}^{t_k<t} \omega\tau_G\,\mathrm{e}^{-V_{\text{B}}(z,t)/k_B\Delta T}\prod_{a=1}^{m}\lim_{\sigma_a\to 0}\frac{\mathrm{e}^{-\frac{[z^a-z_k^a]^2}{2\sigma_a^2}}}{\sqrt{2\pi}\sigma_a} \quad , \end{aligned}$$

in which we made the substitution of the δ-function with finite width Gaussians and of the integration step $\mathrm{d}t$ with discrete deposition time τ_G. The conversion from a continuous formulation to a discrete one is useful to recover a bias potential similar to the one of standard metadynamics (not to mention the computational necessity of working with discrete time-steps and positions). Omitting the limit symbols, the bias potential now reads

$$V_{\text{B}}(\xi(r),t) = \sum_{\substack{t_k=k\tau_G \\ k\in\mathbb{N}}}^{t_k<t} \underbrace{\omega\tau_G \, \mathrm{e}^{-V_{\text{B}}(\xi(r),t)/k_B\Delta T}}_{A(t)} \prod_{a=1}^{n} \mathrm{e}^{-\frac{[\xi^a(r)-z_k^a]^2}{2\sigma_a^2}} \quad , \tag{4.4}$$

where we recognize in Eq. (4.4) the same structure of Eq. (2.45) but with a *time-dependent Gaussians height* $A(t)$. Its time evolution follows the time evolution of $\dot{V}_{\text{B}}$. The factor ω is (proportional to) the initial deposition rate[5]. For increasing time t we have $\dot{V}_{\text{B}} \sim 1/t$, because of the $N(\xi(r),t) \sim t$ term in the denominator. Note that $\dot{V}_{\text{B}}(z,t) \xrightarrow{t\to+\infty} 0$ but not uniformly in z space, because the biased dynamics will spend different time at different value of z. As a whole, $A(t)$ decreases with time.

This different implementation has the advantage that the convergence of the reconstruction to the exact profile $F(z)$ can be proved rigorously (a convergence proof is given in Appendix D): the free energy estimate now reads

$$\boxed{F_{\text{WTM}}(z) = -\frac{T+\Delta T}{\Delta T} V_{\text{B}}(z,t) \xrightarrow{t\to+\infty} F(z)} \quad . \tag{4.5}$$

Here the bias potential does not totally compensate for the underlying free energy surface; rather, we have

$$F(z) + V_{\text{B}}(z,t) = F(z) - \frac{\Delta T}{T+\Delta T} F(z) = \frac{T}{T+\Delta T} F(z) \quad , \tag{4.6}$$

leading to the following biased probability distribution for z:

$$\rho_{\text{B}}(z,t) = \mathrm{e}^{-\beta F(z)/\alpha} = \mathrm{e}^{-\beta' F(z)} \quad , \quad \beta' = \frac{\beta}{\alpha} = \frac{1}{k_B(T+\Delta T)} \quad , \tag{4.7}$$

where we have defined the *bias factor* $\alpha = \dfrac{T+\Delta T}{T}$. Thus, the probability distribution $\rho_{\text{B}}(z,t)$ obtained towards the end of a well-tempered metadynamics run is altered and the z space sampling near convergence resembles a canonical sampling at the enhanced temperature $T+\Delta T$. In this sense, tuning the value ΔT selects the temperature window in which the metadynamics exploration will be performed. In the limit $\Delta T \to +\infty$ the standard metadynamics scheme is restored.

4.2.2 Umbrella Sampling refinement

The basic assumption of metadynamics is that in

$$F(z) = -V_{\text{B}}(z,t) - \frac{1}{\beta} \ln \rho_{\text{B}}(z,t) \quad \Leftrightarrow \quad \rho_{\text{B}}(z,t) = \mathrm{e}^{-\beta[F(z)+V_{\text{B}}(z,t)]} \tag{4.8}$$

[5] With respect to standard metadynamics with fixed height w we have $w = \omega\tau_G$ at $t = 0$, in this sense $\omega = w/\tau_G$ is the initial deposition rate.

the term $\ln \rho_{\text{B}}(z,t)$ is negligible for a sufficiently long time t. This is true if the biased "probability density" $\rho_{\text{B}}(z,t)$ is (even approximately) uniform, because in this case only a global additive constant is neglected in the estimate of Eq. (4.8). However, in practice the free energy landscape $F(z)+V_{\text{B}}(z,t)$ is not completely flat by construction, because of the finite-height Gaussians used.

A method to handle this problem, proposed first in [12] and then in [10], is not to neglect the logarithmic term in Eq. (4.8) but, rather, to estimate it with Umbrella Sampling-like simulations. After a metadynamics reconstruction is performed for a suitable time t^*, the function $F(z)+V_{\text{B}}(z,t^*)$ will have more or less only low (thermal) free energy structures. Thus, a standard MD estimation of the residual "free energy" of this system is feasible:

$$\rho_{\text{B}}(z,t^*) = \mathrm{e}^{-\beta[F(z)+V_{\text{B}}(z,t^*)]} = \lim_{\mathcal{T}\to+\infty} \frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} \mathrm{d}t \delta(z-\xi(r_{\text{B}}(t))) \tag{4.9}$$

(cfr Eq. (2.44)). The bias potential V_{B} is kept fixed in this phase, and so it is used as in techniques like Umbrella Sampling [175] for a direct estimation of the biased probability density. The advantage is that the external bias V_{B} is not arbitrary, but has been created on the system itself, much more in the spirit of Adaptive Umbrella Sampling [121]. Thus the sampling becomes better with growing precision of V_{B}.

The free energy estimate from Eq. (4.8) now reads

$$\boxed{F(z) \simeq F_{\text{M+US}}(z) = -V_{\text{B}}(z,t^*) - \frac{1}{\beta} \ln H_{\text{B}}(z,t^\star)} \quad , \tag{4.10}$$

where

$$V_{\text{B}}(z,t^*) = \sum_{\substack{t_k = k\tau_G \\ k=1}}^{\mathcal{M}} w \prod_{a=1}^{m} \mathrm{e}^{-\frac{[z^a - z_k^a]^2}{2\sigma_a^2}} \qquad (t^* = \mathcal{M}\tau_G) \quad , \tag{4.11a}$$

$$H_{\text{B}}(z,t^\star) = \frac{1}{t^\star} \int_0^{t^\star} \mathrm{d}t \delta(z - z_{\text{B}}(t)) \simeq \frac{n(z,t^\star)}{\mathcal{N} I(z)} \qquad (t^\star = \mathcal{N}\Delta t) \quad . \tag{4.11b}$$

A metadynamics run builds a basic estimate $V_{\text{B}}(z,t^*)$ for $F(z)$ with $\mathcal{M}$ Gaussians function. The residual probability density is evaluated as a normalized[6] histogram $H_{\text{B}}(z,t^\star)$, where $n(z,t^\star)$ is the fraction of the total number $\mathcal{N}$ of sampled points of the trajectory $z_{\text{B}}(t) = \xi(r_{\text{B}}(t))$ that belongs to the spatial discretization of the δ-function in finite intervals $I(z) = \Delta z^1 \cdot \ldots \cdot \Delta z^m$.

The evaluation of $\rho_{\text{B}}(z,t)$ not only corrects the free energy estimate, but also allows an estimation of the reconstruction error. Indeed, if the deposition protocol may introduce some spurious features in the reconstructed free energy, then they are reflected in the residual distribution of Eq. (4.9). The standard sampling of the function $\rho_{\text{B}}(z,t)$ allows to solve the problem and to increase

[6] As long as we use always $\ln H_{\text{B}}(z,t^\star)$, the normalization factors $\mathcal{N}$ and $I(z)$ turns to be additive constant to the free energy and so are negligible for ΔF calculations.

accuracy up to the statistical precision of the sampling itself. This happens because the evaluation of the residual "probability density" converts the systematic error which depends on the deposition protocol into statistical errors from the sampling of $\rho_{\text{B}}(z,t)$. In the standard hypothesis of independent sampling[7], from the Poisson statistics the error for a given counting is

$$n(z,t^\star) \quad \longrightarrow \quad \delta n(z,t^\star) = \sqrt{n(z,t^\star)} \quad . \tag{4.12}$$

From this starting point we have that

$$\delta F_{\text{M+US}}(z) = 1/\beta\sqrt{n(z,t^\star)} \tag{4.13}$$

is a local estimation of the standard error on the free energy reconstruction.

We want to stress that with this scheme a refinement of the metadynamics estimate $V_{\text{B}}(z,t)$ is realized, and an error estimate $\delta F_{\text{M+US}}$ is performed. Remarkably, the estimate could be done from a single simulation.

Correct measure in computing histograms by spherical coordinates

For Cremer-pople coordinates the sampling of the residual probability density

$$\rho_{\text{B}}(z^{\text{CP}}) = \lim_{\mathcal{T}\to+\infty} \frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} \delta(\xi^{\text{CP}}(r_{\text{B}}(t)) - z^{\text{CP}}) \mathrm{d}\tau \tag{4.14}$$

requires the proper usage of the δ-function. Given Cartesian (q_x, q_y, q_z) and spherical (Q,θ,ϕ) coordinate representations in the six-membered ring puckering space $\mathbb{R}^3$, the δ-function reads

$$\delta(z^{\text{CP}}) = \delta(q_x)\delta(q_y)\delta(q_z) = \frac{\delta(Q)}{Q^2}\frac{\delta(\theta)}{\sin\theta}\delta(\phi) \tag{4.15}$$

where the extra factors in the denominator account for the *Jacobian factor* $\mathcal{J}(Q,\theta,\phi)$ that changes the volume element

$$\mathrm{d}V = \mathrm{d}q_x\mathrm{d}q_y\mathrm{d}q_z = \underbrace{Q^2\sin\theta}_{\mathcal{J}(Q,\theta,\phi)}\,\mathrm{d}Q\mathrm{d}\theta\mathrm{d}\phi \quad . \tag{4.16}$$

In our calculations, we are interested only in the the angular (θ,ϕ) Cremer-Pople variables. In this case, the plain histogram

$$H_{\text{B}}(\theta,\phi,t^\star) = \frac{1}{t^\star}\int_0^{t^\star} \delta(\theta_{\text{B}}(\tau)-\theta)\delta(\phi_{\text{B}}(\tau)-\phi)\mathrm{d}\tau = \frac{n(\theta,\phi,t^\star)}{\mathcal{N}\Delta\theta\Delta\phi} \tag{4.17}$$

of visited points along the biased trajectory $\{(\theta_{\text{B}}(t),\phi_{\text{B}}(t))\}$ could be still simply accumulated by direct counting as indicated in Eq. (4.11b). Then, to reconstruct the proper estimate of the residual probability density

$$\rho_{\text{B}}(\theta,\phi) = \lim_{\mathcal{T}\to+\infty}\frac{1}{\mathcal{T}}\int_0^{\mathcal{T}} \frac{\delta(\theta_{\text{B}}(\tau)-\theta)}{\sin\theta}\delta(\phi_{\text{B}}(\tau)-\phi)\mathrm{d}\tau \simeq \frac{H_{\text{B}}(\theta,\phi,t^\star)}{\sin\theta} \tag{4.18}$$

[7] Standard techniques to assure that the counting events are uncorrelated have to be taken into account. Non-standard sources of correlation can occur if the biased free energy landscape has still small structures, and has to be corrected coherently.

the corresponding factor $\frac{1}{\sin\theta}$ from the Jacobian is needed. The total free energy now could be written as

$$F(z) \simeq \boxed{F_{\text{M+US}}(z) = -V_{\text{B}}(z,t^*) - \frac{1}{\beta}\ln H_{\text{B}}(z,t^\star) + \frac{1}{\beta}\ln\sin\theta} \tag{4.19}$$

where the last term is the Jacobian contribution $\frac{1}{\beta}\ln\sin(\theta)$, that vanishes in standard (Cartesian) cases where $\mathcal{J} = 1$ (as in Eq. (4.10)).

Besides, it has to be mentioned that the same consideration about Jacobian factors holds true for all quantities calculated with the presence of the volume element $\mathrm{d}V$. In particular, the aforementioned population probabilities are now calculated as

$$P[S] = \frac{\int\limits_{z\in S\subset\mathrm{R}^3} \mathrm{e}^{-\beta F(z)}\,\mathrm{d}z}{\int\limits_{\mathrm{R}^3} \mathrm{e}^{-\beta F(z)}\,\mathrm{d}z} = \frac{\int\limits_{S\subset[0,\pi]\times[0,2\pi)} \mathrm{e}^{-\beta F(\theta,\phi)}\sin\theta\mathrm{d}\theta\mathrm{d}\phi}{\int\limits_{[0,\pi]\times[0,2\pi)} \mathrm{e}^{-\beta F(\theta,\phi)}\sin\theta\mathrm{d}\theta\mathrm{d}\phi} \tag{4.20}$$

from the profile $F(\theta,\phi)$. In the specific case of an the Umbrella Sampling refinement of metadynamics estimates, we have

$$P[S] = \frac{\int\limits_{S\subset[0,\pi]\times[0,2\pi)} \mathrm{e}^{-\beta V_{\text{B}}(\theta,\phi,t^*)}\,H_{\text{B}}(\theta,\phi,t^\star)\mathrm{d}\theta\mathrm{d}\phi}{\int\limits_{[0,\pi]\times[0,2\pi)} \mathrm{e}^{-\beta V_{\text{B}}(\theta,\phi,t^*)}\,H_{\text{B}}(\theta,\phi,t^\star)\mathrm{d}\theta\mathrm{d}\phi} \tag{4.21}$$

with the simplification using the Jacobian factor in the definition of $P[S]$ and the Jacobian factor from the free energy evaluation.

Analogous calculations are needed also in the case of plain MD free energy evaluation (see Eq. (2.44)). The free energy estimate is then

$$F(z) \simeq F_{\text{D}}(z) = -\frac{1}{\beta}\ln H(z,t^\star) + \frac{1}{\beta}\ln\sin\theta \tag{4.22}$$

where $H(z,t^\star)$ is a plain histogram like the one of Eq. (4.17) but along an unbiased trajectory $\{\theta(t),\phi(t)\}$.

4.3 Spherical coordinates and metadynamics

In this Section we want to specifically explore possible problems of metadynamics performed in spherical coordinates. A simple observation indicates that metadynamics does not reproduce properly the behavior in the close vicinity of the poles. Within the toy-model N.1 of Section 4.1.1, a comparison between direct dynamics[8] $F_{\text{D}}(z)$ and metadynamics $F_{\text{M}}(z)$ free energy estimates

[8] For details in the calculation of $F_{\text{D}}(z)$ see Section 4.2.2, in particular Eq. (4.22).

is possible. In this case, the MD result could be considered as a reference for metadynamics. To this end, a metadynamics simulation was performed with the accelerated sampling protocol of Panel 4.3 on the indicated toy-model. The result, showed in Fig. 4.16, highlights a critical behavior in metadynamics for the regions very close to poles: a spurious increment in free energy values is systematically found within a region close to the poles position.

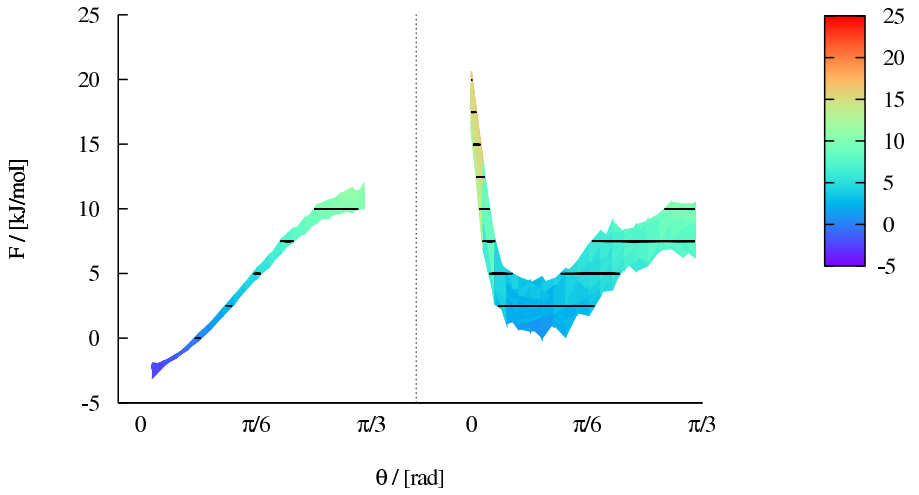


Figure 4.16: Toy-model N.1: free energy landscapes (detail). Side view of the free energy surface, in the close vicinity of the north pole, from direct MD estimate (left) and from metadynamics (right). The profiles are shifted in order to have the surfaces aligned approximately at $\theta = \pi/3$. Isolines are drawn every $k_B T$ ($T = 300\,\mathrm{K}$) starting from $F = 0$. The behavior at the south pole is equivalent.

This problem is quite natural for two reasons:

1. the standard problem in the reconstruction of sharp edges of a given CV with a sum of finite width functions. Since the variable θ is intrinsically limited, this means that at the edges $\theta = 0, \pi$ the free energy is discontinuous ($F(\theta, \phi) = 0$ for $\theta \notin [0, \pi]$);

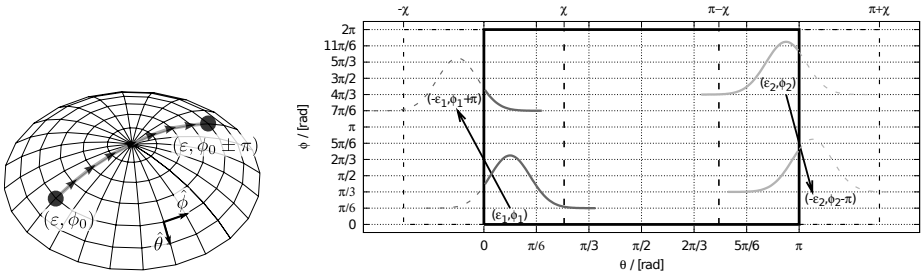
2. the nature of the poles to be somehow critical points[9] in the coordinate system. In computer simulation the sphere surface is "rectified" as a rectangular domain. The sampling of the poles (that are in the real space small regions) then turns to be the sampling of a quite large rectangular domain of fictitious bins.

The Umbrella Sampling refinement, that affects globally the free energy estimate, could in principle correct this problem too. In the following two extra solutions will be also presented.

[9] We are not referring to the problem of poles definition in spherical coordinates. The exact position of poles, which are not defined by construction, are only two sets of null measure, and then the occurrence of two specific points in a simulation is negligible.

4.3.1 Periodicity at poles

The spherical Cremer-Pople angles have ranges $(\theta, \phi) \in [0, \pi] \times [0, 2\pi)$. The domain is periodic in ϕ direction and non-periodic in θ direction. However, at the border of the θ domain a periodicity is somehow present. Consider the region $(\theta, \phi) \in [0, \varepsilon] \times [0, 2\pi)$ of Fig. 4.17(a), with ε a small angular value, around the north pole (the same consideration holds for the south pole, too). Now move a test point, starting from a position (ε, ϕ_0), towards the north pole following the direction of the meridian ϕ_0. Once the pole is crossed, the point "jumps" to the corresponding meridian $\phi_0 \pm \pi$. The sign is related to the periodicity of ϕ: the angle $\phi_0 \pm \pi$ that falls in $[0, 2\pi]$ is the chosen one. Eventually, following the meridian $\phi_0 \pm \pi$ the point reaches the position $(\varepsilon, \phi_0 \pm \pi)$ symmetrically located with respect to the starting point. The proposed trajectory across the poles is regularly continuous even if the coordinates exhibit a clear discontinuity.



(a) (b)

Figure 4.17: Periodicity at poles. (a): coordinate discontinuity in pole crossing; (b): modification of the standard Gaussians deposition to account poles periodicity. The extended range for θ variable is $[-\chi, \pi + \pi]$ (between the dashed lines). Gaussians deposited in the regions $[0, \chi]$ and $[\pi - \chi, \pi]$ are replicated to account the periodicity. Only solid portion of Gaussian functions can affect the system via meta-forces.

A (meta)dynamics calculation must permit such trajectories[10], *i.e.* a Gaussians function located at (ε, ϕ_0) should in principle be able to pull the representative point of the system towards $(\varepsilon, \phi_0 \pm \pi)$. However, this is not completely the case. The bias potential $V_B(\theta, \phi, t)$ in computer simulations is defined on the rectangular domain $[0, \pi] \times [0, 2\pi)$, as depicted in Fig. 4.17(b). Within the *minimum image convention*[11], that accounts for the ϕ periodicity in standard MD, Gaussian functions near the ϕ borders are effectively extended across the periodic boundary. On the contrary, a Gaussian function located near the north pole at (ε_1, ϕ_1) extends towards the unphysical region of $\theta < 0$. This area is correctly unreachable by the system, because the coordinate definition cannot give negative values of θ by construction. Nevertheless, the tail of this Gaussian in the region $\theta < 0$ is not counted to affect the system along the $\phi_1 + \pi$ meridian. The only way to have indirectly such effect is to add a different Gaussian function at $(\varepsilon^*, \phi_1 + \pi)$ with $0 < \varepsilon^* < \varepsilon_1$, but the closer is the Gaussian to

[10] Unless energetic and/or entropic constraints prevent it explicitly. This is not the case for puckering, as showed in MD tests (see for example Fig. 4.3).

[11] The minimum image keep all differences $|\phi_1 - \phi_2| < \pi$ for a 2π-periodic ϕ angle.

the pole the lower is the possibility of this compensating deposition. Thus, the standard deposition strategy is not wrong in principle, but could add avoidable (even if not so severe) systematic errors and extra computational cost to the reconstruction.

To take into account this periodicity at poles in a simpler way, we can modify the deposition procedure as follows (see Fig. 4.17(b) for a pictorial view of the procedure). When a Gaussian function is deposited at (θ_i, ϕ_i) within a region $[0, \chi] \times [0, 2\pi)$ near the border $\theta = 0$ of the θ domain, then an extra Gaussian function is deposited in $[-\chi, 0] \times [0, 2\pi)$ at the position $(-\theta_i, \phi_i \pm \pi)$ (the sign $\pm$ discriminated as before, thus chosen to have the new angle $\phi_i \pm \pi$ in $\in [0, 2\pi)$). In a similar way extra Gaussian functions are added near the $\theta = \pi$ border, giving the compact prescription

$$\begin{array}{ccc} (\theta_i, \phi_i) & \xrightarrow[\text{then add}]{\text{if}(|\theta_i - B| < \chi)} & (2B - \theta_i, \phi_i \pm \pi) \\ \cap & & \cap \\ [0, \pi] \times [0, 2\pi) & & [-\chi, \pi + \chi] \times [0, 2\pi) \end{array} \tag{4.23}$$

for the poles $B = 0, \pi$, as shown in Fig. 4.17(b). The effective domain of the bias potential is then extended with two unphysical regions of width χ at each border. This extension has to be commensurate with the Gaussians width in θ direction, typically with values $\chi > 3\sigma_\theta$. In this way, the left/right tail contribution of the standard Gaussians in (θ_i, ϕ_i) is still lost, but is effectively counted by the right/left tail of the corresponding periodic Gaussians in $(-\theta_i, \phi_i \pm \pi)$ that falls in the physical region. The bias potential now reads

$$\begin{aligned} V_{\text{B}}(\theta, \phi, t) = & \sum_{\substack{t_k = k\tau_G \\ \text{standard}}} w \exp\left\{ -\frac{[\theta - \theta_k]^2}{2\sigma_\theta^2} \right\} \exp\left\{ -\frac{[\phi - \phi_k]^2}{2\sigma_\phi^2} \right\} + \\ & + \sum_{\substack{t_k = k\tau_G \\ \text{periodic}}} w \exp\left\{ -\frac{[\theta - (2B - \theta_k)]^2}{2\sigma_\theta^2} \right\} \exp\left\{ -\frac{[\phi - (\phi_k \pm \pi)]^2}{2\sigma_\phi^2} \right\} \end{aligned} \tag{4.24}$$

and the system could be pushed towards the north pole by a Gaussian in (ε, ϕ_0) (meta-force in the direction of $\theta - \varepsilon < 0$), and when the pole is crossed the system is pushed away from the pole by the corresponding periodic Gaussian $(-\varepsilon, \phi_0 \pm \pi)$ (meta-force in the direction of $\theta + \varepsilon > 0$), like in the path showed in Fig. 4.17(a).

4.3.2 Alternative spherical angles

The problem with the under-sampling of poles is important not only *per se*, but also because in the Cremer-Pople framework the two main thermodynamic states are located at poles: the ideal chairs conformers occupy exactly the north and the south poles of the puckering sphere. However, as shown before a spherical framework is still desirable.

A possible way to circumvent this problem, or at least to evaluate its order of magnitude, is to "rotate" the puckering sphere. Since we are interested in a thin, spherically symmetric domain, we could in principle use a generic pair (ϑ, φ) of colatitude and longitude angles to describe the puckering or to perform

metadynamics. We recall here that geographical angles (ϑ, φ) are simply given by

$$\begin{cases} r_1 = \varrho \sin\vartheta \cos\varphi \\ r_2 = \varrho \sin\vartheta \sin\varphi \\ r_3 = \varrho \cos\vartheta \end{cases} \iff \begin{cases} \varrho = \sqrt{r_1^2 + r_2^2 + r_3^2} \\ \vartheta = \arccos\left[r_3/\varrho\right] \\ \varphi = \arctan\left[r_2/r_1\right] \end{cases} \tag{4.25}$$

where (r_1, r_2, r_3) are Cartesian coordinates. The corresponding spherical coordinates $(\varrho, \vartheta, \varphi)$ have the $\hat{\mathrm{r}}_3$ direction as the polar axis and $r_1 r_2$ as the equatorial plane. The idea is to start from the Cartesian puckering representation (q_x, q_y, q_z) (see Table 3.1 in Section 3.4.2 for details), to select the polar axis in each Cartesian directions and to build the corresponding (geographical) angle pair. We have then 3 possibilities[12]

$$\begin{cases} \theta = \arccos\left[q_z/Q\right] \\ \phi = \arctan\left[q_y/q_x\right] \end{cases} , \begin{cases} \gamma = \arccos\left[q_x/Q\right] \\ \eta = \arctan\left[q_z/q_y\right] \end{cases} , \begin{cases} \mu = \arccos\left[q_y/Q\right] \\ \nu = \arctan\left[q_x/q_z\right] \end{cases} \tag{4.26}$$

that correspond to three spherical coordinate sets:

1. the (Q, θ, ϕ) set, where $\hat{\mathrm{q}}_z$ is the polar axis and $q_x q_y$ is the equatorial plane (the standard Cremer-Pople spherical coordinates);
2. the (Q, γ, η) set, where $\hat{\mathrm{q}}_x$ is the polar axis and $q_y q_z$ is the equatorial plane;
3. the (Q, μ, ν) set, where $\hat{\mathrm{q}}_y$ is the polar axis and $q_z q_x$ is the equatorial plane.

Obviously, from the point of view of a simple representation of conformers, the three sets have the same capability, like for Cartesian or cylindrical CP representations. However, in the planar projections of Figs. 4.18(b) to 4.18(d) the "localization" of the conformers somehow changes. Indeed, the conformers at poles changes from ${}^4C_1$ and ${}^1C_4$ for (θ, ϕ) coordinates, to ${}^{3\mathrm{O}}B$ and $B_{3\mathrm{O}}$ for (γ, η) coordinates, and to ${}^5S_1$ and ${}^1S_5$ for (μ, ν) coordinates.

To perform metadynamics in these alternative spherical sets, the extra gradients

$$\nabla_i\gamma \quad , \quad \nabla_i\eta \quad , \quad \nabla_i\mu \quad , \quad \nabla_i\nu \tag{4.27}$$

are needed. By direct calculation, we have

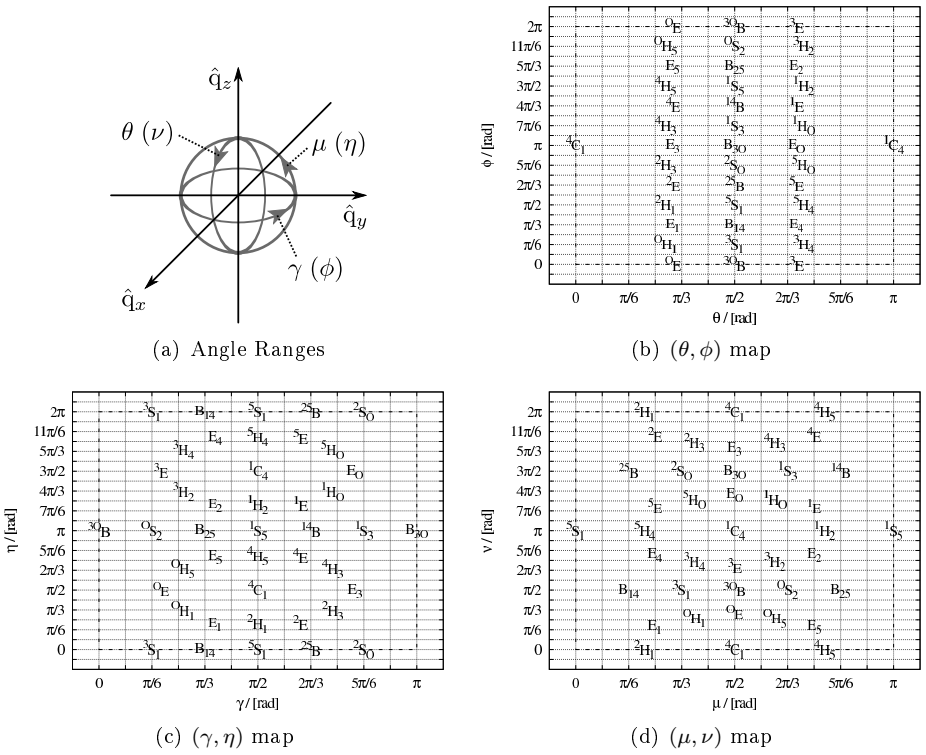
$$\nabla_i\gamma = \frac{1}{3\sqrt{2}Q^2} \frac{1}{\sqrt{2\mathcal{A}_2^2 + \mathcal{C}^2}} \Big[\mathcal{B}_2 \left(2\mathcal{A}_2 \nabla_i\mathcal{A}_2 + \mathcal{C}\nabla_i\mathcal{C}\right) - \left(2\mathcal{A}_2^2 + \mathcal{C}^2\right)\nabla_i\mathcal{B}_2\Big] \tag{4.28}$$

$$\nabla_i\eta = \sqrt{2}\frac{-\mathcal{A}_2\nabla_i\mathcal{C} + \mathcal{C}\nabla_i\mathcal{A}_2}{2\mathcal{A}_2^2 + \mathcal{C}^2} \tag{4.29}$$

$$\nabla_i\mu = -\frac{1}{3\sqrt{2}Q^2} \frac{1}{\sqrt{2\mathcal{B}_2^2 + \mathcal{C}^2}} \Big[\mathcal{A}_2 \left(2\mathcal{B}_2 \nabla_i\mathcal{B}_2 + \mathcal{C}\nabla_i\mathcal{C}\right) - \left(2\mathcal{B}_2^2 + \mathcal{C}^2\right)\nabla_i\mathcal{A}_2\Big] \tag{4.30}$$

$$\nabla_i\nu = \sqrt{2}\frac{\mathcal{C}\nabla_i\mathcal{B}_2 - \mathcal{B}_2\nabla_i\mathcal{C}}{2\mathcal{B}_2^2 + \mathcal{C}^2} \tag{4.31}$$

[12] The total puckering amplitude $Q = \sqrt{q_2^2 + q_3^2}$ is common to all coordinate set.



(a) Angle Ranges

(b) (θ, ϕ) map

(c) (γ, η) map

(d) (μ, ν) map

Figure 4.18: Alternative spherical maps. (a): Orientation of spherical angles. (b)-(d): Plate Carrèe projections as functions of each proposed spherical set. The localization of the 38 ideal IUPAC conformers are indicated in all projections.

(see Appendix A for details), with the same successive hierarchical construction of the derivative using Eqs. (3.32) and (3.33). If we perform metadynamics in these alternative coordinates, the exploration of the chair and inverted chair basins in principle does not suffer from the aforementioned problems of poles sampling. Indeed, within the alternative spherical representations ${}^4C_1$ and ${}^1C_4$ states are on the equator of the new coordinate system. Two test simulations on the toy-model of Section 4.1.1 were performed, using the (γ, η) and the (μ, ν) coordinate sets, to explore this possibility. The simple metadynamics scheme was used for a short meta-time, with the algorithm for the periodicity at poles described in Section 4.3.1. In Fig. 4.19 the obtained free energy surfaces are shown, and are compared with the corresponding (θ, ϕ) profile and with the plain MD profile. All the profiles have been plotted with respect to the original (θ, ϕ) coordinates, using the corresponding transformation formulas (see Table 4.3 at the end of the Section). This change of representation is useful to have a unified framework of representation, irrespective to the coordinates used in metadynamics. Besides, in the (θ, ϕ) framework the position of conformers and the definition of their basins is more straightforward than the equivalent analysis on the (γ, η) or (μ, ν) representations (as it is clear from Fig. 4.18). Considering Fig. 4.19(c), the ability in reconstruction of the chairs basins (the original poles) for the metadynamics calculations in (γ, η) coordinates is clearly visible, even

if a rough estimate (low number of Gaussians) is made. However, the ${}^{3O}B$ and B_{3O} conformers (corresponding to the new polar regions) are clearly under-sampled. The same scenario is present in the profile of Fig. 4.19(d), produced with respect to the (μ, ν) variables: an improved sampling for the ${}^4C_1$ and ${}^1C_4$ basins is present, with a consequent clear under-sampling of the actual poles position (the ${}^5S_1$ and the ${}^1S_5$ regions).

The general meaning of these findings is that the under-sampling at poles is unavoidable. Thus, a correction of metadynamics estimates in spherical coordinates is needed. Besides the Umbrella Sampling refinement, here the usage of different spherical coordinates is proposed: rotated coordinates could give free energy estimates with different systematic poles under-sampling, allowing to evaluate the order of magnitude of this effect. Moreover, an average procedure of the free energy reconstructions in the different spherical set could be addressed. Considering that each profile suffers from under-sampling in the local polar positions, a weighted mean method is desirable. Suitable chosen weights could indeed shield – or even suppress – the under-sampled areas in a profile with respect to the same area correctly sampled in another. The free energy estimate is the function

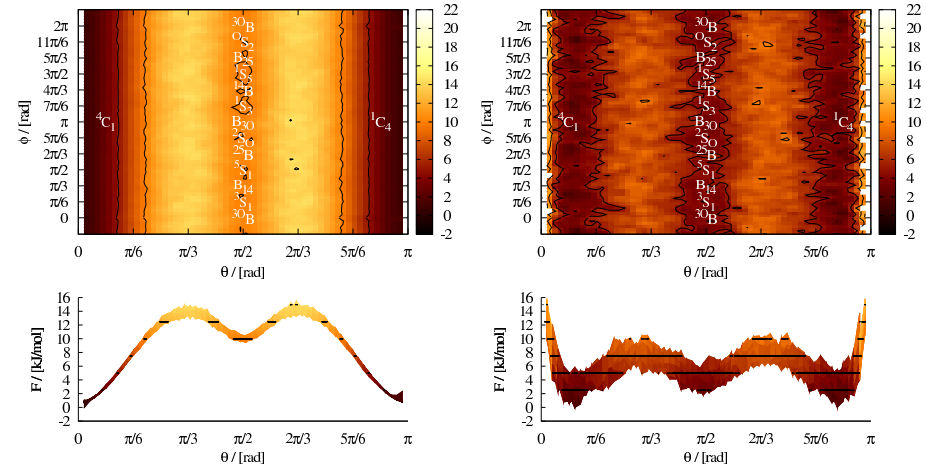
$$\mathcal{F}_{\mathrm{M}}(\theta,\phi) = \frac{F_{\mathrm{M}}(\theta,\phi)w_{\theta\phi} + F_{\mathrm{M}}(\gamma(\theta,\phi),\eta(\theta,\phi))w_{\gamma\eta} + F_{\mathrm{M}}(\mu(\theta,\phi),\nu(\theta,\phi))w_{\mu\nu}}{w_{\theta\phi} + w_{\gamma\eta} + w_{\mu\nu}} \tag{4.32}$$

if for example the profiles $F_{\mathrm{M}}(\theta,\phi)$, $F_{\mathrm{M}}(\gamma,\eta)$ and $F_{\mathrm{M}}(\mu,\nu)$ are at disposal. Since different metadynamics reconstruction are produced with different number of Gaussians, each profile has the structure $\widetilde{F}^i_{\mathrm{M}}(z) \simeq F(z) + c^i$, with a different additive constant c^i. The profiles are then comparable only when this constant is eliminated, otherwise in a weighted mean also the terms c^i will be weighted instead of remaining global additive constants. A simple way to do so is to refer each profile with the local value of a selected structure, *e.g.* the local minimum in the ${}^4C_1$ basin. Accordingly to this consideration, the presented profiles in Figs. 4.19(b) to 4.19(d) are shifted as $F^i_{\mathrm{M}}(z) = \widetilde{F}^i_{\mathrm{M}}(z) - \widetilde{F}^i_{\mathrm{M}}[{}^4C_1]$. In Fig. 4.20 the average $\mathcal{F}_{\mathrm{M}}(\theta,\phi)$ obtained from the metadynamics profiles of Fig. 4.19 with the weights

$$w_{\theta\phi} = \frac{\sin^2\theta}{2} \quad , \quad w_{\gamma\eta} = \frac{1-\sin^2\theta\cos^2\phi}{2} \quad , \quad w_{\mu\nu} = \frac{1-\sin^2\theta\cos^2\phi}{2} \tag{4.33}$$

is presented (also the weights are showed). Now the global aspect of the reconstructed free energy is in much better agreement with the reference MD reconstruction.

To summarize, a synthetic view of all the Cremer-Pople representations introduced in this thesis (in Table 4.2) and of transformation formulas between spherical sets (in Table 4.3) are presented.



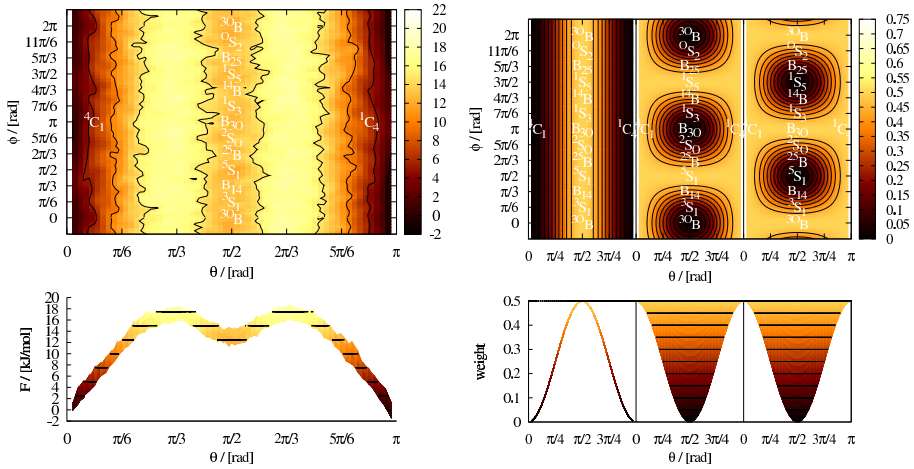
(a) plain MD

(b) Metadynamics in (θ, ϕ) coordinates

(c) Metadynamics in (γ, η) coordinates

(d) Metadynamics in (μ, ν) coordinates

Figure 4.19: Free energy landscapes in alternative spherical coordinates. The comparison of rough metadynamics estimates F_{META} (*i.e.* the estimate of Eq. (2.53) with low number of Gaussians) with the plain MD result on Toy-model N.1 is presented. All profiles are plotted in the (θ, ϕ) coordinates. Each profile is set to zero at the position of the minimum in the $^4C_1$ basin ($\theta < \pi/3$). Top: Plate Carrée projections, with isolines drawn every k_BT ($T = 300\,\text{K}$). The profile is replicated in two thin stripes at $\phi < 0$ and $\phi > 2\pi$ to stress the ϕ periodicity. Stable ideal conformers are also indicated. Bottom: projections of the free energy profile onto the $\phi = 0$ plane, with isolines drawn every k_BT. Darker colors corresponds to lower values of energy.



(a) Weighted average profile. The function $\mathcal{F}_M(\theta, \phi)$ (from Eq. (4.32)) is plotted.

(b) Weighting functions. From left to right the functions $w_{\theta\phi}$, $w_{\gamma\eta}$ and $w_{\mu\nu}$ of Eq. (4.33) are plotted

Figure 4.20: Free energy landscape from weighted average. (a): the profile is set to zero at the position of the minimum in the $^4C_1$ basin ($\theta < \pi/3$). Top: Plate Carrée projections, with isolines drawn every k_BT ($T = 300\,\mathrm{K}$). The profile is replicated in two thin stripes at $\phi < 0$ and $\phi > 2\pi$ to stress the ϕ periodicity. Stable ideal conformers are also indicated. Bottom: projections of the free energy profile onto the $\phi = 0$ plane, with isolines drawn every k_BT. Darker colors corresponds to lower values of energy. (b): the effect of shielding the polar regions can be appreciated.

Table 4.2: Extended Cremer-Pople puckering coordinate sets (summary). The original CP representations of Table 3.1 and the extra spherical representations of this Chapter are reported.

Cremer–Pople coordinates ($N = 6$) interconversion				
original			extended	
cylindrical (q_2, ϕ_2, q_3)	spherical(-q_z) (Q, θ, ϕ)	Cartesian (q_x, q_y, q_z)	spherical-q_x (Q, γ, η)	spherical-q_y (Q, μ, ν)
$q_2 \cos\phi_2$	$Q \sin\theta \cos\phi$	q_x	$Q \cos\gamma$	$Q \sin\mu \sin\nu$
$q_2 \sin\phi_2$	$Q \sin\theta \sin\phi$	q_y	$Q \sin\gamma \cos\eta$	$Q \cos\mu$
q_3	$Q \cos\theta$	q_z	$Q \sin\gamma \sin\eta$	$Q \sin\mu \cos\nu$

Table 4.3: Transformations between spherical Cremer-Pople sets. See Eq. (4.26), Fig. 4.18, and Table 4.2 for angle definitions

	(θ, ϕ)	(γ, η)	(μ, ν)
(θ, ϕ)	–	$\begin{cases} \cos\theta = \sin\gamma \sin\eta \\ \tan\phi = \dfrac{\sin\gamma\cos\eta}{\cos\gamma} \end{cases}$	$\begin{cases} \cos\theta = \sin\mu\cos\nu \\ \tan\phi = \dfrac{\cos\mu}{\sin\mu\sin\nu} \end{cases}$
(γ, η)	$\begin{cases} \cos\gamma = \sin\theta\cos\phi \\ \tan\eta = \dfrac{\cos\theta}{\sin\theta\sin\phi} \end{cases}$	–	$\begin{cases} \cos\gamma = \sin\mu\sin\nu \\ \tan\eta = \dfrac{\sin\mu\cos\nu}{\cos\mu} \end{cases}$
(μ, ν)	$\begin{cases} \cos\mu = \sin\theta\sin\phi \\ \tan\nu = \dfrac{\sin\theta\cos\phi}{\cos\theta} \end{cases}$	$\begin{cases} \cos\mu = \sin\gamma\cos\eta \\ \tan\nu = \dfrac{\cos\gamma}{\sin\gamma\sin\eta} \end{cases}$	–

Chapter 5

Alternative reaction coordinates for puckering

> And red and yellow and green and brown
> And scarlet and black and ochre and peach
> And ruby and olive and violet and fawn
> And lilac and gold and chocolate and mauve
> And cream and crimson and silver and rose
> And azure and lemon and russet and gray
> And purple and white and pink and orange
> And blue!!!
>
> A.L.Webber
> *Joseph's Coat* (Joseph and the Amazing Technicolor Dreamcoat)

In this Chapter metadynamics calculations on hexopyranoses with different puckering parametrizations are reported. This investigation of the characteristics of different coordinate choices in the context of the puckering problem suggests that the spherical Cremer-Pople description is the optimal choice in connection with accelerated sampling methods.

Contents

5.1 Spherical Cremer-Pople against Cartesian Cremer-Pople representation

In all the calculations presented in Chapter 4 we made use of the spherical (θ, ϕ) Cremer-Pople parameter set as collective variables in metadynamics. In this Section we will explore the performances of metadynamics with another Cremer-Pople representation, namely, the Cartesian one. This comparison was stimulated by the investigation of Biarnés and coworkers of the puckering free energy landscape of β-D-glucopyranose performed using Car-Parrinello metadynamics [24]. They showed that a reduction from 3 to only 2 puckering coordinates is somehow possible also in the Cartesian representation. Indeed, projecting the interconversion paths on the equator of the puckering sphere – that is, using the Stoddart representation [169] – a dimensional reduction from the Cartesian set (q_x, q_y, q_z) to the reduced one (q_x, q_y), similar to the usage of the angular subset (θ, ϕ) of spherical CP coordinates, seems feasible. This reduction is appealing, for it allows a metadynamics reconstruction in a phase space of dimension $m = 2$, as in the spherical angular case. However, in the following we will outline how the reduced Cartesian CP representation introduces strong biases in the reconstruction of the free energy profile. On the contrary, the Cremer-Pople coordinates (θ, ϕ) are the only suitable choice as reaction coordinates to perform an accelerated sampling (*i.e.* a free energy reconstruction with metadynamics).

5.1.1 Simulation methods

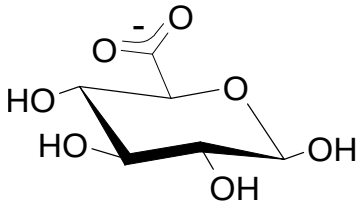


Figure 5.1: Glucuronic acid model. United-atoms model for β-D-GlcUA

The model compound for the comparison of metadynamics behavior against puckering parameterization is the glucuronic acid molecule (β-D-GlcUA). It is a monosaccharide derivated from glucose, where the $-CH_2OH$ group is substituted by a carboxylic group ($-COOH$, that in solution is most of the time found in the deprotonated, resonance stabilized form $-COO^-$, see Fig. 5.1). The free energy reconstructions of β-D-GlcUAp were performed employing both the Cartesian (q_x, q_y) and the spherical (θ, ϕ) CP parameters as reaction coordinates[1]. The molecule was modeled using the classical, united-atoms, G45a4 force field [112]. Partial charges on the side group $-COO^-$ have been neutralized by hand, in order to limit the check to metadynamics with different puckering parametrization (thus without the extra complications from electrostatics). For the other simulation parameters see Panel 5.1.

[1] For the definition of the CP sets see Section 3.4.

MD parameters	
time-step	1 fs
integrator	sd (Langevin)
friction coeff.	0.1 ps^{-1}
ref-T	300 K
non-bonded cutoff	1 nm
ε_r	80

no PBC, COM motion remotion (roto-translation), constrained bonds (SHAKE [153])

Well-tempered metadynamics parameters		
CV (ξ^1, ξ^2)	(θ, ϕ)	(q_x, q_y)
σ_1, σ_2	0.15 rad	0.005 nm
τ_G	200 step	100 step
w	1 kJ/mol	10.2 kJ/mol
ΔT	2000 K	2000 K
(Q, θ, ϕ)	variable monitored (τ_C = 50 step), adaptive width scheme (rescaling every 1000 step to a factor of 0.2 of the RMS-distance in the 1000 step previous window)	

Panel 5.1: Simulation Parameters for glucuronic acid system. Parameters for the reported well-tempered metadynamics simulation of a single molecule in vacuum, performed with the GROMETA [32, 20, 111] software (see Section 3.5). For three-dimensional metadynamics (see the beginning of Section 5.1.3), the extra variable is Q, with $\sigma_Q = 0.001$ nm .

During each run, independently of the collective variable used, the values (Q, θ, ϕ) of the have been collected. The starting conformation has been the $^4C_1$ chair for all simulations, apart from one in which the ergodicity has been tested starting the simulations from the $^1C_4$ conformer.

5.1.2 Direction of the meta-forces

We already stated that in the (θ, ϕ) representation the bias forces are always tangent to the puckering sphere "surface" (along the directions of $\hat{\theta}$ and $\hat{\phi}$). This aspect of the direction of meta-forces is the most critical one with the Cartesian (q_x, q_y, q_z) representation. Indeed, the bias forces associated with q_x and q_y are always pointing in a direction parallel to the equatorial plane. This has two simple but important consequences:

(i) if one only uses (q_x, q_y), the meta-forces are not able to force the system to move along the zenithal direction once the equator is reached. In this scenario, transitions across the equatorial line could be induced only by real forces and thermal fluctuations, without any contribution from the bias potential. If the free energy landscape presents a barrier higher than the thermal energy, it will be virtually impossible for the system to transit the equatorial line. This is why, for example, Biarnés and coworkers found that the system never explored the southern hemisphere during their simulation run: once the equator is reached the system is no more pushed in the right direction to perform the transition to the southern hemisphere. With this coordinates choice the method becomes not only non-ergodic in practice, but also gives severe biases in the reconstruction. Indeed, with this coordinate choice the strength of the meta-forces along the $\hat{\theta}$ direction decreases as $\cos(\theta)$. This means that the depth of free energy wells and height of free energy barriers along the radial direction in the $(q_x q_y)$-plane will be systematically overestimated. The error becomes more severe when the system is approaching the equator. The steep free energy barrier at

high values of $q_r^2 = q_x^2 + q_y^2$ which has been observed in Ref. [24] and which the metadynamics was unable to surmount, is precisely an artifact generated by this sampling with wrongly oriented meta-forces;

(ii) since the bias forces are not only softened along the $\hat{\theta}$ direction, but are also increased along the radial one $\hat{\mathrm{Q}}$, they will start forcing the system to explore regions with values of Q far from the equilibrium ones, as soon as the system departs from the polar regions. The reconstruction of the free energy profile will therefore be unavoidably biased, by the sampling of unwanted conformations at unphysical values of the total puckering amplitude Q. This is true even if the third component, q_z, is also employed. Meta-forces along the $\hat{\mathrm{q}}_z$ direction permit in principle to eliminate the problem of the forbidden transit of the equator. However, in this case one should compute the whole three-dimensional landscape, and then take the free energy density on a shell at constant Q, which is extremely more demanding.

We want to stress that the proposed problems are not related to puckering coordinates in general. On the contrary, these are issues connected to the property of the system to have a density of states concentrated in a thin spherical shell. To asses the proper sampling of this kind of configuration space, meta-forces have to be tangent to the sphere surface. Therefore, only the directions $\hat{\theta}$ and $\hat{\phi}$ – or linear combinations of them – are suitable to define the reduced space for a metadynamics reconstruction.

5.1.3 Simulation results

At first we present a check, for this specific case, of the validity of the assumption that the Q coordinate can be excluded. We performed a metadynamics using all three spherical coordinates (Q, θ, ϕ), and then we averaged out the two remaining degrees of freedom using[2]

$$F(Q) = -\frac{1}{\beta} \ln \int_0^{2\pi} \mathrm{d}\phi \int_0^{\pi} \mathrm{d}\theta \sin\theta \, \mathrm{e}^{-\beta F(Q,\theta,\phi)} \tag{5.1}$$

(for the presence of Jacobian term $\sin\theta$ see Section 4.2.2) thus obtaining the free energy density $F(Q)$, and the estimate for $\rho(Q) = \mathrm{e}^{-\beta F(Q)}$ reported in Fig. 5.2. Indeed, $\rho(Q)$ is characterized by a pronounced, narrow peak located around the value of 0.05 nm and is unimodal even in presence of a large side chain (the $-\mathrm{COO}^-$ side chain at C5 carbon atom).

In Fig. 5.3 the puckering free energy profile calculated with respect to (θ, ϕ) coordinates is presented. The full CV space has been spanned by the metadynamics run. A deep basin at $\theta \simeq 0$ is present, which contains the global minimum in correspondence of the ${}^4C_1$ conformation, around $\phi = 0$. The free energy basin is broader in two regions, providing some slightly distorted chairs around $\phi = \pi/2$ and $3\pi/2$. The next conformer which can be observed is a ${}^{3\mathrm{O}}B$ boat close to the border of the diagram ($\theta \simeq \pi/2$, $\phi \simeq 0$). In the middle

[2] Formally, this is a further dimensional reduction from the probability measure $\mu^{\xi}(\mathrm{d}Q\mathrm{d}\theta\mathrm{d}\phi)$ to the measure $\mu(\mathrm{d}Q) = \rho(\mathrm{d}Q)Q^2\mathrm{d}Q$ which density is $\rho(\mathrm{d}Q) = \mathrm{e}^{-\beta F(Q)}$.

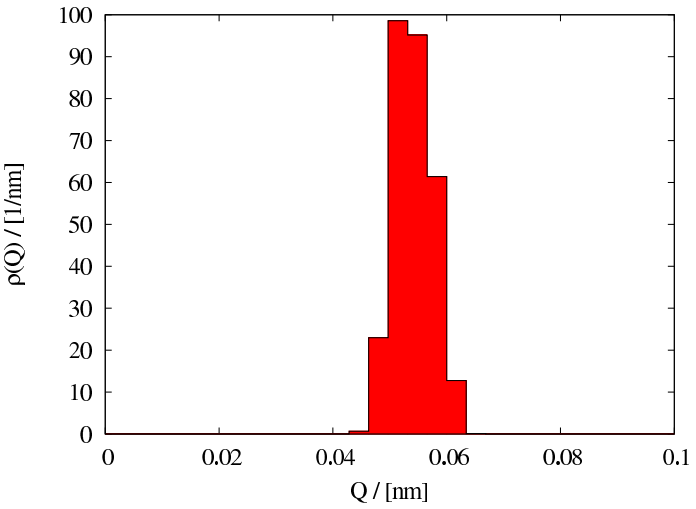


Figure 5.2: Probability density $\rho(Q)$. The probability density for the radial coordinate was obtained by an integral reduction of the three-dimensional profile $F(Q, \theta, \phi)$ (see Eq. (5.1)).

of the diagram ($\theta \simeq \pi/2, \phi \simeq \pi$) a local minimum corresponding to the B_{3O} conformer is also present. A steep barrier has then to be overcome, right after the equatorial line $\theta = \pi/2$, to reach the ${}^1C_4$ conformation, located around $\theta \simeq \pi$. By employing the spherical coordinates, the exploration of the southern hemisphere (the region $\theta > \pi/2$) of the puckering sphere was not a problem. The barrier located close to the equatorial line is definitely high ($\simeq 40\,\mathrm{kJ/mol}$), but the metadynamics technique could easily overcome it.

On the contrary, the problem of ergodicity is evident in metadynamics simulations with reduced Cartesian coordinates (q_x, q_y). During the free energy reconstruction, we tracked the value of $\cos(\theta)$, since the q_x and q_y coordinates alone do not allow to distinguish between northern and southern hemisphere. Therefore, we were able to compute the free energy landscape $F(q_x, q_y)$ as well as the distribution of the Gaussians placed in either of the two hemispheres $G^{\pm}(q_x, q_y)$ (see Fig. 5.4). Since we performed well-tempered metadynamics, the free energy estimate is both the sum of the deposed Gaussian (see Eq. (4.4)) and the logarithmic measure of the number of conformations z visited during a metadynamics run, $F(z) \propto \log\left[\int \delta(z - z_{\mathrm{B}}(t'))\mathrm{d}t'\right]$ (see Eq. (4.1)). However, the plots obtained by separating out the contributions $G^{\pm}(z) \simeq \log\left[\int \delta(z - z(t'))\vartheta(\pm\cos(\theta))\mathrm{d}t'\right]$ of the Gaussians deposited in the northern and southern hemisphere, respectively, are only indicative of the sampled regions (here $\vartheta(x)$ denotes the Heavside function). In other words, $F(q_x, q_y) \neq G^{+}(q_x, q_y) + G^{-}(q_x, q_y)$ by construction, because the two contributions $G^{\pm}$ are not to be considered as free energies. Nevertheless, the functions $G^{\pm}$ are quite informative (see Fig. 5.4, middle and lower panels): their graphs shows that only the northern hemisphere has been explored completely during the metadynamics run. The southern hemisphere has been reached, thanks to natural fluctuations, but the sampled region is relatively small. This result qualitatively reproduces the behavior observed by Biarnés and coworkers, who

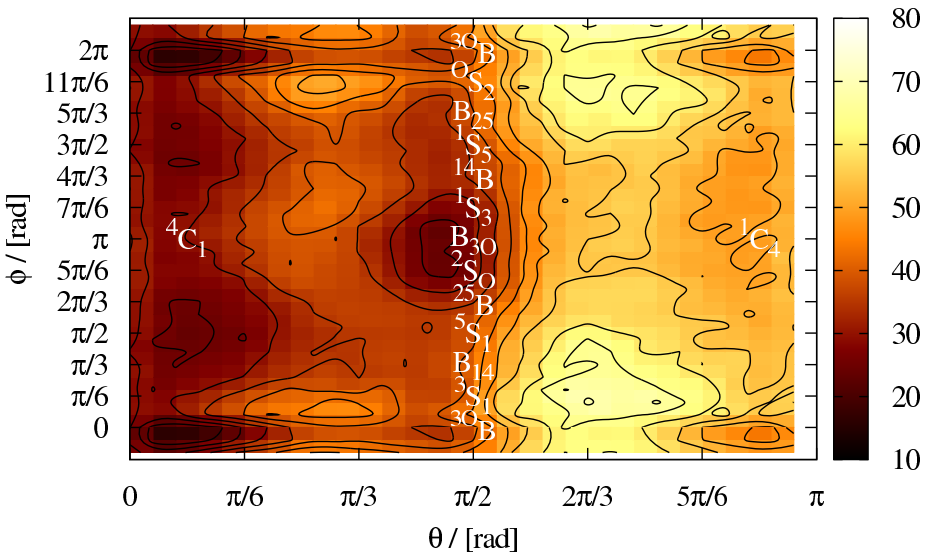


Figure 5.3: Free energy with the spherical CP representation. Puckering free energy $F_{\text{WTM}}(\theta, \phi)$ (see Eqs. (4.1) and (4.5), energies are in kJ/mol) of glucuronic acid model. The profile is set to zero at the position of the minimum in the ${}^4C_1$ basin ($\theta < \pi/3$). Plate Carrée projection, with isolines drawn every k_BT ($T = 300\,\text{K}$). The profile is replicated in two thin stripes at $\phi < 0$ and $\phi > 2\pi$ to stress the ϕ periodicity. Stable ideal conformers are also indicated.

noticed a sampling of the northern hemisphere only: in Ref. [24], the authors attributed this behavior to the presence of an extremely high free energy barrier at the equatorial line. In this way the metadynamics algorithm was definitely not able to be ergodic, even if in our simulations a tenfold longer simulation time and the same Gaussian heights have been used, with respect to the metadynamics with spherical coordinates. For comparison, the estimate of the barrier height at the equatorial line obtained using Cartesian coordinate is more than 90 kJ/mol, that has to be compared with the value of about 40 kJ/mol estimated using spherical coordinates. The presence of this artifact, whose origin was discussed in Section 5.1.2, prevents metadynamics to be ergodic. In Table 5.1 the differences in free energy of some selected conformations, with respect to the ${}^4C_1$ chair, presented in Table 5.1, show the magnitude of the error in ΔF values introduced by using Cartesian coordinates.

Another way to exhibit the artifacts introduced by the use of Cartesian (q_x, q_y) collective variables is to look at the effect that the mixing of tangent and radial coordinates has on the range of sampled conformations at different Q. In Fig. 5.5 the correlation between Q and the magnitude of the projected puckering vector $q_r \equiv \sqrt{q_x^2 + q_y^2}$ is presented. At low values of q_r, that is, at low values of the θ angle, the correlation with Q is only minimal: the radial coordinate Q is distributed around the value of 0.055 nm as in the free case. At values of q_r larger than 0.05 nm, on the contrary, a strong correlation develops, which almost completely correlates the two variables, along the $\sin(\theta) = q_r/Q$ line. This means that when the system is driven by metadynamics close to

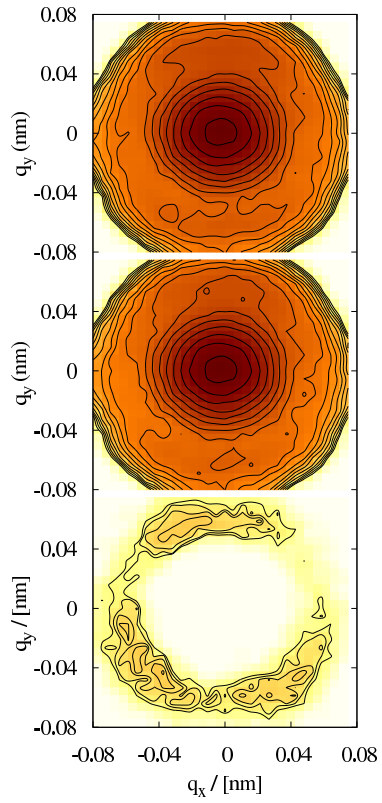


Figure 5.4: Free energy with the Cartesian CP representation. Puckering free energy $F_{\mathrm{WTM}}(q_x, q_y)$ (see Eqs. (4.1) and (4.5)) of glucuronic acid model. Upper panel: free energy landscape. Middle and lower panels: logarithm of the distribution of the Gaussians placed in the northern (G^+) and southern (G^-) hemispheres, respectively, during the metadynamics run. Every isoline corresponds to an increment in energy of 5 kJ/mol ($\simeq 2k_BT$ at room temperature). The figures in the middle and lower panels are not free energy surfaces, and in those cases isolines have to be intended only as a guide to the eye.

Table 5.1: Free energy from different CP set. Free energies ΔF of different conformations, estimated using the two different sets of spherical (†) and Cartesian (‡) coordinates, along with the location of the conformations on the $(\theta\phi)$-plane and (q_xq_y)-plane, respectively.

	$^4C_1$	$^1C_4$	^{3O}B	B_{3O}
$\Delta F^\dagger$	0.0	27.7	9.3	20.0
(θ, ϕ)	$(0.2, 6.1)$	$(2.9, 6.1)$	$(1.4, 0.0)$	$(1.5, 3.4)$
$\Delta F^\ddagger$	0.0	*n.a.*	28.7	33.1
(q_x, q_y)	$(0.0, 0.0)$	*n.a.*	$(0.0, 0.058)$	$(0.0, -0.064)$

Energies ΔF are in kJ/mol, θ and ϕ angles are in rad, q_x and q_y coordinates are in nm.

the equator of the puckering sphere, conformations that are sampled at higher values of θ are necessarily characterized by a non-physical, high total puckering amplitude. Thus, a strong systematic bias in the reconstructed free energy profile occur.

These findings clearly demonstrate that the non-ergodicity and the biases in free energy reconstructions are too a high price to be paid for the simplifications in representing conformations which derive from the use of Cartesian coordinates. In this sense our analysis suggests that only the spherical repre-

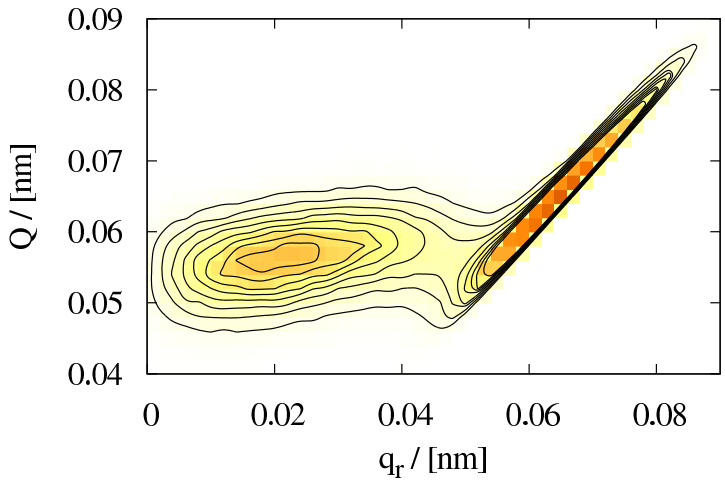


Figure 5.5: Artifacts from Cartesian coordinates. Bivariate distribution $\rho(q_r, Q)$ of the projection $q_r = \sqrt{q_x^2 + q_y^2} = Q\sin(\theta)$ versus the total puckering amplitude Q during a metadynamics run using (q_x, q_y) as CVs. A strong correlation develops while approaching the $\sin(\theta) = q_r/Q$ line. Lighter colors correspond to lower probabilities

sentation of the Cremer-Pople coordinates should be considered as a proper set of collective variables for enhanced sampling techniques, to study the puckered conformations of ring structures.

5.2 Spherical Cremer-Pople against Strauss-Pickett representation

In this Section we address the advantages and disadvantages of two different puckering description, the Cremer-Pople and Strauss-Pickett coordinates, from the perspective of computing the free energy surfaces and population of puckered conformers by means of metadynamics simulations. This comparison was stimulated by the investigation of Hansen and Hünenberger of the relative free energies of glucopyranose ring conformers by means of the Local Elevation Umbrella Sampling method [66]. They performed a complete three-dimensional exploration of the puckering space in the SP coordinates, giving the necessary indication for the counting of states in free energy basins. As we showed in Section 3.2.2, the Cremer-Pople and the Strauss-Pickett approaches (see Sections 3.1 and 3.2, respectively) to ring puckering have proven to be equivalent on the field of group theory analysis of closed rings [26]. Concerning the SP coordinates, the representation is intrinsically three-dimensional, because the three angles $(\alpha_1, \alpha_2, \alpha_3)$ of Fig. 3.1 measure the orientation of three "flaps" involving different atoms of the ring (the ideal IUPAC conformer are at positions indicated in Table 5.2).

We will show in the following how the ring puckering properties are correctly recovered in both the CP and SP framework, and at the same time we will stress

the advantage of the two-dimensional Cremer-Pople representation against the Strauss-Pickett one.

Table 5.2: Standard conformers in the SP representation. Location of carbohydrate conformations in the SP dihedral space (from Rao et al. [146]). The total space of ring conformers is the $[-90°, 90°] \times [-90°, 90°] \times [-90°, 90°]$ sub-volume.

Conformer	α_1	α_2	α_3
${}^4C_1$	$-35°$	$-35°$	$-35°$
${}^1C_4$	$35°$	$35°$	$35°$
B_{3O}	$30°$	$-60°$	$30°$
B_{25}	$30°$	$30°$	$-60°$
B_{14}	$-60°$	$30°$	$30°$
${}^{3O}B$	$-30°$	$60°$	$-30°$
${}^{25}B$	$-30°$	$-30°$	$60°$
${}^{14}B$	$60°$	$-30°$	$-30°$
${}^1S_3$	$60°$	$-60°$	$0°$
${}^1S_5$	$60°$	$0°$	$-60°$
${}^3S_5$	$0°$	$60°$	$-60°$
${}^3S_1$	$-60°$	$60°$	$0°$
${}^5S_1$	$-60°$	$0°$	$60°$
${}^5S_3$	$0°$	$-60°$	$60°$

5.2.1 Simulation details

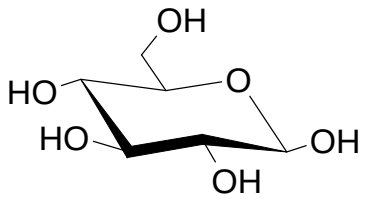


Figure 5.6: Glucupyranose model. United-atoms model for β-D-Glc.

We performed several simulations involving β-D-Glucose: molecular dynamics and combined metadynamics/umbrella sampling simulations, in presence and absence of solvent, and using (θ, ϕ) Cremer-Pople or $(\alpha_1, \alpha_2, \alpha_3)$ Strauss-Pickett coordinates. In all simulations glucose molecules have been modeled using the GROMOS 45A4 force field [112], while the SPC model [21] has been employed for water molecules. Even though newer versions of the GROMOS force field provide a better description of the puckering properties of sugars [10, 11, 65], the G45a4 force field has been chosen because the main goal of these comparisons is the behavior of the free energy reconstructions with respect to the coordinate choice.

The simulations performed will be conventionally called simulation I-III:

Simulation I is a 150 ns long molecular dynamics simulation of 512 glucose molecules in 25512 molecules of solvent. This represent a large enough statistical sample to observe conformational changes, allowing to check the prediction of the metadynamics/umbrella sampling approach by estimating equilibrium population of chair conformers and, in addition, transition rates. Configurations along the trajectory were sampled every 10 ps for further analysis (all other simulation parameters are collected in Panel 5.2);

MD parameters	
time-step	2 fs
simulation time	150 ns
integrator	md (leap-frog)
non-bonded cutoff	1.3 nm
T-coupling	Nosé-Hover [129, 74]
T-ref, τ_T	300 K , 1 ps
P-coupling	Parrinello-Rahamn [138]
p-ref, τ_p	1.013 25 bar , 1 ps
compressibility	4.5×10^{-5} bar^{-1}

PBC, COM motion remotion (roto-translation), dispersion correction (energy and pressure), constrained bonds (SHAKE [153], relative tolerance 0.1 %)

Panel 5.2: Parameters for Simulation I. Parameters for the reported MD simulation of 512 glucose molecules with 25512 SPC [21] water molecules, performed with the GROMACS [179] and PLUMED [27] software (see Section 3.5).

Simulation II is a metadynamics simulation of a single glucose molecule in vacuum. The bias potential was applied on the (θ, ϕ) coordinates. The free energy profile has been generated by calculating the value of the biasing potential on a 60×60 grid in the (θ, ϕ) space (all other simulation parameters are collected in Panel 5.3);

Simulation III is another metadynamics simulation of a single glucose molecule in vacuum, performed by applying the bias potential on the $(\alpha_1, \alpha_2, \alpha_3)$ SP dihedrals.The free energy profile has been generated by calculating the value of the biasing potential on a $60 \times 60 \times 60$ grid in the $(\alpha_1, \alpha_2, \alpha_3)$ space (all other simulation parameters are collected in Panel 5.3).

MD parameters	
time-step	1 fs
integrator	sd (Langevin)
friction coeff.	0.1 ps^{-1}
ref-T	300 K
non-bonded cutoff	1 nm
ε_r	80

no PBC, COM motion remotion (roto-translation), constrained bonds (SHAKE [153], tolerance 0.1 %)

Metadynamics parameters	
τ_G	0.1 ps
w	0.12 kJ/mol
Simulation II (10^5 Gaussians collected)	
CV (ξ^1, ξ^2)	(θ, ϕ)
σ_1, σ_2	0.05 rad
Simulation III (10^5 Gaussians collected)	
CV (ξ^1, ξ^2, ξ^3)	$(\alpha_1, \alpha_2, \alpha_3)$
$\sigma_1, \sigma_2, \sigma_3$	0.2 rad
(Q, θ, ϕ) variable monitored ($\tau_C = 50$ step)	

Panel 5.3: Parameters for Simulations II and III. Parameters for the Metadynamics/Umbrella sampling simulation of a single molecule in vacuum, performed with the GROMACS [179] and PLUMED [27] software (see Section 3.5).

5.2.2 The state-counting problem

Although Cremer-Pople and Strauss-Pickett coordinates are equally able to describe puckered conformers from a purely geometrical point of view, differences arise due to the shape of the puckering free energy hypersurfaces. Moreover, every coordinate system has to be complemented by a suitable assignment scheme for thermodynamic state. In this way different conformers – namely conformers with different values of (CP or SP) puckering coordinates – which are located in the same free energy basin, will represent a single conformer in the thermodynamical sense (see Section 2.2 for the related discussion).

In case of (θ, ϕ) CP variables, simple visual inspection of free energy profiles is sufficient to assign thermodynamic states. Conversely, quantitative schemes for the state counting are required for $(\alpha_1, \alpha_2, \alpha_3)$ SP variables. This can be easily seen by looking at the free energy profiles obtained from simulations II and III, presented in Figs. 5.7 and 5.8, respectively. For simulation II, which employed CP angles, the whole two-dimensional surface is at disposal in a single plot. For simulation III, which employed SP dihedral angles, the three-dimensional volume has to be presented with some relevant cross sections of the free energy volume, namely those at $\alpha_3 = -60°$, $-35°$, $0°$, $35°$ and $60°$. By looking at the free energy profiles, it is readily understandable how different could be the task of identifying features on the free energy profiles (conformers location, free energy basins and barriers) in the two coordinate systems.

Concerning assignment schemes to supplement the SP coordinates, in the Local Elevation Umbrella Sampling investigation of Hansen and Hünenberger [66], the authors proposed three quantitative schemes to assign a conformation, identified by the triplet of SP dihedrals $\beta = (\beta_1, \beta_2, \beta_3)$, to the i-th state $\alpha^i = (\alpha^i_1, \alpha^i_2, \alpha^i_3)$ (See Table 5.2 for the definition of the triplets defining different puckered ideal conformers):

(i) in the closest distance (CD) scheme, a conformation β is assigned to the i-th state if the quantity $d^i(\beta) = 1 - \alpha^i \cdot \beta / (\|\alpha^i\| \|\beta\|)$ is smaller than that of other states $d^j(\beta)$;

(ii) in the angular deviation (AD) scheme, a conformation β is assigned to the i-th state if it belongs to a cubic region around α^i defined by suitably chosen angle ranges. This scheme gives unassigned conformations, because cubic domains have not to overlap (see Ref. [66] for details);

(iii) in the minimum shape (MS) scheme states are identified by ellipsoids enclosing a local minimum and all the nearby conformations within an energy interval of 5 kJ/mol (the details are reported in Ref. [66]). Although the criterion is still partially geometrical – it assumes an ellipsoidal shape of the basins – it makes use of the information from the free energy landscape;

Among the three SP partitioning prescriptions, the MS approach is the counting scheme closest to the visual inspection approach used with CP variables, because it is not purely geometrical as the CD and the AD schemes are.

We chose to compare how three different schemes partition the respective conformational space: on one hand the visual inspection for CP coordinates;

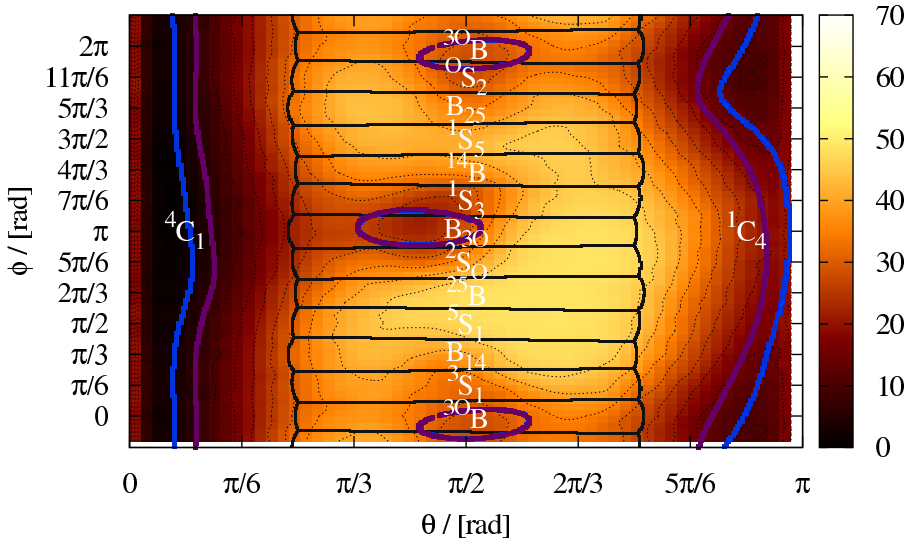


Figure 5.7: Counting schemes on the $(\theta\phi)$-plane. The underlying color map is the free energy surface of β-D-Glucopyranose obtained with the metadynamics/umbrella sampling approach [10] (Plate Carrée projection, with isolines drawn every $2k_BT$ at room temperature). The visual inspection scheme uses this profiles to select rectangular domains, with borders parallel to the θ and ϕ axes. Superimposed solid lines defines the regions determined by the CD scheme (black lines), projected onto the (θ, ϕ) plane. The names of ideal conformers associated to the different regions of the CD scheme are also shown. Blue and violet solid lines represent the projections of the ellipsoids for the MS scheme using a 5 to 10 kJ/mol threshold, respectively.

on the other hand, the CD and MS schemes for SP coordinates[3]. We perform this comparison in the common framework of the $(\theta\phi)$-plane. In this way the rectangular domains (*e.g.* the region $\theta < 60°$ for chairs) that naturally arise from visual inspection could be compared with the three-dimensional CD and MS partitions of the SP space projected onto the CP plane. The results of this projection is shown in Fig. 5.7. On one hand it is possible to see different sub-volumes of the SP space corresponding to the 14 states determined by the CD assignment. The 14 regions that corresponds to the ideal conformers of Table 5.2 have been determined by the following procedure. For each point of a regular grid of 500×500 points in the $(\theta\phi)$-plane, the full set of atomic coordinates (x_i, y_j, z_j) has been reconstructed – using bond length and bond angles of an ideal cyclohexane – following the procedure indicated in Ref. [40] (see also Section 3.2.3 and Appendix C). Then, SP dihedrals have been calculated from the atomic coordinates, and eventually the CD scheme has been applied[4]. From Fig. 5.7 becomes clear that, along the equator, states determined by their

[3] We omit the AD counting scheme because leave unassigned conformation on geometrical basis, and not for energetic reason.

[4] The computer code which computes atomic positions from CP parameters is made available at `http://www.science.unitn.it/~sega/sugars.html` .

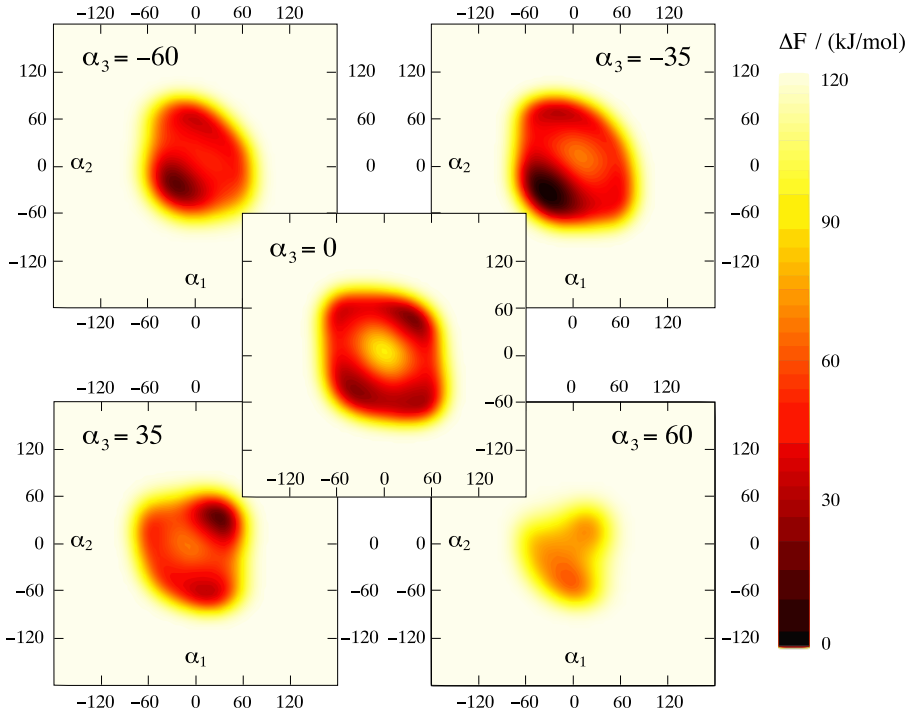


Figure 5.8: SP free energy surface. Cross sections (at $\alpha_3 = -60°$, $-35°$, $0°$, $35°$ and $60°$) of the puckering free energy profile obtained from Simulation III.

thermally accessible conformational volume encompass more than one state defined by geometry considerations. For example, in case of (a) the basin around $\phi = 0°$ covers a group of five geometrically different conformers (namely B_{25}, ${}^{O}S_2$, ${}^{3O}B$, ${}^{3}S_1$, B_{14}); in case of (b) the basin around $\phi = 180°$ groups four different conformers (namely ${}^{2}S_{O}$, B_{3O}, ${}^{1}S_3$, ${}^{14}B$). This is understandable due to ring flexibility on the pseudorotational path. Moreover, this confirms th observations made by Hünenberger and Hansen in their work: purely geometrical criteria such as CD and AD do not allow to identify thermodynamic states (intended as free energy basins) [66]. In this sense the CD scheme produces a miscounting in thermodynamic state, at least for flexible conformers. It has to be mentioned, indeed, that the regions associated with the two chair conformers, at the north and south pole, respectively, seem on the contrary to capture quite nicely the features of the underlying free energy profile. Thus, for rigid conformers no miscounting is registered.

We calculated with the same procedure also the projection of the MS ellipsoids obtained from the SP free energy profile onto the $(\theta\phi)$-plane, using two different threshold values (5 kJ/mol and 10 kJ/mol, respectively). As it can be seen from Fig. 5.7, the MS scheme is able to describe noticeably well the shape of the main four free energy basins. However, with the 5 kJ/mol threshold the full extent of the basins is still not covered. Changing the threshold to 10 kJ/mol affects only the assignment of chair conformer basins, but not the boat and skew ones, as only the latter basins are less deep than 5 kJ/mol.

After this semi-qualitative analysis, in order to check quantitatively the CD and MS schemes for chairs we compared the percentage of $^4C_1$ along the evolution of simulation I, estimated using the CD, MS and visual inspection schemes. In Fig. 5.9, the time evolution of the number of chair conformer, computed using the assignment criteria, is reported. The population of chairs and inverted chairs is practically indistinguishable for the CD and visual inspection schemes (it amounts to a relative root mean square difference between the methods of 0.02 % and 0.03 % for chairs and inverted chairs, respectively). This proves that SP dihedrals and the CD approach are quite robust for the identification of chairs. On the contrary, the MS scheme shows an interesting behavior. When using the 5 kJ/mol threshold, only roughly 66 % of chairs (and inverted chairs) with respect to the visual inspection or CD schemes are observed. This behavior means a clear miscounting of states for this scheme. However, since the measured populations are reduced by roughly the same ratio for both chairs and inverted chairs, the calculation of free energy difference using Boltzmann inversion formula gives a value of 10.0 kJ/mol, compatible with the other schemes (that gives the estimate $\Delta F = 9.2$ kJ/mol). By using the higher 10 kJ/mol threshold the situation could be improved, by means of an enlargement of the assigned basins of chairs (which gives now a roughly 90 % of chairs, much closer to the visual inspection value). However, it has to be noticed that since the smaller basins of the boat and skew conformers do not change shape, the evaluation of their free energy difference will be indeed affected by the choice of the threshold.

From Fig. 5.9 some additional informations could be extracted. Assuming a reaction kinetic of the first order, a fit to an exponential decay is possible. The fit shows that the population of chairs tends to the value of 97.7 % (with a corresponding rate of about 13 ns^{-1}) and showing still large oscillations on the 40 ns time scale. The estimated population corresponds to a free energy $\Delta F = 9.3$ kJ/mol, in good agreement with the estimate of 10 kJ/mol previously obtained by means of combined metadynamics/umbrella sampling [10].

5.2.3 Collective variable choice and convergence

Apart from the state-counting problem discussed so far, the main difference between the use of CP and SP coordinates should be the dimensionality of the puckering space. In general, much more time will be spent to fill the free energy profile with the biasing potential in the SP case, with respect to the CP one, for the simple reason that the SP space is three-dimensional, whereas the CP is only two-dimensional. This is because the time spent to sample a space grows exponentially with the number of its dimensions. More precisely, this is true only if the different subspaces have the same characteristics, from an energetic and geometrical point of view, otherwise the scaling of the reconstruction time could be less dramatic. For example, if we use (Q, θ, ϕ) instead of (θ, ϕ) variables only, the computational cost will not increase dramatically because the amount of phase space accessible at a certain energy will still have the shape of a spherical shell (see Section 4.1.1). Thus, simple dimensional arguments are not sufficient to estimate how the efficiency in filling the free energy profile depends on the set of collective variables in use.

One possible estimate might be obtained by checking how many depositions

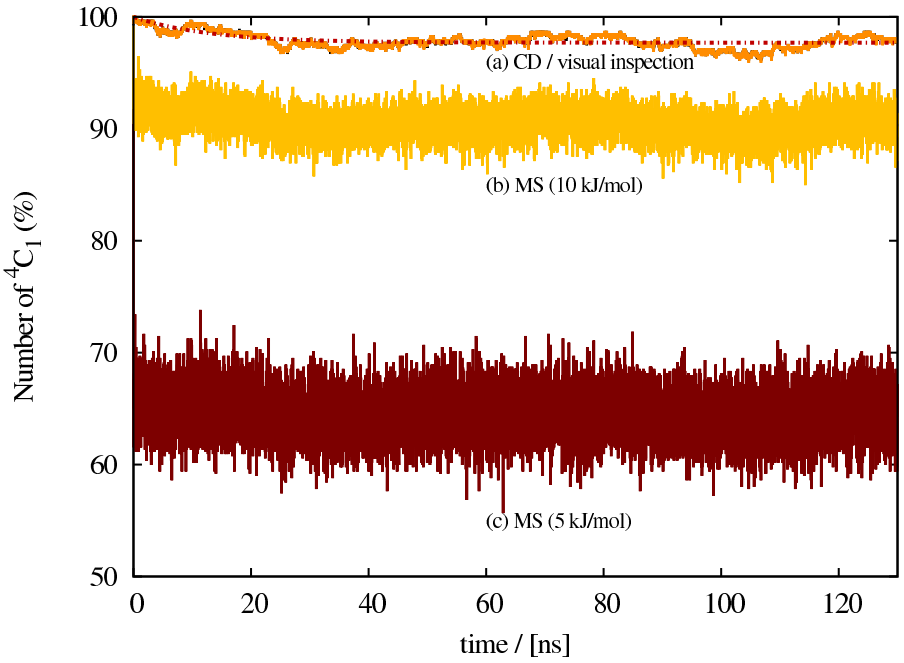


Figure 5.9: Time evolution of population $P[^4C_1]$. Data of simulation I are shown. Solid lines from top to bottom: (a) the CD scheme (almost coincident with the visual inspection scheme result); (b) the MS scheme with the 10 kJ/mol threshold; and (c) the MS scheme with the 5 kJ/mol threshold. A fit to an exponential decay (dot-dashed line) for (a) is also presented.

are needed in order to fill with Gaussians a given amount of the energetically accessible portion of the free energy profile, ΔF, once the system has become ergodic. Obviously, the portion filled by Gaussians for a given number of deposition will depend on the Gaussian height, but also on its width, and different coordinates require in general the use of different values for the Gaussians width. Differently from the heights, they are in general not commensurable. In order to provide a meaningful comparison, the number of deposition have to be rescaled by the volume of a reference feature, in units of the (height and) width of the Gaussians. For the puckering free energy landscape of glucose one could choose the volume accessible within one unit of thermal energy, k_BT from the minimum of a given conformer like, for example, ^{3O}B.

In Fig. 5.10 we present the equivalent number of depositions (that is, normalized according to the procedure described above) that belong to the test basin ^{3O}B (which reference volume was chosen as $0.4\,\mathrm{rad}^2$ and $1.2\times10^{-3}\,\mathrm{rad}^3$ in the CP and SP framework, respectively). The tendency for both CP and SP cases is to reach a plateau value, which denotes an homogeneous filling of the free energy landscape. A fit to an exponential decay has also been performed, which gives an estimate plateau value of about 120 and 2300 equivalent Gaussians for the case of CP and SP variables, respectively. This result shows that using CP variables can lead to an tenfold increase in efficiency in filling the free energy landscape with respect to the SP dihedral angles, when using the same

Gaussian height and an equivalent Gaussian width.

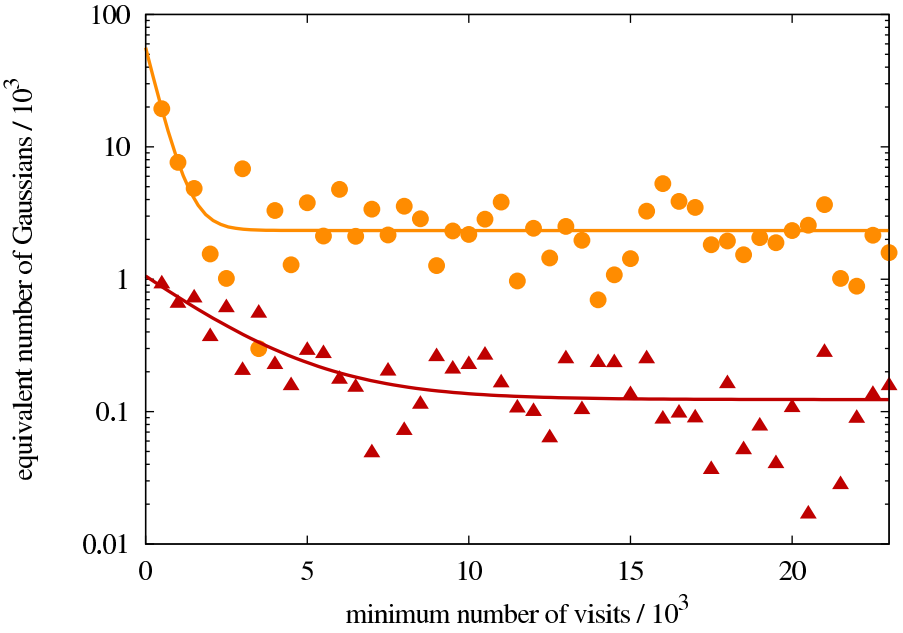


Figure 5.10: Filling equivalent efficiency. Efficiency of the CP and SP variables in filling the conformational free energy profile. Circles: CD scheme for SP dihedrals; Triangles: visual inspection criterion fo CP variables. The result of a fit to an exponential decay is also provided for each dataset.

By and large, CP and SP equivalence on the pure geometrical level does not correspond to a practical equivalence in accelerated dynamics simulations to compute puckering free energy landscapes. The need for only two (angular) puckering coordinates to sample in an ergodic way the conformational space in the Cremer–Pople approach is at the root of two main advantages with respect to the usage of Strauss–Pickett dihedrals: (i) energetically-separated conformational basins are readily identified in the CP scheme, whereas when employing the three SP dihedrals algorithmic partitioning of the conformational space are needed. Two of the checked partitioning schemes, namely the closest distance and the minimum shape, are shown to be prone to a misleading interpretation of the thermally accessible conformers of sugar rings. The reason for the miscounting in the closest distance scheme is related to its purely geometrical nature, while in the minimum shape scheme the miscounting is related to the choice of the energy threshold; (ii) employing CP coordinates rather than SP dihedrals leads to a tenfold decrease of the time needed to homogeneously fill the conformational space.

Chapter 6

Simulating α/β-D-pyranosides with the GROMOS force field

I have failed, I have been forsaken
I've been scorned and misunderstood
I have lost, my life has been taken
I'd surrender if only I could
And as I poise on the edge of life
Were time disappears
I bow in fear to the charm of the seer

Ayreon
The Charm of the Seer (The Final Experiment)

In this Chapter the systematic investigation of puckering properties of D-aldopyranoses with the GROMOS 45a4 force field is presented. A combined metadynamics/umbrella sampling approach is used to reconstruct the Cremer-Pople free energy landscapes for the whole series of α/β-D stereoisomers. A modification GROMOS 45a4 force field parametrization scheme is also presented, that improves the description of the puckering free energy differences between the most populated (chair and inverted chair) conformers.

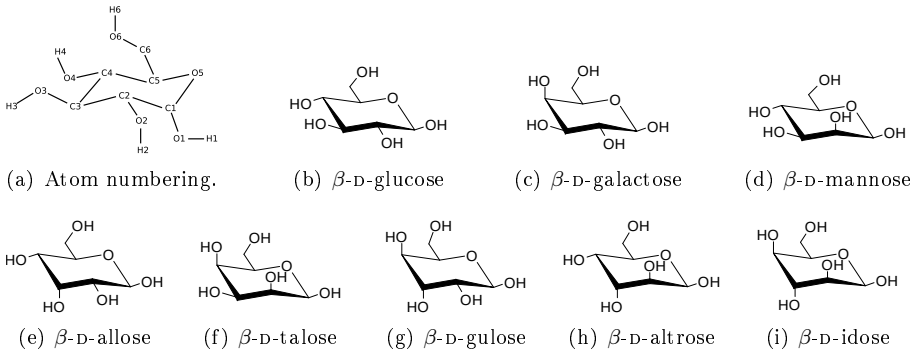
Contents

6.1 The sugar puckering problem in classical force fields

Within the framework of classical force fields, the number of computer experiments on saccharides has grown considerably in recent years, and various systems have been addressed [67, 68, 3, 176, 181, 104, 34, 124, 90, 123, 136, 137, 102, 178, 182, 190, 156, 187, 60, 135, 125, 114, 54, 93, 189, 2, 177]. Devising a realistic model of monosaccharides is obviously a decisive step in order for carbohydrates simulations to have enough predictive power. The accurate description of monosaccharides with classical force fields is not an easy task, because of the delicate interplay of different factors such as the presence of a high number of intramolecular hydrogen bonds, the competition of these hydrogen bonds with water-sugar ones and important steric and electrostatic effects between ring substituents in spatial proximity (see for example Ref. [95] and references within). The problem of reproducing some carbohydrates peculiarities, such as the rotameric distribution of the hydroxymethyl group or the anomeric and exo-anomeric effects, have been addressed in various force fields, and the reader can find some comparative analyses in Refs. [140, 18, 38]. However, considerably less attention has been so far devoted to the correct reproduction of the ring conformational properties. The problem of reproducing some carbohydrates structural properties – such as the rotameric distribution of the hydroxymethyl group or the anomeric and exo-anomeric effects – has been addressed for different force fields (the reader can find some comparative analyses in Refs. [140, 18, 38]). However, considerably less attention has been devoted so far to the correct reproduction of ring conformational properties. Despite the fact that biologically relevant monosaccharides almost always appear only in one stable ring conformer, several authors reported inappropriately high percentages of secondary puckered conformations [28, 96, 165, 112, 36, 109] when modeling carbohydrates using classical force fields such as GROMOS [92, 180, 94, 134, 165, 112] or OPLS-AA [43] for simulations of sugars in solution. For example, regarding one of the latest GROMOS parameter set for carbohydrates (the G45a4 parameter set [112]), conformers different than $^4C_1$ have been shown to be accessible during equilibrium simulation runs of β-D-glucose [95]. Moreover, two recent works [66, 166] estimated the chair/inverted chair free energy difference to be at least 10 kJ/mol lower than most theoretical and ab-initio simulation results. Besides equilibrium simulations, the importance of ring puckering has been proven by simulated pulling experiments, employed to interpret single molecule force-spectroscopy data [103, 125, 71]. In this case, ring conformational transitions simulated using three different carbohydrate force fields (AMBER94 [37], AMBER-GLYCAM [186] and CHARMM-Parm22/SU01 [49]) led to different interpretation of the same experimental data [125].

6.2 D-aldopyranoses: simulations and force field details

Metadynamics and equilibrium simulations have been performed for each of the 16 D-aldopyranoses, using a system composed of the respective sugar ring in a cubic simulation box filled with 504 water molecule. The schematic picture of



(a) Atom numbering. (b) β-D-glucose (c) β-D-galactose (d) β-D-mannose

(e) β-D-allose (f) β-D-talose (g) β-D-gulose (h) β-D-altrose (i) β-D-idose

Figure 6.1: Sugar models. (a): atom numbering scheme for hexopyranoses; (b) to (i): united-atom β-D-aldopyranose models. Models for the α-D-aldopyranoses are simply obtained by changing the orientation of $-$OH group at C1 from equatorial to axial (see Section 1.2).

the simulation models are given in Figs. 6.1(b) to 6.1(i), while the numeration of ring atoms used in the following is given in Fig. 6.1(a). Before starting each run, a 100 ps long molecular dynamics simulation with no bias has been performed to equilibrate the different sugars in their $^4C_1$ conformer. All the other simulation details for the metadynamics/umbrella sampling run are reported in Panel 6.1.

MD parameters	
time-step	0.2 fs
simulation time†	4 ns + 4 ns
integrator	md (leap-frog)
box size	$\sim$ 2 nm
non-bonded cutoff	1.0 nm
T-coupling	Nosé-Hover [129, 74]
ref-T, τ_T	300 K , 1 ps
P-coupling	Parrinello-Rahamn [138]
ref-P, τ_p	1.0 atm , 1 ps
compressibility	4.5×10^{-5} bar^{-1}

PBC, COM motion remotion (roto-translation), constrained bonds (SHAKE [153] for molecules, SETTLE [122] for solvent)

Metadynamics parameters	
CV (ξ^1, ξ^2)	(θ, ϕ)
σ_1, σ_2	0.15 rad
τ_G	200 step
w	0.5 kJ/mol
#Gaussians	10^5

Q variable monitored (τ_C = 200 step), histogram $\rho_B(\theta, \phi)$ from umbrella sampling on a grid of 60×60
† the duration is separated between the metadynamics and the umbrella sampling phase, respectively

Panel 6.1: Simulation Parameters for D-aldohexoses. Parameters for the Metadynamics/Umbrella sampling simulation of a single molecule with 504 SPC [21] water molecules, performed with the GROMETA [32, 20, 111] and PLUMED [27] software (see Section 3.5).

From the point of view of the G45a4 force field, the stereoisomers of glucose differ only slightly:

(i) the order of the two central atoms involved in an improper dihedral interaction (used to force the tetrahedral geometry around carbon atoms in united-atoms FF) at a chiral center has to be inverted, when a residue has to be moved from the equatorial to the axial orientation (*e.g.*: the improper dihedral to handle the chirality at carbon C4 is defined as C4–C3–O4–C5

for glucose and as C4–O4–C3–C5 for galactose);

(ii) α anomers differ from β anomers in torsional interactions on the O5–C1–O1–H1 dihedral, to account for the (exo-)anomeric effect;

(iii) sugars having O4 and C6 located on the same side of the ring plane ("Gal-like" sugars: galactose, talose, gulose, idose) are modeled with different parameters for the O5–C5–C6–O6 and C4–C5–C6–O6 dihedral angles with respect to those having O4 and C6 located on opposite side of the ring plane ("Glc-like" sugars: glucose, mannose, allose, altrose).

For the complete parameters of the G45a4 FF see Ref. [112].

6.3 Puckering free energy of β-D-glucose

Firstly we report on the combined metadynamics/umbrella sampling calculation of the puckering free energy profile of β-D-Glc. In Fig. 6.2 we present the free energy $F_{\text{M+US}}(\theta,\phi)$ as a function of the Cremer-Pople angular variables (θ,ϕ). On the bottom panel of Fig. 6.2, the lower contour of the colored region allows to easily identify the minima along the θ coordinate (the minima of this lower contour) and the transition states, or free energy saddle points (the maxima of the lower contour). On the $(\theta\phi)$-plane, ${}^4C_1$, ${}^1C_4$, and ${}^{3O}B$ conformers are clearly recognizable as minima basins. Due to the periodic nature of ϕ, the ${}^{3O}B$ basin appears duplicated across the $\phi = 0$ ans $\phi = 2\pi$ line (to emphasize this behavior, two thin stripes at $\phi < 0$ and $\phi > 2\pi$ replicates the periodic profile). Another local minimum basin, more shallow than the previous ones is located near the ${}^1S_3$ conformer. Both the basins on the equator line seems to include also some other near boat- and skewboat-like conformers. This kind of occurrence is a natural feature for the high flexible states on the pseudorotational path (located at $\theta = \pi/2$). We will not go in further details in the description of these local minima, because of their very high free energy ($\Delta F[{}^{3O}B] = 20.1(2)\,\text{kJ/mol}$ and $\Delta F[{}^1S_3] = 36.6(3)\,\text{kJ/mol}$, respectively).

In addition to the free energy profile, in Fig. 6.3 we show an analogous plot of the correction term $-k_BT\ln\rho_{\text{B}}(\theta,\phi)$ coming from the umbrella sampling phase. The screening of this contribution allows to check if residual meta-stabilities are left after the metadynamics phase, and whether an ergodic sample of the interesting region in the $(\theta\phi)$-plane has been performed or not. Residual meta-stabilities are recognizable as basins, while poorly sampled regions appear as peaks. In the case reported in Fig. 6.3 it is possible to see that the metadynamics run performed reasonably well. Indeed, the paths connecting the original metastable states do not show residual meta-stabilities, and the complete space of puckering angles has been sampled properly. Corrections to the free energy difference between ${}^4C_1$ and ${}^1C_4$ are $\simeq 4\,\text{kJ/mol} \sim 2k_BT$ in modulus. In case of the next most populated state, ${}^{3O}B$, the correction is again of the same order, while for other states the correction can reach $\simeq 14\,\text{kJ/mol} \sim 6k_BT$ with respect to the free energy of ${}^4C_1$. Even if the correction results to be quite high for some regions, the height of the residual saddle points (the maxima of the lower contour, bottom panel of Fig. 6.3) is definitly lower ($\sim 4k_BT$) than the same points on the free energy profile ($\sim 12k_BT$, from the maxima of the lower

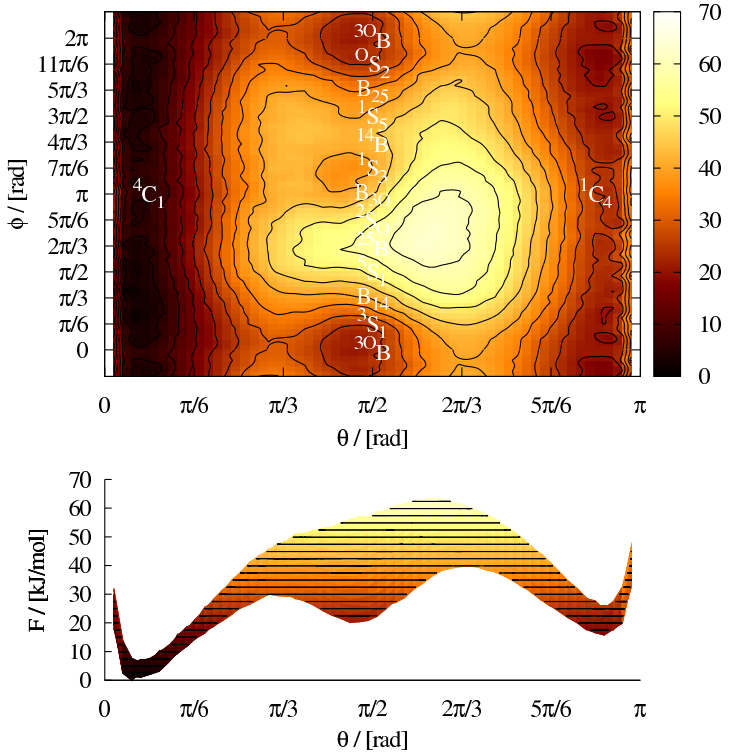


Figure 6.2: β-D-glucose free energy landscape. Puckering free energy $F_{\mathrm{M+US}}(\theta, \phi)$ (see Eq. (4.19)) of β-D-Glc using the GROMOS 45a4 parameter set is presented. The profile is set to zero at the position of the local minimum in the $^{4}C_{1}$ basin ($\theta < \pi/3$). Top panel: Plate Carrêe projection, with isolines drawn every $2k_BT$ ($T = 300\,\mathrm{K}$). The profile is replicated in two thin stripes at $\phi < 0$ and $\phi > 2\pi$ to stress the ϕ periodicity. Bottom panel: projection of the free energy profile onto the $\phi = 0$ plane, with isolines drawn every $1k_BT$. Darker colors correspond to lower energies.

contour in the bottom panel of Fig. 6.2). Moreover, it is worth mentioning that the largest corrections occur at transition states and around flexible conformers. This behavior is expected, as for those regions the assumption of the reaction coordinates being slows degrees of freedom ceases to be valid. The umbrella phase thus contributes in a significant way to the estimate of free energy differences, and appears to be more important for less populated states and for transition states.

As noted before, both the $^{1}C_{4}$ and ^{3O}B puckering free energies (local minima at about 15.8(2) kJ/mol and 20.1(2) kJ/mol, respectively) appear to be lower than what one would expect. In particular, the free energy of the inverted chair is about 10 kJ/mol lower than both our reference values of Refs. [5, 183] (see Table 1.2). This free energy difference leads to a inverted chair population of about 0.1 %, which can be observed during regular molecular dynamics runs at equilibrium. While not strictly incompatible with NMR experiments, which usually can predict populations with a $\simeq 2\,\%$ accuracy, this value is certainly strikingly

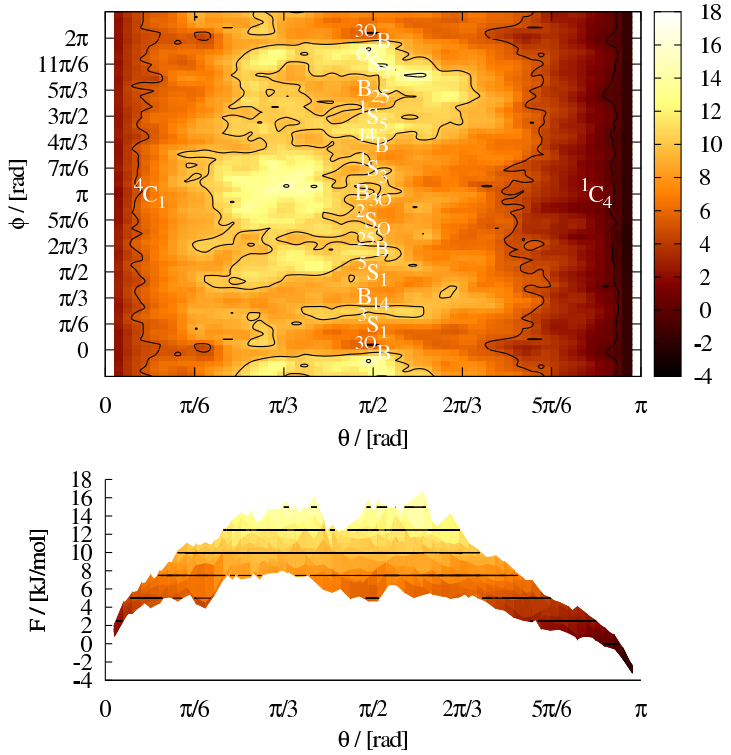


Figure 6.3: Contribution from equilibrium sampling after metadynamics. The sampling of $-k_BT \ln \rho_B(\theta, \phi)$ after metadynamics is presented for the β-D-Glc (G45a4 FF). The correction term is zero at the position of the local minimum in the $^4C_1$ basin ($\theta < \pi/3$). Top panel: Plate Carrèe projection, with isolines drawn every $2k_BT$ ($T = 300\,\mathrm{K}$). Bottom panel: projection of the free energy profile onto the $\phi = 0$ plane, with isolines drawn every $1k_BT$. Darker colors correspond to lower energies.

lower than all other theoretical estimates. This fact is even more important if one considers that β-D-Glc is supposed to have the largest $^1C_4$ free energy among aldopyranoses. It is thus expected that the $^1C_4$ conformers of the other stereoisomers could be characterized by even lower free energy differences[1]. A more direct connection with experimentally measurable quantities is performed through the calculation of the populations $P[S]$ of the thermodynamic basins S associated with each of the recognizable conformers. To this aim, the $(\theta\phi)$-plane has been partitioned in four regions (see Fig. 3.4(b)):

1. $\theta \in [0, \pi/3]$, associated to $^4C_1$;
2. $\theta \in [2\pi/3, \pi]$, associated to $^1C_4$;
3. $(\theta, \phi) \in [\pi/3, 2\pi/3] \times [-2\pi/3, 2\pi/3]$, associated to ^{3O}B;

[1] As it will be discussed in Section 6.4, this scenario is only partially correct, as unexpected patterns of the chair/inverted chair free energy difference appear along the series with the G45a4 FF.

4. $(\theta, \phi) \in [\pi/3, 2\pi/3] \times [2\pi/3, 4\pi/3]$, associated to the region around $^1S_3$.

With this partitioning the separating lines are located with good approximation along the maxima of the free energy surface (as is evident especially from the side-view of the free energy landscape, see Fig. 6.2) and, therefore, also close to the transition states. The conformer population $P[S]$ of a region S (calculated using Eq. (4.21)), for the $^4C_1$, $^1C_4$, ^{3O}B and $^1S_3$ conformations account to 99.858(1) %, 0.124(1) %, 0.0169(2) % and $3.23(4) \times 10^{-5}$ % of the total population, respectively (see also Fig. 6.4).

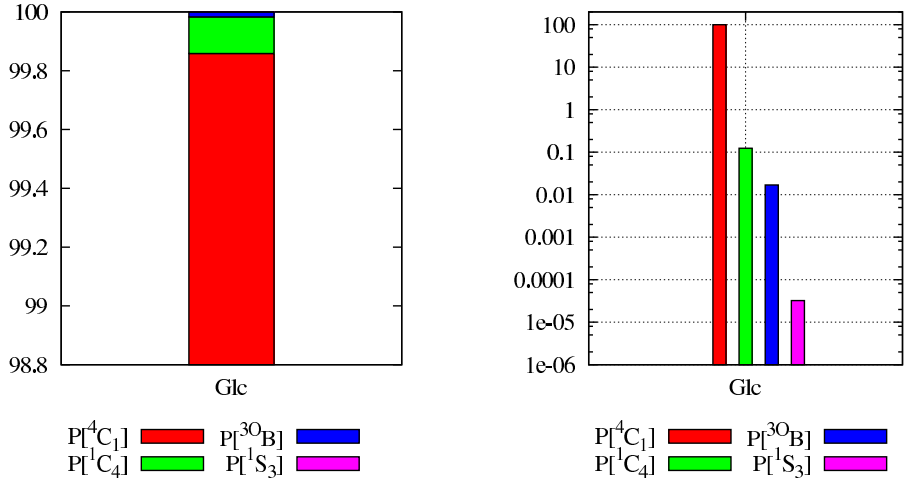


Figure 6.4: Population percentages for β-D-glucose conformers. To make the populations $P[S]$ (in %) of conformers S other than the dominant $^4C_1$ one visible, the histograms are reported in the zoomed view (left) and in the log-scale view (right).

A direct comparison of these data on β-D-Glc could be addressed with the values obtained by Hansen and Hüneneberger [66, 65] on the same system. In their works they reported values for the inverted chair free energy difference covering the range from 14.5 kJ/mol ± 0.8 kJ/mol to 16.5 kJ/mol ± 0.2 kJ/mol. Since in the approach of Hansen and Hünenberger free energies are derived from populations (by inverting Boltzmann formula[2] by means of Eq. (2.43)), we applied the same procedure and obtained a value of 16.67(2) kJ/mol. Thus, the different estimates are in agreement despite if the accelerated sampling approach and the collective variables used were different (the authors of Refs. [66, 65] used the Local Elevation Umbrella Sampling approach with the Strauss-Pickett CVs). Notice that, Spiwok and coworkers [166] in a recent metadynamics comparison of FF performances against β-D-Glc found the $^1C_4$ free energy of β-D-Glc to be 11.6 kJ/mol ± 1.8 kJ/mol for the G45a4 force field. Since they employed the (q_x, q_y, q_z) Cartesian Cremer-Pople coordinates as collective variables, we ascribe this important difference to the aforementioned problems related to the

[2]In principle our value of $\Delta F = 15.8(2)$ kJ/mol is not directly comparable, even if is compatible, because it is a point estimate $\Delta F = F(z_1) - F(z_0)$ that does not account for the shape of the free energy basin as the Boltzmann inversion does.

direction of the bias forces along the chosen collective variables (see the discussion in Section 5.1.2), and thus to possible systematic errors (that in our approach are under control by means of the equilibrium sampling, see the related discussion in Section 4.2.2).

6.4 D-Aldopyranoses with the G45a4 parameter set

The theoretical estimates of Table 1.2 predict a great variety in the conformational free energies for the 16 stereoisomers of glucose. For these reasons we decided to compute in a systematic way the puckering free energy landscape for the whole series of α/β-D-pyranoses, to check FF capabilities separately for each of the possible stereoisomers. Simulations of the remaining 15 steroisomers of glucose have been performed using the same protocol employed for β-D-Glc (see Panel 6.1). The results are summarized in Fig. 6.5 and in Table 6.1. The differences between data presented here and the data of Ref. [8] are due to further refinements of free energy estimates.

In Fig. 6.5 we report the free energy difference $\Delta F[^1C_4]$, with respect to the $^4C_1$ conformer, of α/β-D-pyranoses modeled using the G45a4 force field. Along with simulation data, the theoretical estimates of Angyal [5] and of Vijayalakshmi and Rao [183], and the experimental finding of Snyder and Serianni [163] and ours [10] are reported. Three horizontal lines are also drawn at $2k_BT$ intervals from the $F = 0$ level, with $T = 300\,\mathrm{K}$. Between them, the two lines at 0 and $2k_BT$ values indicate the thresholds below which the inverted chair population becomes greater than the chair one, and below which the inverted chair population becomes noticeable, respectively.

In Table 6.1, free energy differences and populations for the complete α and β series are listed. The free energy differences of the next leading conformer after $^1C_4$ are also reported. Concerning the populations of the next leading conformer, corresponding values were calculated on the $\theta \in [\pi/3, 2\pi/3]$ region, namely the whole equatorial region. This choice takes into account in some cases also possible other basins different than the next leading one, but their free energy is always so large (as it has been shown for β-D-Glc), that this approximation does not change substantially the population of the next leading conformer. Actually, such populations are all under 0.01 %, so are omitted from Table 6.1 for brevity.

The differences between the theoretical estimates and the simulation results obtained using the GROMOS 45a4 force field are striking. First of all, none of the 16 sugars investigated presents a chair/inverted chair free energy difference in quantitative agreement with the theory. Differences from the theoretical estimates are most of the time larger than 5 kJ/mol, and, in some cases, even larger than 10 kJ/mol. More important, many of these values are in marked qualitative disagreement not only with the theoretical results, but also with experimental evidence. In fact, α-D-Glc, α-D-Gal, α-D-Man, β-D-Tal present an inverted chair free energy which is between the thresholds of $4k_BT$ and $2k_BT$ at room temperature (see the dashed lines in Fig. 6.5). This means that a tiny population of inverted chairs, of the order of 2 % to 10 %, is expected at equilibrium, in

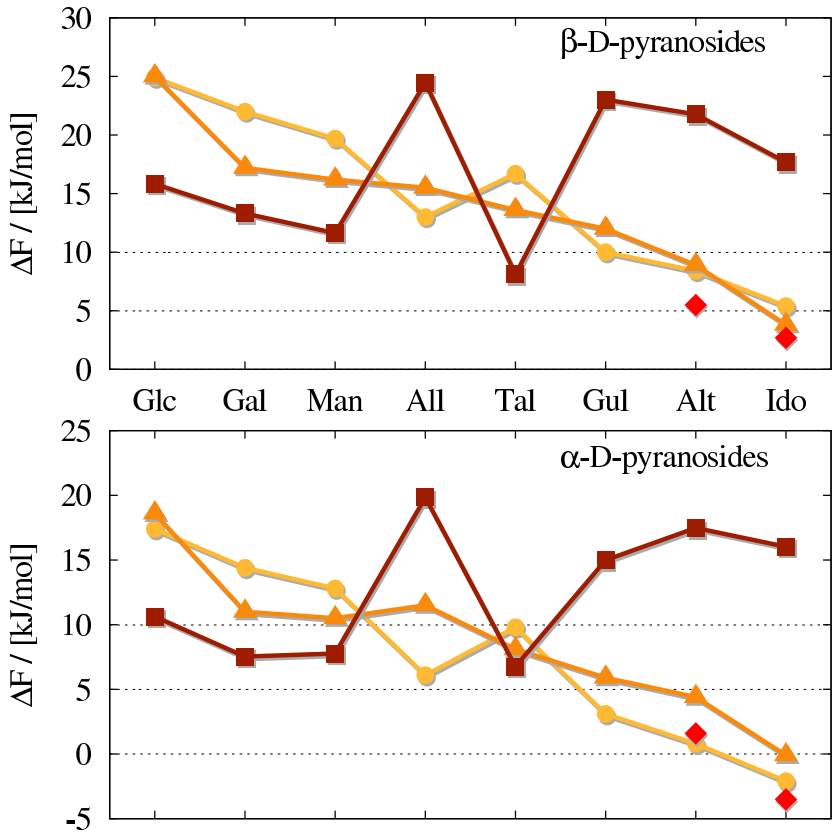


Figure 6.5: Inverted chair free energy difference with G45a4 FF. The β (upper panel) and α (lower panel) series of aldopyranoses are shown. The three curves refer to the simulation results obtained using the G45a4 parameter set (squares) and the predictions of Ref. [5] (triangles) and Ref. [183] (circles). Diamonds correspond to the estimates from NMR measurements (idose, Ref. [163]; altrose, Ref. [10]). Lines are a guide to the eye, and error bars are always smaller than the symbols (<0.5 kJ/mol).

contrast with no experimental evidence of the occurrence of this conformer. On the contrary, the puckering free energy of α/β-D-Gul, α/β-D-Alt, α/β-D-Ido inverted chairs results to be greater than 15 kJ/mol ($\sim 6k_BT$ at room temperature) within the G45a4 parameter set. This result rules out the possibility of observing the presence of inverted chairs in equilibrium simulations, in evident disagreement with theoretical and experimental findings. Finally, the general qualitative behavior of ΔF values against stereoisomery is poorly reproduced: only few indications of the theoretical trends reported in Fig. 6.5 are present, and the difference between α and β series is missed in some cases, like talose and idose.

Table 6.1: Free energy and population of different conformers using the G45a4 parameter set. The identification of the next leading conformer are indicated. Populations of the next leading conformers are in any case less than 10^{-4} %.

Isomer	$\Delta F[^1C_4]$	Next	ΔF[Next]	P$[^4C_1]$	P$[^1C_4]$
β-D-Glc	15.8(2)	^{3O}B	20.1(2)	99.858(1)	0.124(1)
β-D-Gal	13.3(2)	$^3S_1$	28.5(2)	99.62(3)	0.382(3)
β-D-Man	11.6(2)	$^OS_2$	25.5(2)	99.54(3)	0.453(3)
β-D-All	24.5(4)	^{3O}B	33.5(3)	99.9966(1)	0.0034(1)
β-D-Tal	8.1(2)	$^1S_3$	46.7(1)	97.03(2)	3.0(2)
β-D-Gul	23.0(2)	$^OS_2$	43.1(3)	99.9902(1)	0.0098(1)
β-D-Alt	21.8(2)	$^OS_2 \div {}^{3O}B$	37.2(4)	99.987(1)	0.0129(1)
β-D-Ido	17.7(3)	$^1S_5 \div B_{25}$	38.1(3)	99.92(1)	0.079(1)
α-D-Glc	10.6(3)	$^OS_2 \div {}^{3O}B$	31.8(3)	98.88(1)	1.12(1)
α-D-Gal	7.5(3)	$^3S_1$	39.9(2)	95.32(2)	4.68(5)
α-D-Man	7.7(3)	$^OS_2$	20.5(2)	94.5(1)	5.45(5)
α-D-All	19.8(4)	$^OS_2 \div {}^{3O}B$	40.8(3)	99.972(5)	0.028(5)
α-D-Tal	6.8(2)	$^1S_3$	35.4(3)	93.80(5)	6.2(1)
α-D-Gul	15.0(3)	$B_{25} \div {}^OS_2$	47.1(3)	99.828(2)	0.172(2)
α-D-Alt	17.5(3)	$^1S_3$	24.5(2)	99.901(1)	0.096(1)
α-D-Ido	16.0(2)	$^1S_3 \div {}^{14}B$	31.3(1)	99.783(1)	0.217(1)

Free energies ΔF are in kJ/mol, populations $P[S]$ are in %. Populations reported here differ qualitatively in terms of no more than 5 % with respect Ref. [10].

6.4.1 Equilibrium and out-of-equilibrium simulations

The GROMOS 45a4 force field appears to be unable not only to compare quantitatively with experimental and theoretical results, but – even more important – to reproduce the qualitative behavior of any of the two series. The inability of the force field to prevent appearance of inverted chairs at room temperature seems to be a severe drawback. Besides the static properties (*i.e.* the free energy differences), the kinetics of conformational transitions (*i.e.* the transition rates) is also an important aspect, for which a proper modeling of puckering free energy is relevant. In principle, if the inverse transition rate is much longer than the typical time interval spanned by a simulation, then alternative conformers might not be seen during equilibrium simulations. For example, Kräutler and coworkers [95] reported that in 200 ns long simulation runs of β-D-Glc, β-D-Gal, β-D-Man and β-D-Tal, all sugars but glucose remained for more than 99.9 % of the time in the chair conformation, while glucose was found in boat and twisted conformation for the 0.7 % of the time (giving a rough estimate of the characteristic time of escape from the chair conformer basin of 10 ns). However, even if 200 ns is a time much longer than that of most simulations, the pragmatic assumption that "unwanted" conformers do not appear likely is hazardous. This is because conformational transitions are stochastic events, and even a characteristic time of 10 ns might leads to a considerable amount of unphysical conformers in simulation runs much shorter than 200 ns, but with more than just one sugar

molecule in solution. We tested a setting which we consider to be representative of a typical simulation of medium to large size: a 25 ns long run of 512 β-D-Glc and 25×10^3 water molecules in a simulation box with an edge ~ 9.6 nm. With the exception of a 2 fs time-step and a non-bonded cutoff of 1.3 nm, the other simulation parameters are the same of Panel 6.1 without the metadynamics protocol. The appearance of both boats and inverted chairs conformers was observed, with a statistical frequency of 0.06 % and 1.1 %, respectively (see the upper panel of Fig. 6.6. These values are close to the one expected from the free energy calculations (see Table 6.1). Even if the time evolution of the number of $^1C_4$ conformers (lower panel of Fig. 6.6) indicates that equilibrium has not been reached yet, this result gives some information about the kinetics of ring conformational transitions in β-D-Glc. It shows that it is not unlikely to observe inverted chair conformers in equilibrium simulations of conventional size, using the G45a4 parameter set. We observed the appearance of inverted chairs also in similar simulations of β-D-Gal and α-D-Glc (data not shown), although to a much smaller extent, showing that the kinetics of the chair to inverted chair transition is much slower in these cases (consistently with the results of Ref. [95]).

6.5 Refining the force field: the 45a4-ASPG parameter set

6.5.1 Parametrization procedure

Following the results presented so far, we planned to re-parametrize the force field, to better reproduce puckering properties. The idea is to find a minimal set of changes that could fix at least the qualitative aspects discussed so far. The number of parameters on which puckering free energy depends is quite high. Indeed, in principle it depends directly on all ring atoms, and indirectly – but possibly to a considerable extent – on all ring substituents. A completely automated procedure is out of question, and one needs to adopt a heuristic approach with the following criteria:

(i) only FF parameters not directly involved in inter-molecular interactions should be tuned;

(ii) the changes should not be sugar-dependent;

(iii) the changes have to preserve previously known or already tuned molecular properties (rotameric distribution of the $-CH_2OH$ group and the conformation of the glycosidic linkages in disaccharides);

(iv) the inverted chair free energy of most common sugars (*i.e.* glucose, galactose, mannose) should be higher than 10 kJ/mol;

(v) the trend of the inverted chair free energy as a function of the sugar type, shown in Fig. 6.5, has to be reproduced;

(vi) the approximate offset between the inverted chair free energies of α and β anomers has to be reproduced.

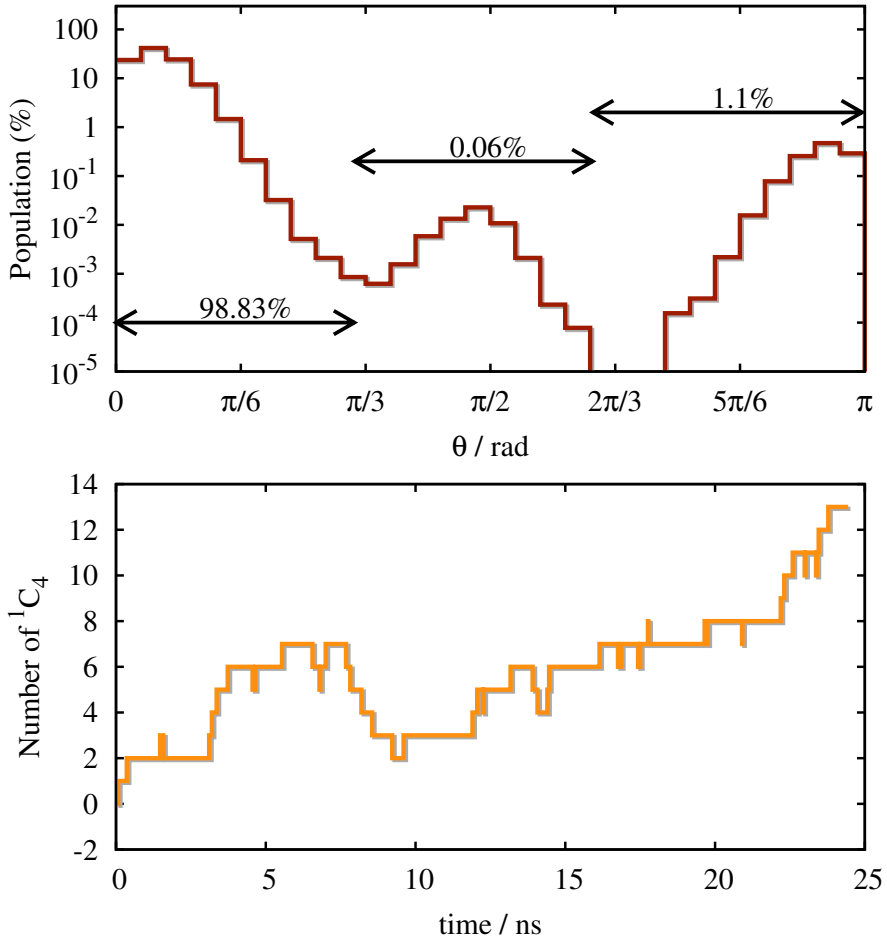


Figure 6.6: Equilibrium populations. Results of the unbiased molecular dynamics simulation of β-D-Glc. Histogram of the population along the θ puckering coordinate (upper panel) and time evolution of the number of inverted chairs (lower panel).

Within this framework, a complete, quantitative agreement for every sugar type is most probably not feasible. However, we found that quite reasonable results can be achieved with minimal parameter changes.

Point (ii) requires Lennard-Jones parameters and partial charges of the G45a4 parameter set to be preserved. Together with the requirement indicated at point (vi), this suggests that only angular or torsional interactions involving three or more ring atoms should be re-optimized. Changes in the stiffness of angular interactions did not change the puckering free energy in a significant way: even if Spieser and coworkers [165] showed that the height of the energetic barrier between $^4C_1$ and $^1C_4$ can be increased by stiffening the angular interactions, this only affects the free energy of the transition states, and not the free energy difference of the metastable conformers. Following point (iii), dihedral interactions directly involved in the rotameric distribution of the hydroxymethyl

group were thus not changed. Concerning the other torsional interactions, we noticed that every term involving three ring atoms (C1,C2,C3,C4,C5 or O5) and either C6 or any hydroxyl oxygen (O1,O2,O3,O4) was either not present or present with a phase term $\cos(\delta) = +1$ and multiplicity $n = 2$ in the torsional potential of Eq. (2.12) (see Table 6.2). As it was pointed out in Ref. [66], such a phase term favors the axial conformation of the substituent, with respect to the equatorial one, whereas a negative value of $\cos(\delta)$ would favor the equatorial conformation with respect to the axial one. Indeed, one of those interactions (O2–C2–C1–O5) has been eliminated in the 45a4 version of the GROMOS force field, for the precise purpose of stabilizing the $^4C_1$ conformer. While this is true for glucose, it is not, *e.g.*, for mannose, whose substituent in C2 is axial in the chair conformer. Therefore, the change in the G45a4 set that stabilized the glucose chair, acted in the opposite direction for mannose, stabilizing the inverted chair.

From the previous consideration on the single interaction on O2–C2–C1–O5, it is clear that any change affecting torsional parameters will have a profound (and sugar-specific) effect on the whole series of pyranoses. Therefore, to make grounded changes to the force field, one has first to understand how the pattern of axial/equatorial substituent in the $^4C_1$ conformer, for given torsional interactions involving the five chiral centers, can influence the properties of the $^1C_4$ free energy curves of the α and β series. In the following the main findings on global effects on puckering properties due to torsional parameters tuning are presented:

- first, tentative modification of torsional terms involving the O3 substituent were tried. The idea was to mimic a 1,3-diaxial interaction with these torsional terms. We found that changing the sign of $\cos(\delta)$ in the C1–C2–C3–O3 and C5–C4–C3–O3 interactions, one can at the same time rise the $^1C_4$ free energy of glucose, galactose, mannose, and β-talose, and lower that of allose, gulose, altrose and idose. Looking at the location along the free energy series of these sugars, these interactions are leading candidates to recover the approximate monotonous trend in the $^1C_4$ free energy and, consequently, to fix consequently point (v);

- another important role in determining the shape of the free energy curve is played by a dihedral interaction involving the substituent at the C5 chiral center (not present in the G45a4 FF). The aim was again to mimic a 1,3-diaxial interaction. However, the chirality of C5 is the same for all 16 steroisomers of D-Glc, and thus can be actually exploited to introduce a global shift for the $^1C_4$ free energies of the whole series of 16 D-pyranoses. This gives some freedom in our parametrization process, and it should be kept in mind that, most probably, parameters optimized this way are not valid to model the series of L-pyranoses;

- in addition to the interactions discussed so far, in order to fulfill point (vi) and reproduce correctly the gap between the $^1C_4$ free energy curves of the α and β series, the torsional interaction for the C3–C2–C1–O1 present in the G45a4 set has to be modified. The chirality of the C1 carbon atom differs only between the α and β anomers, and is certainly playing a role in modulating the height of the free energy gap between α and β anomers;

- corrections to the energy term associated to the dihedral C4–C3–C2–O2 were proven to slightly enhance the agreement to the theoretical estimates.

In summary, the phase $\cos(\delta)$ and amplitude k_ϕ of the C3–C2–C1–O1, C4–C3–C2–O2, C1–C2–C3–O3, C5–C4–C3–O3 and C1–O5–C5–C6 torsional interaction were tuned in order to obtain $^1C_4$ free energy curves (*i.e.* the ΔF against stereoisomery curves of Fig. 6.5) in better agreement with experimental and theoretical data. The set of torsional interactions and their parameters for the proposed modification with respect to the G45a4 parameter set are reported in Table 6.2, and the full topologies in GROMACS format are provided at `http://www.science.unitn.it/~sega/sugars.html`. Notice that the strength of

Table 6.2: FF parameters. The list of possible torsional interaction involving three ring atoms for D-aldopyranoses are presented. The functional form for the torsional interaction is that of Eq. (2.12). Parameters of the G45a4 torsional parameters and modified ones (accounted as 45a4-ASPG) are presented. Values for the constant k_ϕ are in kJ/mol

dihedral angle	45a4			45a4-ASPG		
	k_ϕ	$\cos\delta$	n	k_ϕ	$\cos\delta$	n
O5–C1–C2–O2	–	–	–	–	–	–
O1–C1–C2–C3	0.418	1	2	0.5	−1	2
O2–C2–C3–C4	0.418	1	2	0.5	−1	2
C1–C2–C3–O3	0.418	1	2	2.4	−1	2
O3–C3–C4–C5	0.418	1	2	2.4	−1	2
C2–C3–C4–O4	0.418	1	2	0.418	1	2
O4–C4–C5–O5	–	–	–	–	–	–
C3–C4–C5–C6	–	–	–	–	–	–
C6–C5–O5–C1	–	–	–	0.5	1	2

the C1–C2–C3–O3 and C5–C4–C3–O3 torsional interactions had to be set to a much higher value ($k_\phi = 2.4\,\mathrm{kJ/mol}$) than that of the other ones involving three ring atoms and one hydroxyl oxygen. Given the similar chemical nature of the quadruplets of atoms involved (beside that of the anomeric oxygen), such asymmetry appears to be peculiar. This might originate either from the fact that the O3 oxygen can be involved in the 1,3-trans-diaxial interaction with the hydroxymethyl group (stronger than that with any other OH groups) or from other interaction terms already present in the G45a4 set, whose influence on the puckering free energy is not yet understood, and that might deserve a separate investigation.

It is worth mentioning that in very recent times a full re-optimization of force field parameters for carbohydrates has been proposed by Hansen and coworkers [65]. It is a parametrization procedure that, starting from scratch, aim to tune force field parameters not only on previously considered properties (*e.g.* quantomechanical charges, rotameric distributions, ...) but also for puckering properties. To the best of our knowledge, this is the first attempt to use free energy information (obtained with the Local Elevetion Umbrella Sampling accelerated sampling method) in force field parametrizations.

6.5.2 Free energy landscape with the new 45a4-ASPG parameter set

The 45a4-ASPG parameter set for sugars was devised with the aim not to change properties other than puckering. This choice alone, however, does not guarantee that quantities other than puckering are not affected. We checked explicitly that these modifications did not affect the rotameric distribution of the hydroxymethyl group, calculating the free energy profile of the C4–C5–C6–O6 torsional angle (ω) for the G45a4 and 45a4-ASPG sets of parameters. The results obtained with the two parameter sets did not differ more than 2% between each other (data not shown). The free energy surfaces have been calculated using the same metadynamics/umbrella sampling approach employed for the calculation of the puckering free energy (see Panel 6.1), but using a Gaussian width of 0.1 rad for biasing the ω variable.

We summarized the results obtained using our modified set of parameters in Figs. 6.7 and 6.8 and Table 6.3. In Fig. 6.7 the free energy differences between inverted chair and chair conformers are compared again with the theoretical predictions of Ref. [183] and Ref. [5]. The improvement with respect to the results of Fig. 6.5 is striking. Both the α and β series reproduce now the qualitative trend of the theoretical estimates. galactose, mannose, and talose are not anymore between the $4k_BT$ and the $2k_BT$ thresholds, and the value of gulose, altrose and idose free energies diminished considerably. Also the gap between the α and β anomers is now reproduced reasonably well, being on average $\simeq 7$ kJ/mol. Concerning the conformer populations calculated from the free energy landscape, the new parameter set gives a sensible improvement in reliability with respect to the G45a4 one. To highlight this feature, in Fig. 6.8 the comparison between the population of chair and inverted chair are presented. As it clearly visible, with the G45a4 FF the preferred conformer is almost in any case the $^4C_1$, with very tiny populations of $^1C_4$ conformers for α/β-D-Tal, α-D-Gal and α-D-Man. In particular, there is no clear difference between anomers α and β in conformational preferences, nor any sing of a variability along the series of stereoisomers with given anomery. On the contrary, with the 45a4-ASPG FF the differences between anomers and the features of stereoisomers show a much better agreement with theoretical predictions. In particular, for β-D-aldopyranoses only altrose and idose seem to populate the $^1C_4$ conformation, while for α-D-aldopyranoses the probability $P[^1C_4]$ increases as expected progressively for stereoisomers with higher number of axial ring substituents in the chair form. Unfortunately, the population of inverted chairs in α-D-Ido is still lower than that of chairs, whereas both theory and experiment show a preference for the $^1C_4$ conformer. However, given the fact that these changes are not sugar-specific, the result obtained is, to our opinion, still remarkable, as the ability to reproduce puckering properties has increased dramatically, with respect to the G45a4 set.

To conclude, the complete set of all free energy profiles obtained with the 45a4-ASPG parameter set are presented in Figs. 6.9 and 6.10. On these profiles, the next leading conformer listed in Table 6.3 could be identified as metastable basins, and possibly transition states can be evaluated. However, we will not go in further details with these analysis, as the main goal of this refinement for the force field parametrization is the correct accounting of chair/inverted chair properties. By and large, our parametrization corrects the main inaccuracies of

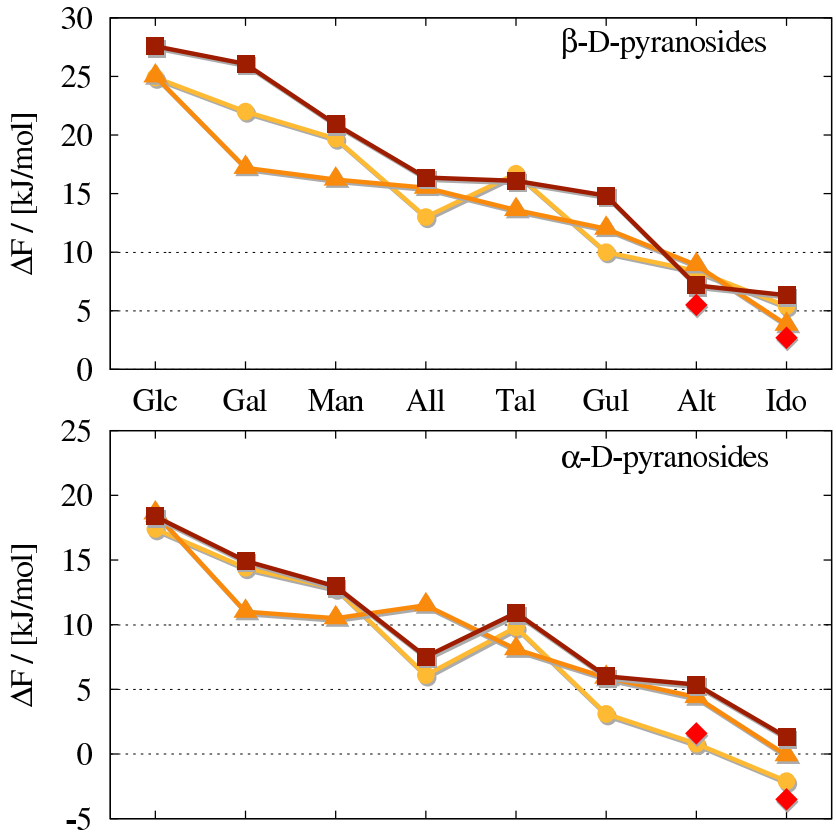


Figure 6.7: Inverted chair free energy difference with 45a4-ASPG FF. The β (upper panel) and α (lower panel) series of aldopyranoses are shown. The three curves refer to the simulation results obtained using the 45a4-ASPG parameter set (squares) and the predictions of Ref. [5] (triangles) and Ref. [183] (circles). Diamonds corresponds to the estimates from NMR measurements (idose, Ref. [163]; altrose, Ref. [10]). Lines are a guide to the eye, and error bars are always smaller than the symbols ($< 0.5\,\mathrm{kJ/mol}$)

the G45a4 FF against puckering properties. Free energy differences are now in accordance with experimental and theoretical data always within 5 kJ/mol, but in most cases within 2.5 kJ/mol ($\sim 1 k_B T$). Within this re-parametrization, a closer agreement with the theoretical models is probably not needed, given the uncertainties to which they are subjected. Indeed, while the theoretical models do not provide any confidence interval, an approximate picture can be obtained by comparing them with the experimental data on idose [163] and altrose [8], which roughly fall within a 3 kJ/mol interval. The improvement attained with the introduction of this new parameter set is to our opinion substantial, and reached the goal of reproducing known puckering free energy differences while keeping other properties, such as inter-molecular interactions and the rotameric distribution of the hydroxymethyl group, unchanged. These modifications to the GROMOS force field will allow to perform more realistic simulations of D-aldopyranoses. They represent a certain improvement in the study of carbohy-

Table 6.3: Free energy and populations of different conformers using the 45a4-ASPG parameter set. The identification of the next leading conformer are indicated (see also Figs. 6.9 and 6.10). Populations of the next leading conformers are in any case less than 10^{-4} %.

Isomer	$\Delta F[^1C_4]$	Next	ΔF[Next]	P[$^4C_1$]	P[$^1C_4$]
β-D-Glc	27.6(3)	$^{O}S_2 \div ^{3O}B$	32.6(3)	99.998 43(1)	0.001 40(1)
β-D-Gal	26.1(3)	$^{3}S_1$	39.7(2)	99.997 58(2)	0.002 41(2)
β-D-Man	20.9(4)	$^{O}S_2$	35.6(3)	99.9789(3)	0.0210(3)
β-D-All	16.4(2)	^{3O}B	25.0(4)	99.810(1)	0.118(1)
β-D-Tal	16.1(3)	$^{1}S_3$	48.5(4)	99.910(1)	0.0904(7)
β-D-Gul	14.8(2)	$^{O}S_2 \div ^{3O}B$	35.5(2)	99.684(2)	0.316(2)
β-D-Alt	7.2(2)	$^{O}S_2$	28.7(2)	95.70(4)	4.30(3)
β-D-Ido	6.3(2)	$B_{25} \div ^{O}S_2$	31.8(2)	94.91(4)	5.09(4)
α-D-Glc	18.4(2)	$^{1}S_3 \div ^{14}B$	35.6(3)	99.9530(3)	0.0470(3)
α-D-Gal	14.9(2)	$^{1}S_3$	42.9(2)	99.756(2)	0.245(2)
α-D-Man	13.0(3)	$^{O}S_2$	26.3(2)	99.552(4)	0.447(4)
α-D-All	7.5(2)	$^{O}S_2 \div ^{3O}B$	29.3(2)	95.38(4)	4.62(3)
α-D-Tal	10.9(2)	$^{1}S_3 \div B_{3O}$	35.5(2)	98.67(1)	1.33(1)
α-D-Gul	6.0(2)	$^{O}S_2$	36.8(3)	91.95(7)	8.05(7)
α-D-Alt	5.4(3)	^{3O}B	19.1(2)	86.1(1)	13.9(1)
α-D-Ido	1.3(2)	$^{O}S_2$	24.0(4)	53.0(2)	47.0(2)

Free energies ΔF are in kJ/mol, populations $P[S]$ are in %. Populations reported here differ qualitatively in terms of no more than 5 % with respect Ref. [10].

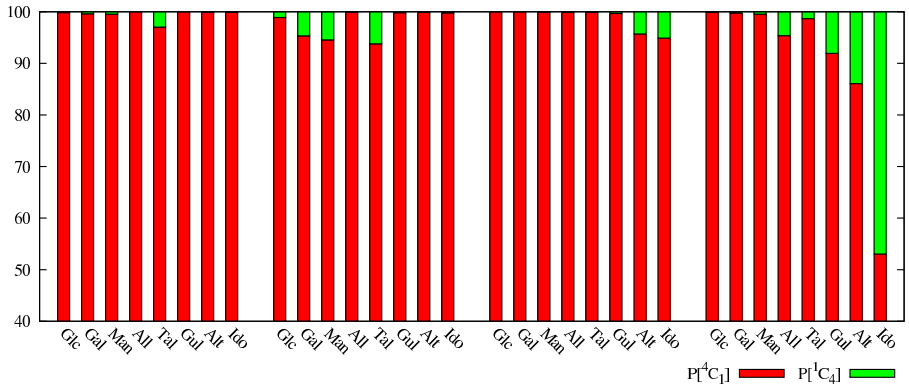
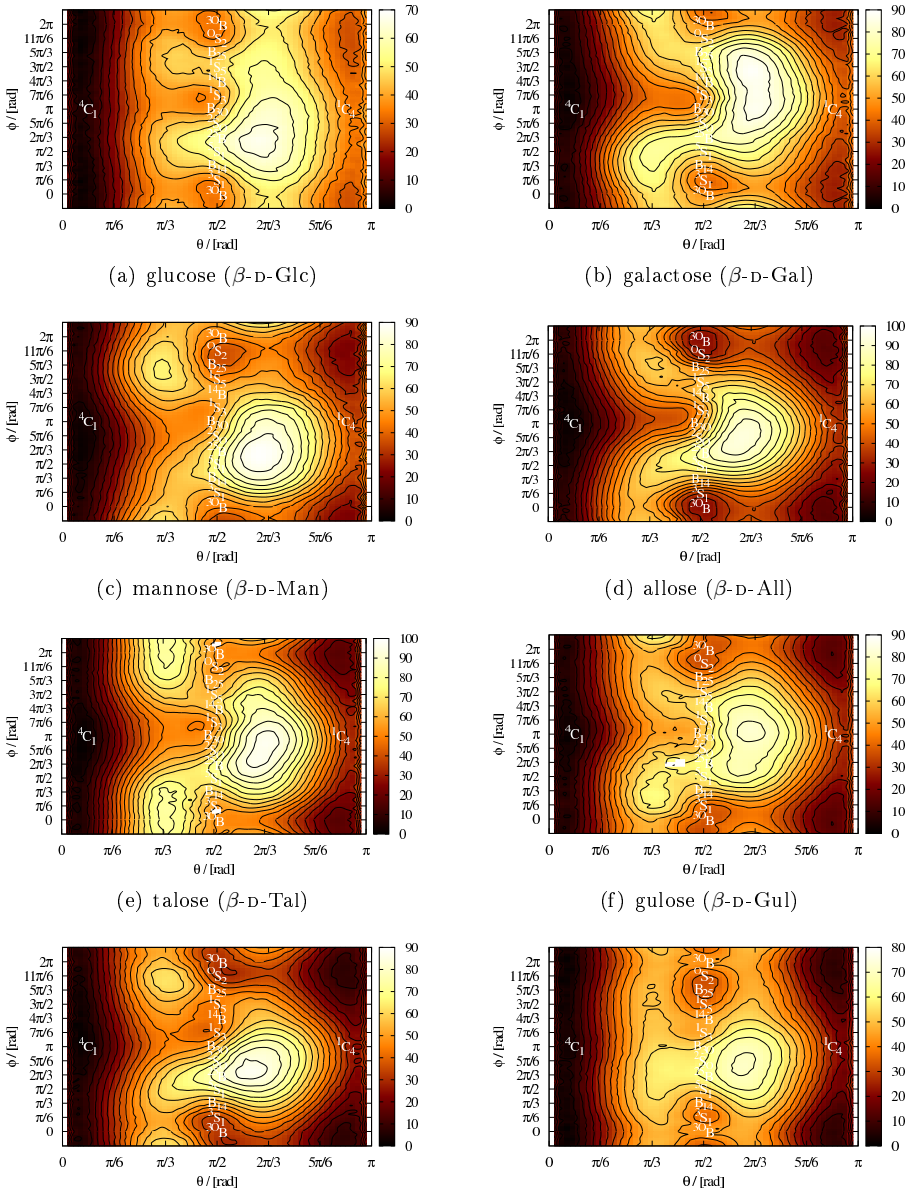


Figure 6.8: Comparison of conformer populations. Conformer populations of $^4C_1$ (red) and $^1C_4$ (green) conformers calculated from metadynamics simulations are presented. The region from 0 % to 40 % is omitted to emphasize the $^1C_4$ populations. Left half: G45a4 FF results. Right half: 45a4-ASPG results

drate equilibrium properties, but will have an even more important impact on the evaluation of out-of-equilibrium properties, such as in the case of simulated AFM pulling experiments. Still, much has to be done regarding the puckering

properties of carbohydrates. In particular, the role of skew conformations – possibly detected in NMR experiments [163, 78] but not significantly present in both the G45a4 and new parameter sets – has to be clarified.



(a) glucose (β-D-Glc)

(b) galactose (β-D-Gal)

(c) mannose (β-D-Man)

(d) allose (β-D-All)

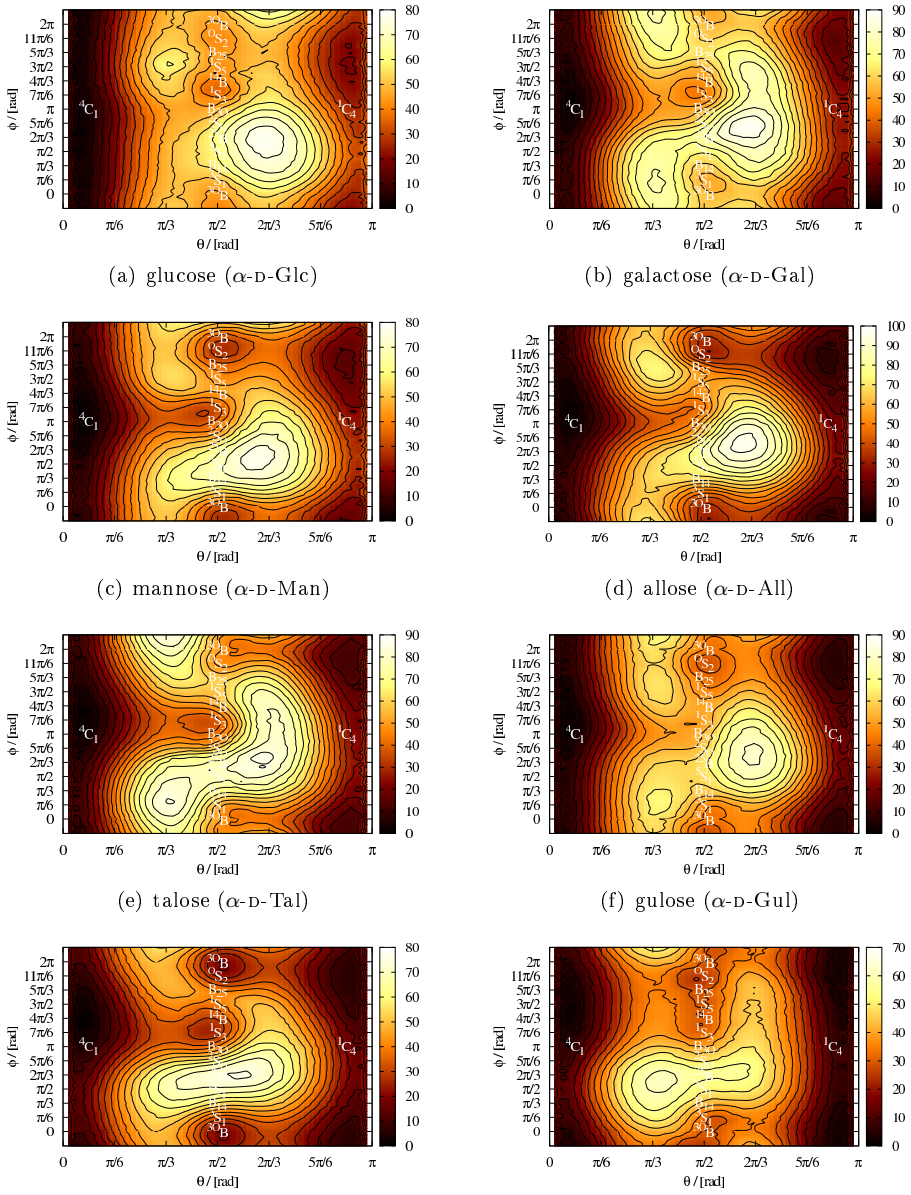
(e) talose (β-D-Tal)

(f) gulose (β-D-Gul)

(g) altrose (β-D-Alt)

(h) idose (β-D-Ido)

Figure 6.9: Free energy landscape of β-D-aldopyranoses. Puckering free energy $F_{\mathrm{M+US}}(\theta, \phi)$ (see Eq. (4.19)) of β-D-aldopyranoses using the 45a4-ASPG parameter set are presented. The profile is set to zero at the position of the local minimum in the $^4C_1$ basin ($\theta < \pi/3$). Plate Carrèe projections, with isolines drawn every $2k_BT$ ($T = 300\,\mathrm{K}$). Profile are replicated in two thin stripes at $\phi < 0$ and $\phi > 2\pi$ to stress the ϕ periodicity. Darker colors correspond to lower energies.



(a) glucose (α-D-Glc)

(b) galactose (α-D-Gal)

(c) mannose (α-D-Man)

(d) allose (α-D-All)

(e) talose (α-D-Tal)

(f) gulose (α-D-Gul)

(g) altrose (α-D-Alt)

(h) idose (α-D-Ido)

Figure 6.10: Free energy landscape of α-D-aldopyranoses. Puckering free energy $F_{\mathrm{M+US}}(\theta, \phi)$ (see Eq. (4.19)) of α-D-aldopyranoses using the 45a4-ASPG parameter set are presented. The profile is set to zero at the position of the local minimum in the ${}^4C_1$ basin ($\theta < \pi/3$). Plate Carrèe projections, with isolines drawn every $2k_BT$ ($T = 300\,\mathrm{K}$). Profile are replicated in two thin stripes at $\phi < 0$ and $\phi > 2\pi$ to stress the ϕ periodicity. Darker colors correspond to lower energies.

Chapter 7

Conclusions and further work

> What I need now is an honest answer
> To make things better
> You can see now
> My hands are tied and I surrender
> So I'll wait here for your final answer
> Yeah, your final answer
>
> The Calling
> *Final Answer*

In the present work we performed a systematic study of the conformational properties of hexopyranose rings by using different variants of the metadynamics algorithm (direct metadynamics, well-tempered metadynamics, metadynamics with umbrella sampling corrections) and different puckering coordinates (spherical and Cartesian Cremer-Pople representations, and Strauss-Pickett dihedrals). We investigated the suitability of several different coordinate sets as collective variables for metadynamics, and we addressed the reliability of the GROMOS 45a4 force field, providing moreover a new parametrization of this force field which leads to much more realistic ratios of the chair and inverted chair conformers.

Using both toy-models and regular models of six membered ring molecules, we showed clearly that the puckering phase space is, for any realistic structure, a spherical shell with no metastabilities in the radial direction of Cremer-Pople spherical coordinates. This feature of the puckering free energy landscape implies that any restricted coordinate set to be used as collective variable for metadynamics should always be able to span the surface of the puckering sphere. To this end, the Cremer-Pople framework permits this exploration, taking advantage of its spherical representation, with the natural (θ, ϕ) angles or with every equivalent spherical pair, like the alternative (γ, η) and $(\mu.\nu)$ angles proposed in Section 4.3.2.

We have taken into account two puckering coordinate sets other than the spherical Cremer-Pople ones: the Cartesian reduced Cremer-Pople (q_x, q_y) and the Strauss-Pickett angles $(\alpha_1, \alpha_2, \alpha_3)$. We investigated their respective advantages and drawbacks as collective variables for the conformational analysis of pyranoses:

(i) the reduced Cartesian set of Cremer-Pople coordinates can not be used as proper CVs to explore the puckering conformational space, even if the analysis is restricted to one of the two emispheres of the puckering sphere. We showed how metadynamics performed using reduced Cartesian CVs – contrarily to the spherical case – is non-ergodic, and how the reconstructed free energy profile suffers from strong biases. The conformations sampled close to the equatorial line were characterized by unphysical values of the total puckering amplitude, which has been shown to be markedly correlated with the proximity to the equatorial line.

(ii) for the Strauss-Pickett angles the geometrical equivalence with the spherical Cremer-Pople ones does not correspond to an equivalence on the practical level in metadynamics simulations. We showed that the need of three variables for the Strauss-Pickett scheme to reach an ergodic sampling in puckering free energy reconstructions is the main drawback with respect to the two variable scheme of the spherical Cremer-Pople approach. This happens for two main reasons: the first problem is that the algorithmic partitioning schemes for three-dimensional free energy landscapes have proven to be prone to a misleading interpretation of the thermally accessible conformers of sugar rings (by means of a miscounting of states in thermodynamic basins); secondly, employing Cremer-Pople coordinates rather than Strauss-Pickett dihedrals leads to a tenfold decrease of the time needed to homogeneously fill the conformational space, for typical systems.

Within the spherical Cremer-Pople framework, we studied in a systematic way the whole series of D-aldopyranoses. Our findings pointed out the deficiencies of the GROMOS 45a4 force field concerning the description of puckering properties: the chair/inverted chair free energy differences are in clear disagreement with theoretical prediction and, even worse, with the few experimental data at disposal. We proposed a transferable, not sugar-specific re-parametrization of the GROMOS 45a4 parameter set, that improves its accuracy in describing the chair/inverted chair free energy difference. The improvement obtained by using the refined set of parameters (referred to as 45a4-ASPG) is substantial: the free energy difference of the inverted chair against stereoisomery is now in agreement with theoretical predictions and experimental data, while other structural and dynamical properties are retained.

The general results of this work make the simulation protocol developed in it a grounded application of the metadynamics technique for puckering conformational analysis. Further refinements of the simulation protocol are yet possible, for example the aforementioned combined use of different spherical sets in free energy reconstructions (see Section 4.3.2). Indeed, some preliminary results in this direction have been recently done: the values of chair/inverted chair free energy difference for α-D-Alt and α-D-Ido, with the same simulation condition described in Chapter 6, were found to be 3.6(2) kJ/mol and 0.6(4) kJ/mol, respectively, in agreement within the errors with the values indicated in Table 6.3. Taking into consideration that this improvement is intended to help against the under-sampling of polar regions of the puckering coordinates, the agreement of these preliminary results with the already presented results stress once more the robustness of our approach.

By and large, the proposed application of metadynamics is in our opinion a valuable approach to test the performances of force fields specifically designed for carbohydrates, like the recently proposed versions of the CHARMM [61] and the GLYCAM [91] force fields for mono- and oligo-saccharides. Indeed, for GROMOS force field, as for many others, the investigations of ring conformer populations in water are scarce, if not missing at all. Moreover, to the best of our knowledge no attempts in testing the ability of force fields in describing puckering properties in solvents different from water has been addressed. The use of other solvents (like methanol CH_3OH, or dimethilsulfoxide $(CH_3)_2SO$) lead to different conformational equilibrium in solution (as can be seen experimentally, see Ref. [10]) by means of different intermolecular interactions that influence the stability of ring conformers. With the help of the two possible benchmark systems, altrose and idose, this could be an extremely valuable testing field for force field capabilities.

The extension of the proposed scheme to study furanoses (five-membered ring sugars) is also quite straightforward. The importance of this study is easly understandable, considering for example the role of furanoid sugars as backbone elements in molecules like DNA and RNA. Our hope is that besides the usefulness related to the specific case of the GROMOS force field paramterization, this work could also serve to attract attention on the importance of the puckering problem in carbohydrate simulations, and to stimulate further investigations.

Appendix A

Cremer-Pople coordinates – Definitions and gradients

We present here some details of the calcualtions regarding the Cremer-Pople puckering parameters.

Position vectors

Consider the position vectors $\mathbf{r}_j$ of the atoms $j = 1, \ldots, N$ of a closed ring. The vectors $\mathbf{R}_j$ with respect to the geometric center $\mathbf{R}$ of the ring are

$$\begin{aligned}\mathbf{R}_j = \mathbf{r}_j - \mathbf{R} = \mathbf{r}_j - \frac{1}{N}\sum_{i=1}^{N}\mathbf{r}_i = \frac{1}{N}N\mathbf{r}_j - \frac{1}{N}\sum_{i=1}^{N}\mathbf{r}_i = \\ = \frac{1}{N}\sum_{i=1}^{N}N\delta_{ij}\mathbf{r}_i - \frac{1}{N}\sum_{i=1}^{N}\mathbf{r}_i = \sum_{i=1}^{N}\frac{N\delta_{ij}-1}{N}\mathbf{r}_i \equiv \sum_{i=1}^{N}\Delta_{ij}\mathbf{r}_i \quad \text{(A.1)}\end{aligned}$$

where the following conditions (used in Eqs. (3.7) and (3.8) for the definition of the mean plane of the ring) applies

$$\sum_{j=1}^{N}\Delta_{ij} = 0 \quad \forall\, i = 1, \ldots, N \quad \Leftrightarrow \quad \sum_{j=1}^{N}\mathbf{R}_j = 0 \quad . \qquad \text{(A.2)}$$

Range of m

All the Cremer-Pople definitions of Section 3.2 are defined using the weights of

$$\begin{cases} w_{m,j} = \sin(m\alpha_j) \\ v_{m,j} = \cos(m\alpha_j) \end{cases} \quad , \quad \alpha_j = \frac{2\pi(j-1)}{N} \quad , \qquad \text{(A.3)}$$

where $j = 1, \ldots, N$ is the atom label and the values of m could span in principle values from 0 to $N-1$. Indeed, after $m > N$ the weights are redundant by construction, for we can always write $m = m' + kN$ with $k \in \mathbb{N}$ and $0 < m' < N$,

and this gives

$$m\alpha_j = m\frac{2\pi(j-1)}{N} = m'\frac{2\pi(j-1)}{N} + k\cancel{N}\frac{2\pi(j-1)}{\cancel{N}} \tag{A.4}$$

from which we have $(m' + kN)\alpha_j \to m'\alpha_j$ in the goniometric functions of Eq. (A.5).

However, if we use the whole $m = 0, \ldots, N-1$ values, half of them are redundant and, accordingly, in Section 3.2.2 we stated that only

$$m \in I_N \quad , \quad I_N = \begin{cases} \{2, 3, \ldots, \frac{N-1}{2}\} & N \text{ odd} \\ \{2, 3, \ldots, \frac{N}{2} - 1\} & N \text{ even} \end{cases} \tag{A.5}$$

are sufficient (in Table A.1 for the list of their symbolic values is given). Indeed,

Table A.1: Weights for Cremer-Pople definitions. List of weights used in CP definitions in symbolic compact form (the actual values are given by substituting proper m and j values in Eq. (A.5)).

						N odd	N even
$m \in I_N$	0	1	2	3	...	$(N-1)/2$	$N/2-1$
$v_{m,j}$	1	$v_{1,j}$	$v_{2,j}$	$v_{3,j}$	...	$v_{(N-1)/2,j}$	$v_{N/2-1,j}$
$w_{m,j}$	0	$w_{1,j}$	$w_{2,j}$	$w_{3,j}$	...	$w_{(N-1)/2,j}$	$w_{N/2-1,j}$

skipping the $m = N/2$ term of the N even case (which gives the $q_{N/2}$ coordinate), we can in principle procede with greater m values until $m = N-1$. However, we can index these extra term with $m^* = N - m$, $m \in I_N$, obtaining the weights reported in Table A.2. The equivalences hold by means of goniometric identities for $\cos(m^*\alpha_j)$ and $\sin(m^*\alpha_j)$ towards the corresponding $\cos(m\alpha_j)$ and $\sin(m\alpha_j)$ terms.

Table A.2: Redundant weights. List of weights obtained for m^* values in symbolic compact form (the equivalence are with the symbols of Table A.1).

	N odd	N even				
$m \in I_N$	$(N-1)/2$	$N/2-1$	...	3	2	1
$m^* = N - m$	$(N+1)/2$	$N/2+1$	...	$N-3$	$N-2$	$N-1$
$v_{m^*,j}$	$v_{(N-1)/2,j}$	$v_{N/2-1,j}$	...	$v_{3,j}$	$v_{2,j}$	$v_{1,j}$
$w_{m^*,j}$	$-w_{(N-1)/2,j}$	$-w_{N/2-1,j}$	...	$-w_{3,j}$	$-w_{2,j}$	$-w_{1,j}$

All the $v_{m^*,j}$ values are then redundant, and the only relevant changes are the signs of $w_{m^*,j} = -w_{m,j}$. This difference, however, does not create new indipendent Cremer-Pople coordinates: the possible new coordinate pairs are

$$q^*_m \cos\phi^*_m = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} z_j v_{m^*,j} = q_m \cos\phi_m \tag{A.6a}$$

$$q^*_m \sin\phi^*_m = -\sqrt{\frac{2}{N}} \sum_{j=1}^{N} z_j w_{m^*,j} = -q_m \sin\phi_m \tag{A.6b}$$

and thus $q_{m^*} = q_m$ and $\phi_{m^*} = -\phi_m$ by construction. For this reason, the ranges Eq. (3.19) are sufficient for Cremer-Pople definitions (Eqs. (3.16) to (3.17)).

Gradients for Cremer-Pople parameters ($N = 6$)

For the total puckering amplitude $Q = \sqrt{\sum_{j=1}^{6} z_j^2}$ we have

$$\nabla_i Q \;=\; \nabla_i \sqrt{\sum_{j=1}^{6} z_j^2} \;=\; \frac{1}{2Q} \nabla_i \sum_{j=1}^{6} z_j^2 \;=\; \frac{1}{2Q} \sum_{j=1}^{6} \nabla_i z_j^2 \;=\; \frac{1}{Q} \sum_{j=1}^{6} z_j \nabla_i z_j \quad \text{(A.7)}$$

(this formula is nevertheless valid for generic N values). For the spherical angles, we recall their definitions $\theta = \arctan[q_2/q_3] = \arctan[\sqrt{2(\mathcal{A}_2^2 + \mathcal{B}_2^2)}/\mathcal{C}]$ and $\phi = \arctan[-\mathcal{A}_2/\mathcal{B}_2]$ as functions of the re-summations

$$\mathcal{A}_2 = \sum_{j=1}^{6} z_j w_{2,j} \quad , \quad \mathcal{B}_2 = \sum_{j=1}^{6} z_j v_{2,j} \quad , \quad \mathcal{C} = \sum_{j=1}^{6} z_j (-1)^{j-1} \qquad \text{(A.8)}$$

(that are the one of Eq. (3.20) with $N = 6$ and with the weights Eq. (3.9) with $m = 2$). We will make use of

$$\nabla_i \tan\alpha = [1 + (\tan\alpha)^2] \nabla_i \alpha \longrightarrow \nabla_i \alpha = \frac{1}{1 + (\tan\alpha)^2} \nabla_i \tan\alpha \qquad \text{(A.9)}$$

to calculate the gradients. For θ

$$\nabla_i \theta = \frac{1}{1 + \left(\dfrac{q_2}{q_3}\right)^2} \nabla_i \left(\frac{q_2}{q_3}\right) = \frac{q_3^2}{q_2^2 + q_3^2} \left(q_2 \nabla_i \frac{1}{q_3} + \frac{1}{q_3} \nabla_i q_2 \right) =$$

$$= \frac{1}{Q^2} \left(-q_2 \nabla_i q_3 + q_3 \nabla_i q_2 \right) = (\star) \quad \text{(A.10)}$$

and using

$$\nabla_i q_2 = \nabla_i \sqrt{(\mathcal{A}_2^2 + \mathcal{B}_2^2)/3} = \frac{1}{3q_2} \left[\mathcal{A}_2 (\nabla_i \mathcal{A}_2) + \mathcal{B}_2 (\nabla_i \mathcal{B}_2) \right] \quad , \quad \nabla_i q_3 = \frac{\nabla_i \mathcal{C}}{\sqrt{6}}$$

$$\frac{q_3}{3q_2} = \frac{\mathcal{C}}{3\sqrt{2}\sqrt{\mathcal{A}_2^2 + \mathcal{B}_2^2}}$$

we obtain

$$(\star) = \frac{1}{3\sqrt{2}Q^2} \left\{ \frac{\mathcal{C}}{\sqrt{\mathcal{A}_2^2 + \mathcal{B}_2^2}} \left[\mathcal{A}_2 (\nabla_i \mathcal{A}_2) + \mathcal{B}_2 (\nabla_i \mathcal{B}_2) \right] - \sqrt{\mathcal{A}_2^2 + \mathcal{B}_2^2} \nabla_i \mathcal{C} \right\} \quad \text{(A.11)}$$

For ϕ

$$\nabla_i\phi = -\frac{1}{1+\left(\frac{\mathcal{A}_2}{\mathcal{B}_2}\right)^2}\nabla_i\left(\frac{\mathcal{A}_2}{\mathcal{B}_2}\right) = \frac{-\mathcal{B}_2^2}{\mathcal{B}_2^2+\mathcal{A}_2^2}\left[\nabla_i\left(\frac{1}{\mathcal{B}_2}\right)\mathcal{A}_2 + \frac{1}{\mathcal{B}_2}(\nabla_i\mathcal{A}_2)\right] =$$
$$= \frac{-\mathcal{B}_2^2}{\mathcal{B}_2^2+\mathcal{A}_2^2}\left[-\frac{1}{\mathcal{B}_2^2}(\nabla_i\mathcal{B}_2)\mathcal{A}_2 + \frac{1}{\mathcal{B}_2}(\nabla_i\mathcal{A}_2)\right] = \frac{1}{\mathcal{A}_2^2+\mathcal{B}_2^2}\left[\mathcal{A}_2(\nabla_i\mathcal{B}_2) - \mathcal{B}_2(\nabla_i\mathcal{A}_2)\right] \tag{A.12}$$

We report here for completeness the gradients of Cremer-Pople coordinates for $N = 6$, namely the gradients of Cartesian Cremer-Pople coordinates

$$\nabla_i q_x = \nabla_i\mathcal{B}_2 \tag{A.13a}$$
$$\nabla_i q_y = -\nabla_i\mathcal{A}_2 \tag{A.13b}$$
$$\nabla_i q_z = \nabla_i\mathcal{C} \tag{A.13c}$$

and the gradients of cylindrical Cremer-Pople coordinates

$$\nabla_i q_2 = \frac{\mathcal{A}_2\nabla_i\mathcal{A}_2 + \mathcal{B}_2\nabla_i\mathcal{B}_2}{\sqrt{3(\mathcal{A}_2^2+\mathcal{B}_2^2)}} \tag{A.13d}$$
$$\nabla_i\phi_2 = \nabla_i\phi \tag{A.13e}$$
$$\nabla_i q_3 = \nabla_i q_z \quad . \tag{A.13f}$$

All the reported gradients are functions of $\nabla_i\mathcal{A}_2$, $\nabla_i\mathcal{B}_2$ and $\nabla_i\mathcal{C}$, then are all eventually functions of $\nabla_i z_j$ (see Eq. (3.32)).

Gradient of the elevations z_j

The following formulae are valid for generic $N = 6$ values. From the definition

$$z_j = \mathbf{R}_j \cdot \frac{\mathbf{R}'\times\mathbf{R}''}{|\mathbf{R}'\times\mathbf{R}''|} \quad , \quad \begin{cases}\mathbf{R}' = \sum_{j=1}^N \mathbf{R}_j w_{1,j} \\ \mathbf{R}'' = \sum_{j=1}^N \mathbf{R}_j v_{1,j}\end{cases} \tag{A.14}$$

by direct calculation we have[1]

$$\begin{aligned}\nabla_i z_j &= \nabla_i\left(\mathbf{R}_j\cdot\frac{\mathbf{R}'\times\mathbf{R}''}{|\mathbf{R}'\times\mathbf{R}''|}\right) \\ &= \nabla_i\left(\frac{1}{|\mathbf{R}'\times\mathbf{R}''|}\right)\left[\mathbf{R}_j\cdot\left(\mathbf{R}'\times\mathbf{R}''\right)\right] + \frac{\nabla_i\left[\mathbf{R}_j\cdot\left(\mathbf{R}'\times\mathbf{R}''\right)\right]}{|\mathbf{R}'\times\mathbf{R}''|} \\ &= -\frac{\left[\mathbf{R}_j\cdot\left(\mathbf{R}'\times\mathbf{R}''\right)\right]}{|\mathbf{R}'\times\mathbf{R}''|^2}\nabla_i|\mathbf{R}'\times\mathbf{R}''| + \frac{\nabla_i\left[\mathbf{R}_j\cdot\left(\mathbf{R}'\times\mathbf{R}''\right)\right]}{|\mathbf{R}'\times\mathbf{R}''|} \\ &= -\frac{\left[\mathbf{R}_j\cdot\left(\mathbf{R}'\times\mathbf{R}''\right)\right]}{2|\mathbf{R}'\times\mathbf{R}''|^3}\nabla_i\left[(\mathbf{R}'\times\mathbf{R}'')\cdot(\mathbf{R}'\times\mathbf{R}'')\right] + \frac{\nabla_i\left[\mathbf{R}_j\cdot\left(\mathbf{R}'\times\mathbf{R}''\right)\right]}{|\mathbf{R}'\times\mathbf{R}''|}\end{aligned} \tag{A.15}$$

and thus we obtain Eq. (3.33).

[1] using $|\mathbf{a}| = \sqrt{\mathbf{a}\cdot\mathbf{a}}$

In the following calculation of $\nabla_i z_j$, roman letters $(i,j,k,l,m,n = 1,\dots,N)$ are used for particle indexes, while Greek letters $(\alpha,\beta,\gamma,\delta,\eta,\theta = 1,2,3)$ are used as vector component indexes. To lighten the notation, the contracted symbols $w_j \equiv w_{1,j}$ and $v_j \equiv v_{1,j}$ are used, and every repeated index is intended to be summed (Einstein summation convention). Given the position vectors $\mathbf{R}_j = \hat{\mathrm{e}}_\alpha R_j^\alpha$ (with respect to the original reference frame $\{\hat{\mathrm{e}}_\alpha\}_{\alpha=1,2,3}$) the explicit vector product[2]

$$\mathbf{R}' \times \mathbf{R}'' = w_l v_k \mathbf{R}_l \times \mathbf{R}_k = w_l v_k \hat{\mathrm{e}}_\alpha \varepsilon^{\alpha\beta\gamma} R_l^\beta R_k^\gamma \tag{A.16}$$

($\varepsilon^{\alpha\beta\gamma}$ is the totally-antisymmetric Levi-Civita symbol) and its modulus[3]

$$|\mathbf{R}' \times \mathbf{R}''|^2 = (\mathbf{R}' \times \mathbf{R}'') \cdot (\mathbf{R}' \times \mathbf{R}'') = w_l v_k w_a v_b R_l^\beta R_k^\gamma (R_a^\beta R_b^\gamma - R_a^\gamma R_b^\beta) \tag{A.17}$$

and the triple product[4]

$$\mathbf{R}_j \cdot \big(\mathbf{R}' \times \mathbf{R}''\big) = w_l v_k \mathbf{R}_j \cdot \big(\mathbf{R}_l \times \mathbf{R}_k\big) = w_l v_k \varepsilon^{\alpha\beta\gamma} R_j^\alpha R_l^\beta R_k^\gamma \tag{A.18}$$

can be computed. In the terms $\nabla_i z_j$ the gradient ∇_i acts only on the coordinates R_j^α:

$$\nabla_i R_j^\alpha = \hat{\mathrm{e}}_\beta \frac{\partial R_j^\alpha}{\partial r_i^\beta} = \hat{\mathrm{e}}_\beta \frac{\partial}{\partial r_i^\beta} \Delta_{kj} r_k^\alpha = \hat{\mathrm{e}}_\beta \Delta_{kj} \frac{\partial r_k^\alpha}{\partial r_i^\beta} = \hat{\mathrm{e}}_\beta \Delta_{kj} \delta^{\beta\alpha} \delta^{ik} = \hat{\mathrm{e}}_\alpha \Delta_{ij} \tag{A.19}$$

(with different indexes according to the specific term).

The calculation of

$$\nabla_i \big[\mathbf{R}_j \cdot \big(\mathbf{R}' \times \mathbf{R}''\big)\big] = w_l v_k \nabla_i \big[\mathbf{R}_j \cdot (\mathbf{R}_l \times \mathbf{R}_k)\big] \tag{A.20}$$

is a resummation of the terms

$$\begin{aligned}
\nabla_i \big[\mathbf{R}_j \cdot (\mathbf{R}_l \times \mathbf{R}_k)\big] &= \nabla_i \left(R_j^\alpha \varepsilon^{\alpha\beta\gamma} R_l^\beta R_k^\gamma\right) = \varepsilon^{\alpha\beta\gamma} \nabla_i \left(R_j^\alpha R_l^\beta R_k^\gamma\right) = \\
&= \varepsilon^{\alpha\beta\gamma} \Big[\big(\nabla_i R_j^\alpha\big) R_l^\beta R_k^\gamma + R_j^\alpha \big(\nabla_i R_l^\beta\big) R_k^\gamma + R_j^\alpha R_l^\beta \big(\nabla_i R_k^\gamma\big)\Big] = \\
&= \varepsilon^{\alpha\beta\gamma} \Big[\Delta_{ij} \hat{\mathrm{e}}_\alpha R_l^\beta R_k^\gamma + R_j^\alpha \Delta_{il} \hat{\mathrm{e}}_\beta R_k^\gamma + R_j^\alpha R_l^\beta \Delta_{ik} \hat{\mathrm{e}}_\gamma\Big] = \\
&= \Delta_{ij} \hat{\mathrm{e}}_\alpha \varepsilon^{\alpha\beta\gamma} R_l^\beta R_k^\gamma + \Delta_{il} \hat{\mathrm{e}}_\beta \varepsilon^{\alpha\beta\gamma} R_j^\alpha R_k^\gamma + \Delta_{ik} \hat{\mathrm{e}}_\gamma \varepsilon^{\alpha\beta\gamma} R_j^\alpha R_l^\beta \\
&= \Delta_{ij} \big(\mathbf{R}_l \times \mathbf{R}_k\big) + \Delta_{il} \big(\mathbf{R}_k \times \mathbf{R}_j\big) + \Delta_{ik} \big(\mathbf{R}_j \times \mathbf{R}_l\big)
\end{aligned}$$

where in the last equality we recombined the summations to three vector products (*e.g.* $\hat{\mathrm{e}}_\alpha \varepsilon^{\alpha\beta\gamma} R_l^\beta R_k^\gamma = \mathbf{R}_l \times \mathbf{R}_k$). Then, the total gradient is

$$\begin{aligned}
&\nabla_i \big[\mathbf{R}_j \cdot \big(\mathbf{R}' \times \mathbf{R}''\big)\big] = \\
&\quad = w_l v_k \Big[\Delta_{ij} \big(\mathbf{R}_l \times \mathbf{R}_k\big) + \Delta_{il} \big(\mathbf{R}_k \times \mathbf{R}_j\big) + \Delta_{ik} \big(\mathbf{R}_j \times \mathbf{R}_l\big)\Big] = \\
&\quad = \Delta_{ij} \left(w_l v_k \mathbf{R}_l \times \mathbf{R}_k\right) + \left(\Delta_{il} w_l\right) \left(v_k \mathbf{R}_k \times \mathbf{R}_j\right) + \left(\Delta_{ik} v_k\right) \left(\mathbf{R}_j \times w_l \mathbf{R}_l\right)
\end{aligned}$$

[2] using $\mathbf{a} \times \mathbf{b} = \hat{\mathrm{e}}_\alpha \varepsilon^{\alpha\beta\gamma} a_\beta b_\gamma$

[3] using $(\mathbf{a} \times \mathbf{b}) \cdot (\mathbf{c} \times \mathbf{d}) = \varepsilon^{\alpha\beta\gamma} \varepsilon^{\alpha\eta\theta} a_\beta b_\gamma c_\eta d_\theta = a_\beta b_\gamma (c_\beta d_\gamma - c_\gamma d_\beta)$ in which we used $\varepsilon^{\alpha\beta\gamma} \varepsilon^{\alpha\eta\theta} = \delta_{\beta\eta}\delta_{\gamma\theta} - \delta_{\beta\theta}\delta_{\gamma\eta}$

[4] using $\mathbf{c} \cdot (\mathbf{a} \times \mathbf{b})) = c_\alpha \varepsilon^{\alpha\beta\gamma} a_\beta b_\gamma$

and with the definitions

$$\mathcal{E}_i = \Delta_{il} w_l \quad , \quad \mathcal{F}_i = \Delta_{ik} v_k \tag{A.21}$$

we eventually obtain Eq. (3.34b).

The calculation of

$$\nabla_i \big[(\mathbf{R}' \times \mathbf{R}'') \cdot (\mathbf{R}' \times \mathbf{R}'') \big] = w_l v_k w_a v_b \nabla_i \big[R_l^\beta R_k^\gamma (R_a^\beta R_b^\gamma - R_a^\gamma R_b^\beta) \big] \tag{A.22}$$

is a resummation of the terms

$$\begin{aligned}
&\nabla_i \big[R_l^\beta R_k^\gamma (R_a^\beta R_b^\gamma - R_a^\gamma R_b^\beta) \big] = \\
&\quad = \Big[\big(\nabla_i R_l^\beta\big) R_k^\gamma (R_a^\beta R_b^\gamma - R_a^\gamma R_b^\beta) + R_l^\beta \big(\nabla_i R_k^\gamma\big)(R_a^\beta R_b^\gamma - R_a^\gamma R_b^\beta) + \\
&\qquad R_l^\beta R_k^\gamma \big[\big(\nabla_i R_a^\beta\big) R_b^\gamma + R_a^\beta \big(\nabla_i R_b^\gamma\big) - \big(\nabla_i R_a^\gamma\big) R_b^\beta - R_a^\gamma \big(\nabla_i R_b^\beta\big) \big] \Big] = \\
&\quad = \Big[\hat{\mathrm{e}}_\beta \Delta_{il} R_k^\gamma (R_a^\beta R_b^\gamma - R_a^\gamma R_b^\beta) + \hat{\mathrm{e}}_\gamma \Delta_{ik} R_l^\beta (R_a^\beta R_b^\gamma - R_a^\gamma R_b^\beta) + \\
&\qquad R_l^\beta R_k^\gamma \big[\hat{\mathrm{e}}_\beta \Delta_{ia} R_b^\gamma + \hat{\mathrm{e}}_\gamma \Delta_{ib} R_a^\beta - \hat{\mathrm{e}}_\gamma \Delta_{ia} R_b^\beta - \hat{\mathrm{e}}_\beta \Delta_{ib} R_a^\gamma \big] \Big] = \\
&\quad = \Delta_{il} \left[\hat{\mathrm{e}}_\beta R_a^\beta R_k^\gamma R_b^\gamma - \hat{\mathrm{e}}_\beta R_b^\beta R_k^\gamma R_a^\gamma \right] + \Delta_{ik} \left[\hat{\mathrm{e}}_\gamma R_b^\gamma R_l^\beta R_a^\beta - \hat{\mathrm{e}}_\gamma R_a^\gamma R_l^\beta R_b^\beta \right] + \\
&\qquad + \Delta_{ia} \left[\hat{\mathrm{e}}_\beta R_l^\beta R_k^\gamma R_b^\gamma - \hat{\mathrm{e}}_\gamma R_k^\gamma R_l^\beta R_b^\beta \right] + \Delta_{ib} \left[\hat{\mathrm{e}}_\gamma R_k^\gamma R_l^\beta R_a^\beta - \hat{\mathrm{e}}_\beta R_l^\beta R_k^\gamma R_a^\gamma \right] \quad .
\end{aligned}$$

By re-combining the terms with the correct indexes, the summation transforms to scalar products[5] (*e.g.* $R_k^\gamma R_b^\gamma = \mathbf{R}_k \cdot \mathbf{R}_b$) and to vectors (*e.g.* $\hat{\mathrm{e}}_\beta R_a^\beta = \mathbf{R}_a$) giving

$$\begin{aligned}
&\nabla_i \big[(\mathbf{R}' \times \mathbf{R}'') \cdot (\mathbf{R}' \times \mathbf{R}'') \big] = \\
&\qquad = w_l v_k w_a v_b \Big\{ \Delta_{il} \big[\mathbf{R}_a \big(\mathbf{R}_k \cdot \mathbf{R}_b\big) - \mathbf{R}_b \big(\mathbf{R}_k \cdot \mathbf{R}_a\big) \big] + \\
&\qquad\qquad + \Delta_{ik} \big[\mathbf{R}_b \big(\mathbf{R}_l \cdot \mathbf{R}_a\big) - \mathbf{R}_a \big(\mathbf{R}_l \cdot \mathbf{R}_b\big) \big] + \\
&\qquad\qquad + \Delta_{ia} \big[\mathbf{R}_l \big(\mathbf{R}_k \cdot \mathbf{R}_b\big) - \mathbf{R}_k \big(\mathbf{R}_l \cdot \mathbf{R}_b\big) \big] + \\
&\qquad\qquad + \Delta_{ib} \big[\mathbf{R}_k \big(\mathbf{R}_l \cdot \mathbf{R}_a\big) - \mathbf{R}_l \big(\mathbf{R}_k \cdot \mathbf{R}_a\big) \big] \Big\} = \\
&\qquad = w_l \Delta_{il} \big[w_a \mathbf{R}_a \big(v_k \mathbf{R}_k \cdot v_b \mathbf{R}_b\big) - v_b \mathbf{R}_b \big(v_k \mathbf{R}_k \cdot w_a \mathbf{R}_a\big) \big] + \\
&\qquad\quad + v_k \Delta_{ik} \big[v_b \mathbf{R}_b \big(w_l \mathbf{R}_l \cdot w_a \mathbf{R}_a\big) - w_a \mathbf{R}_a \big(w_l \mathbf{R}_l \cdot v_b \mathbf{R}_b\big) \big] + \\
&\qquad\quad + w_a \Delta_{ia} \big[w_l \mathbf{R}_l \big(v_k \mathbf{R}_k \cdot v_b \mathbf{R}_b\big) - v_k \mathbf{R}_k \big(w_l \mathbf{R}_l \cdot v_b \mathbf{R}_b\big) \big] + \\
&\qquad\quad + v_b \Delta_{ib} \big[v_k \mathbf{R}_k \big(w_l \mathbf{R}_l \cdot w_a \mathbf{R}_a\big) - w_l \mathbf{R}_l \big(v_k \mathbf{R}_k \cdot w_a \mathbf{R}_a\big) \big] \quad .
\end{aligned}$$

where in the last manipulation we can substitute the vectors $\mathbf{R}'$ e $\mathbf{R}''$ and the symbols $\mathcal{E}_i$ and $\mathcal{F}_i$ of Eq. (A.21), eventually obtaining Eq. (3.34a).

Gradient of extra spherical Cremer-Pople angles ($N = 6$)

For the alternative spherical angles, we recall their definitions

$$\begin{cases} \gamma = \arctan \left[\sqrt{2\mathcal{A}_2^2 + \mathcal{C}^2} / \sqrt{2} \mathcal{B}_2 \right] \\ \eta = \arctan \left[-\mathcal{C} / \sqrt{2} \mathcal{A}_2 \right] \end{cases} , \begin{cases} \mu = \arctan \left[-\sqrt{2\mathcal{B}_2^2 + \mathcal{C}^2} / \sqrt{2} \mathcal{A}_2 \right] \\ \nu = \arctan \left[\sqrt{2} \mathcal{B}_2 / \mathcal{C} \right] \end{cases} \tag{A.23}$$

[5] using $\mathbf{a} \cdot \mathbf{b} = a_\alpha b_\alpha$

here given as explicit functions of the re-summations

$$\mathcal{A}_2 = \sum_{j=1}^{6} z_j w_{2,j} \quad , \quad \mathcal{B}_2 = \sum_{j=1}^{6} z_j v_{2,j} \quad , \quad \mathcal{C} = \sum_{j=1}^{6} z_j (-1)^{j-1} \tag{A.24}$$

(that are the one of Eq. (3.20) with $N = 6$ and with the weights Eq. (3.9) with $m = 2$). We will make use of

$$\nabla_i \tan\alpha = [1 + (\tan\alpha)^2]\nabla_i\alpha \longrightarrow \nabla_i\alpha = \frac{1}{1 + (\tan\alpha)^2}\nabla_i \tan\alpha \tag{A.25}$$

to perform the direct calculatio of the gradients. For γ

$$\begin{aligned}
\nabla_i\gamma &= \frac{1}{1 + \dfrac{2\mathcal{A}_2^2 + \mathcal{C}^2}{2\mathcal{B}_2^2}} \nabla_i \left(\frac{\sqrt{2\mathcal{A}_2^2 + \mathcal{C}^2}}{\sqrt{2}\mathcal{B}_2} \right) = \\
&= \frac{2\mathcal{B}_2^2}{2(\mathcal{A}_2^2 + \mathcal{B}_2^2) + \mathcal{C}^2} \left[\frac{2\mathcal{A}_2\nabla_i\mathcal{A}_2 + \mathcal{C}\nabla_i\mathcal{C}}{\sqrt{2}\mathcal{B}_2\sqrt{2\mathcal{A}_2^2 + \mathcal{C}^2}} - \frac{\sqrt{2\mathcal{A}_2^2 + \mathcal{C}^2}}{\sqrt{2}\mathcal{B}_2^2}\nabla_i\mathcal{B}_2 \right] = \\
&= \frac{\sqrt{2}}{6Q^2\sqrt{2\mathcal{A}_2^2 + \mathcal{C}^2}} \left[\mathcal{B}_2 \left(2\mathcal{A}_2\nabla_i\mathcal{A}_2 + \mathcal{C}\nabla_i\mathcal{C} \right) - \left(2\mathcal{A}_2^2 + \mathcal{C}^2 \right) \nabla_i\mathcal{B}_2 \right] \quad .
\end{aligned} \tag{A.26}$$

For η

$$\begin{aligned}
\nabla_i\eta &= \frac{1}{1 + \dfrac{\mathcal{C}^2}{2\mathcal{A}_2^2}} \frac{-1}{\sqrt{2}} \nabla_i \left(\frac{\mathcal{C}}{\mathcal{A}_2} \right) = -\frac{2\mathcal{A}_2^2}{2\mathcal{A}_2^2 + \mathcal{C}^2} \frac{\mathcal{A}_2\nabla_i\mathcal{C} - \mathcal{C}\nabla_i\mathcal{A}_2}{\sqrt{2}\mathcal{A}_2^2} = \\
&= \sqrt{2}\frac{\mathcal{C}\nabla_i\mathcal{A}_2 - \mathcal{A}_2\nabla_i\mathcal{C}}{2\mathcal{A}_2^2 + \mathcal{C}^2} \quad .
\end{aligned} \tag{A.27}$$

For μ

$$\begin{aligned}
\nabla_i\mu &= \frac{1}{1 + \dfrac{2\mathcal{B}_2^2 + \mathcal{C}^2}{2\mathcal{A}_2^2}} \nabla_i \left(-\frac{\sqrt{2\mathcal{B}_2^2 + \mathcal{C}^2}}{\sqrt{2}\mathcal{A}_2} \right) = \\
&= -\frac{2\mathcal{A}_2^2}{2(\mathcal{A}_2^2 + \mathcal{B}_2^2) + \mathcal{C}^2} \left[\frac{2\mathcal{B}_2\nabla_i\mathcal{B}_2 + \mathcal{C}\nabla_i\mathcal{C}}{\sqrt{2}\mathcal{A}_2\sqrt{2\mathcal{B}_2^2 + \mathcal{C}^2}} - \frac{\sqrt{2\mathcal{B}_2^2 + \mathcal{C}^2}}{\sqrt{2}\mathcal{A}_2^2}\nabla_i\mathcal{A}_2 \right] = \\
&= -\frac{\sqrt{2}}{6Q^2\sqrt{2\mathcal{B}_2^2 + \mathcal{C}^2}} \left[\mathcal{A}_2 \left(2\mathcal{B}_2\nabla_i\mathcal{B}_2 + \mathcal{C}\nabla_i\mathcal{C} \right) - \left(2\mathcal{B}_2^2 + \mathcal{C}^2 \right) \nabla_i\mathcal{A}_2 \right] \quad .
\end{aligned} \tag{A.28}$$

For ν

$$\begin{aligned}
\nabla_i\nu &= \frac{1}{1 + \dfrac{2\mathcal{B}_2^2}{\mathcal{C}^2}} \sqrt{2}\nabla_i \left(\frac{\mathcal{B}_2}{\mathcal{C}} \right) = \sqrt{2}\frac{\mathcal{C}^2}{2\mathcal{B}_2^2 + \mathcal{C}^2} \frac{\mathcal{C}\nabla_i\mathcal{B}_2 - \mathcal{B}_2\nabla_i\mathcal{C}}{\mathcal{C}^2} = \\
&= \sqrt{2}\frac{\mathcal{C}\nabla_i\mathcal{B}_2 - \mathcal{B}_2\nabla_i\mathcal{C}}{2\mathcal{B}_2^2 + \mathcal{C}^2} \quad .
\end{aligned} \tag{A.29}$$

It is interesting to observe that the reported gradients, together with the corresponding formulae Eqs. (3.30) and (3.31) for the (θ, ϕ) set, have the same pattern of terms. In fact, the gradients could be calculated with the following general approach. Given the generic coordinates of $\mathbb{R}^3$ in spherical and Cartesian representations

$$\begin{cases} x = \varrho \sin\vartheta \cos\varphi \\ y = \varrho \sin\vartheta \sin\varphi \\ z = \varrho \cos\vartheta \end{cases} \quad \Leftrightarrow \quad \begin{cases} \varrho = \sqrt{x^2+y^2+z^2} \\ \vartheta = \arccos\left[z/\sqrt{x^2+y^2+z^2}\right] \\ \varphi = \arctan\left[y/x\right] \end{cases} \tag{A.30}$$

and considering all the components (x, y, z) and $(\varrho, \vartheta, \varphi)$ as functions of some underlying coordinates $\mathbf{r}_i = \hat{\mathrm{e}}_\alpha r_i^\alpha$ (in our case, the atomics positions in the original reference frame), the gradients ∇_i with respect to $\mathbf{r}_i$ of the spherical angles reads

$$\nabla_i \vartheta = \frac{z\left[x\nabla_i x + y\nabla_i y\right] - (x^2+y^2)\nabla_i z}{\varrho^2\sqrt{x^2+y^2}} \tag{A.31a}$$

$$\nabla_i \varphi = \frac{x\nabla_i y - y\nabla_i x}{x^2+y^2} \quad . \tag{A.31b}$$

As can be noticed, the common structure (number of terms and their pattern in the formula) of the gradients of θ, γ, μ and ϕ, η, μ, respectively, is again visible. Indeed, the connection with previously calculated gradients is made with the substitutions

$$(\theta,\phi): \begin{cases} x \leftrightarrow q_x = \mathcal{B}_2/\sqrt{3} \\ y \leftrightarrow q_y = -\mathcal{A}_2/\sqrt{3} \\ z \leftrightarrow q_z = \mathcal{C}/\sqrt{6} \end{cases} \quad , \quad (\gamma,\eta): \begin{cases} x \leftrightarrow q_y = -\mathcal{A}_2/\sqrt{3} \\ y \leftrightarrow q_z = \mathcal{C}/\sqrt{6} \\ z \leftrightarrow q_x = \mathcal{B}_2/\sqrt{3} \end{cases}$$

$$(\mu,\nu): \begin{cases} x \leftrightarrow q_z = \mathcal{C}/\sqrt{6} \\ y \leftrightarrow q_x = \mathcal{B}_2/\sqrt{3} \\ z \leftrightarrow q_y = -\mathcal{A}_2/\sqrt{3} \end{cases} \quad . \tag{A.32}$$

Moreover, with this general formula it is possible to calculate the analytical gradients of a generic pair of spherical CP angles (ϑ, φ), provided the proper transformation from the starting Cartesian CP coordinates (q_x, q_y, q_z) to any rotated Cartesian CP framework (q'_x, q'_y, q'_z) of interest. In this way for example will be possible to select the position of the puckering axes $\{\hat{\mathrm{q}}'_x, \hat{\mathrm{q}}'_y, \hat{\mathrm{q}}'_z\}$ in directions such that the polar regions are located in regions of very low interest, and thus to lower the effect of under-sampling of polar regions.

Appendix B

Cremer-Pople coordinates – Re-numbering of ring atoms

Definition of a re-numbering transformations

Every quantity in the Cremer-Pople framework is defined as weighted averages: given a starting clockwise numeration $\{j = 1, \ldots, N\}$, the weights[1]

$$\begin{cases} w_{m,j} = \sin(m\alpha_j) \\ v_{m,j} = \cos(m\alpha_j) \end{cases} \quad , \quad \alpha_j = \frac{2\pi(j-1)}{N} \quad , \tag{B.1}$$

are assigned to atom properties[2] $\mathcal{Q}_j$. The weights assigment (*i.e.* the actual values of $\mathcal{Q}_j w_{m,j}$ and $\mathcal{Q}_j v_{m,j}$) depends on atom numbering.

A *renumbering transformation* ReN assigns to the value $\mathcal{Q}_j$ a different weight in summations:

$$\sum_{j=1}^{N} \mathcal{Q}_j w_{m,j} \xrightarrow{\mathsf{ReN}} \sum_{j=1}^{N} \mathcal{Q}_j w_{\mathsf{ReN}(j),m} \quad , \quad \sum_{j=1}^{N} \mathcal{Q}_j v_{m,j} \xrightarrow{\mathsf{ReN}} \sum_{j=1}^{N} \mathcal{Q}_j v_{\mathsf{ReN}(j),m} \tag{B.2}$$

(or equivalently, $w_{m,j}$ and $v_{m,j}$ are assigned $\mathcal{Q}_{\mathsf{ReN}^{-1}(j)}$).

Given a starting, clockwise numeration ON $\{j = 1, \ldots, N\}$, a renumeration ReN could:

1. move the labelling by n places in clockwise direction:

$$j \xrightarrow{\mathsf{ShN_n}} k = [j - n + N] \bmod N \qquad \forall j = 1, \ldots, N \quad . \tag{B.3}$$

   After the *numbering shift* ShN_n the first atom of the new numeration will be the atom $n + 1$ of the previous numeration;

2. invert the sense of numeration:

$$j \xrightarrow{\mathsf{InN}} l = [N + 2 - j] \bmod N \qquad \forall j = 1, \ldots, N \quad . \tag{B.4}$$

[1] the weights $(-1)^{j-1}$ for the N even case willbe treated hereafter as a special case of $v_{m,j}$ with $m = N/2$

[2] *e.g.* the position vectors $\mathbf{R}_j$ and atoms elevations z_j from the mean plane, see Chapter 3

> After the *numbering inversion* InN the first atom of the new numeration is unchanged but the numeration is counterclockwise.

The refolding operation $\bmod N$ is needed to assess that the transformation gives $\{1,\ldots,N\} \xrightarrow{\mathsf{ReN}} \{1,\ldots,N\}$, namely that is a 1:1 map on positive integers from 1 to N. The refolding is by means of adding/subtracting extra N terms.

A renumbering ReN transforms the angles α_j as

$$\alpha_j = \frac{2\pi(j-1)}{N} \xrightarrow{\mathsf{ShN}_n} \alpha_k = \frac{2\pi(k-1)}{N} = \frac{2\pi(j-1)}{N} - \frac{2\pi n}{N} + \frac{2\pi\cancel{N}}{\cancel{N}} \tag{B.5a}$$

$$\xrightarrow{\mathsf{InN}} \alpha_l = \frac{2\pi(l-1)}{N} = \frac{2\pi\cancel{N}}{\cancel{N}} - \frac{2\pi(j-1)}{N} \tag{B.5b}$$

where the 2π terms are negligible for the periodicity of goniometric functions (possible refolding terms are negligible for the same reason). Thus, for our calculations $\mathsf{ShN}_n(j) = j - n$ and $\mathsf{InN}(j) = 2 - j$ (we will use the labels only for gonimetric functions of the α_j angles).

Transformation under shifting

From Eq. (B.5a) weights changes to

$$w_{m,j} \xrightarrow{\mathsf{ShN}_n} w_{m,j-n} = \sin m\alpha_j \underbrace{\cos m\frac{2\pi n}{N}}_{v_{m,n+1}} - \cos m\alpha_j \underbrace{\sin m\frac{2\pi n}{N}}_{w_{m,n+1}} \tag{B.6a}$$

$$v_{m,j} \xrightarrow{\mathsf{ShN}_n} v_{m,j-n} = \cos m\alpha_j \overbrace{\cos m\frac{2\pi n}{N}} + \sin m\alpha_j \overbrace{\sin m\frac{2\pi n}{N}} \tag{B.6b}$$

namely the new weights are linear combination of the old weights with terms connected to the shift angle $\alpha_{n+1} = 2\pi n/N$ (see Eqs. (3.9)).

The mean plane definition remains however unaltered: the summations

$$\sum_{j=1}^{N} z_j = 0 \xrightarrow{\mathsf{ShN}_n} \sum_{j=1}^{N} z_j = 0$$

$$\sum_{j=1}^{N} z_j w_{0,j} = 0 \xrightarrow{\mathsf{ShN}_n} \sum_{j=1}^{N} z_j w_{0,j-n} = v_{0,n+1}\sum_{j=1}^{N} z_j w_{0,j} - w_{0,n+1}\sum_{j=1}^{N} z_j v_{0,j} = 0$$

$$\sum_{j=1}^{N} z_j v_{0,j} = 0 \xrightarrow{\mathsf{ShN}_n} \sum_{j=1}^{N} z_j v_{0,j-n} = v_{0,n+1}\sum_{j=1}^{N} z_j v_{0,j} + w_{0,n+1}\sum_{j=1}^{N} z_j w_{0,j} = 0$$

are all unchanged because in the original numeration the summations are simultaneously zero by hypothesis.

The mean plane orientation is unchanged too:

$$\mathbf{R}' \xrightarrow{\mathsf{ShN}_n} \sum_{j=1}^{N} \mathbf{R}_j w_{0,j-n} = v_{0,n+1}\mathbf{R}' - w_{0,n+1}\mathbf{R}''$$

$$\mathbf{R}'' \xrightarrow{\mathsf{ShN}_n} \sum_{j=1}^{N} \mathbf{R}_j v_{0,j-n} = v_{0,n+1}\mathbf{R}'' + w_{0,n+1}\mathbf{R}'$$

from which we calculate

$$\mathbf{R}' \times \mathbf{R}'' \xrightarrow{\mathsf{ShN}_n} [v_{0,n+1}\mathbf{R}' - w_{0,n+1}\mathbf{R}''] \times [v_{0,n+1}\mathbf{R}'' + w_{0,n+1}\mathbf{R}'] =$$
$$= (v_{0,n+1}^2 + w_{0,n+1}^2)\mathbf{R}' \times \mathbf{R}'' + v_{0,n+1}w_{0,n+1}(\cancelto{0}{\mathbf{R}' \times \mathbf{R}'} - \cancelto{0}{\mathbf{R}'' \times \mathbf{R}''}) = \mathbf{R}' \times \mathbf{R}''$$

and the last equality holds because $v_{0,n+1}^2 + w_{0,n+1}^2 = 1$. Even if the mean plane is defined and oriented irrespective to the numbering scheme, the other axes of the Cremer-Pople frame $\{\hat{\mathrm{l}}, \hat{\mathrm{m}}, \hat{\mathrm{n}}\}$ changes. The new y axes by definition is the projection of the atom 1 on the mean plane, thus the unit vectors $\hat{\mathrm{l}}$ and $\hat{\mathrm{m}} = \hat{\mathrm{l}} \times \hat{\mathrm{n}}$ changes. Explicit formulae for this transformation are beyond the scope of this appendix.

The general definitions of (q_m, ϕ_m) and $q_{N/2}$ transform as

$$q_m \cos\phi_m \xrightarrow{\mathsf{ShN}_n} q'_m \cos\phi'_m = v_{m,n+1}q_m \cos\phi_m - w_{m,n+1}q_m \sin\phi_m \tag{B.8a}$$
$$q_m \sin\phi_m \xrightarrow{\mathsf{ShN}_n} q'_m \sin\phi'_m = v_{m,n+1}q_m \sin\phi_m + w_{m,n+1}q_m \cos\phi_m \tag{B.8b}$$
$$q_{N/2} \xrightarrow{\mathsf{ShN}_n} q'_{N/2} = (-1)^n q_{N/2} \quad . \tag{B.8c}$$

The renumbering affects clearly the amplitude $q_{N/2}$, if present. The other amplitudes are unchanged

$$\begin{aligned} q'_m &= \big[(v_{m,n+1}q_m \cos\phi_m - w_{m,n+1}q_m \sin\phi_m)^2 + \\ &\quad + (v_{m,n+1}q_m \sin\phi_m + w_{m,n+1}q_m \cos\phi_m)^2\big]^{1/2} \\ &= \sqrt{(v_{m,n+1}^2 + w_{m,n+1}^2)q_m^2} = q_m \end{aligned}$$

and the total puckering amplitude is invariant too:

$$Q = \sqrt{q_{N/2}^2 + \sum_m q_m^2} \xrightarrow{\mathsf{ShN}_n} \sqrt{(-1)^{2n}q_{N/2}^2 + \sum_m q_m^2} = Q \quad .$$

For the phase angles we have

$$\cos\phi'_m = v_{m,n+1}\cos\phi_m - w_{m,n+1}\sin\phi_m = \cos\left[\phi_m + m\frac{2\pi n}{N}\right]$$
$$\sin\phi'_m = v_{m,n+1}\sin\phi_m + w_{m,n+1}\cos\phi_m = \sin\left[\phi_m + m\frac{2\pi n}{N}\right]$$

that means $\phi'_m = \phi_m + m\dfrac{2\pi n}{N} = \phi_m + m\alpha_{n+1}$, namely the pseudorotational angles are shifted by means of the shift angle α_{n+1}.

With $N = 6$, for the three-dimensional Cartesian representations the transformation reads

$$q_x \xrightarrow{\mathsf{ShN}_n} q'_x = v_{m,n+1}q_x - w_{m,n+1}q_y \tag{B.9a}$$
$$q_y \xrightarrow{\mathsf{ShN}_n} q'_y = v_{m,n+1}q_y + w_{m,n+1}q_x \tag{B.9b}$$
$$q_z \xrightarrow{\mathsf{ShN}_n} q'_z = (-1)^n q_z \quad . \tag{B.9c}$$

which means a α_{n+1} rotation of the axes on the $q_x q_y$-plane, and an inversion of the polar axis $\hat{\mathrm{q}}_z$ if n is odd (*i.e.* an inversion of the poles on the puckering sphere: the chair conformer moves from the north to the south pole). For the three-dimensional spherical representation, to transform the θ angle we go back to the original definition

$$
\begin{aligned}
Q\sin\theta = q_2 &\xrightarrow{\mathsf{ShN}_n} \sin\theta' = \sin\theta \\
Q\cos\theta = q_3 &\xrightarrow{\mathsf{ShN}_n} \cos\theta' = (-1)^n\cos\theta = \cos(n\pi)\cos\theta = \cos(\theta \pm n\pi)
\end{aligned}
$$

and we obtain

$$
\begin{array}{ccc}
\theta & \xrightarrow[n \text{ even}]{\mathsf{ShN}_n} & \theta' = \theta \\
\cap & & \cap \\
[0,\pi] & & [0,\pi]
\end{array}
\quad , \quad
\begin{array}{ccc}
\theta & \xrightarrow[n \text{ odd}]{\mathsf{ShN}_n} & \theta' = \pi - \theta \\
\cap & & \cap \\
[0,\pi] & & [\pi,0]
\end{array}
\quad .
$$

The "inversion" of the θ' domain in the N odd case is understandable if we interpret $\theta \in [0,\pi]$ as a movement along θ from the chair (north pole) to the inverted chair (south pole), and $\theta' \in [\pi,0]$ a movement along θ' from the chair (now at the south pole) to the inverted chair (now at the north pole). In other words, what is calculated moving "from left to right" in $[\pi,0]$ is the same as what is calculated moving "from right to left" in $[0,\pi]$ In a compact form the colatitude angle transform as

$$
\theta \xrightarrow{\mathsf{ShN}_n} \theta' = (-1)^n\theta + n\pi \quad . \tag{B.10}
$$

Transformation under inversion

From Eq. (B.5b) weights changes to

$$
w_{m,j} \xrightarrow{\mathsf{InN}} w_{2-j,m} = \cancelto{0}{\sin m2\pi}\cos\alpha_j - \cancelto{1}{\cos m2\pi}\sin\alpha_j = -w_{m,j} \tag{B.11a}
$$

$$
v_{m,j} \xrightarrow{\mathsf{InN}} v_{2-j,m} = \cancelto{1}{\cos m2\pi}\cos\alpha_j + \cancelto{0}{\sin m2\pi}\sin\alpha_j = v_{m,j} \tag{B.11b}
$$

The mean plane definition remains again unaltered:

$$
\begin{aligned}
\sum_{j=1}^{N} z_j = 0 &\xrightarrow{\mathsf{InN}} \sum_{j=1}^{N} z_j = 0 \\
\sum_{j=1}^{N} z_j w_{0,j} = 0 &\xrightarrow{\mathsf{InN}} \sum_{j=1}^{N} z_j w_{2-j,0} = \sum_{j=1}^{N} z_j w_{0,j} = 0 \\
\sum_{j=1}^{N} z_j v_{0,j} = 0 &\xrightarrow{\mathsf{InN}} \sum_{j=1}^{N} z_j v_{0,j-n} = -\sum_{j=1}^{N} z_j v_{0,j} = 0
\end{aligned}
$$

are all unchanged because in the original numeration the summations are simultaneously zero by hypothesis.

The mean plane orientation instead is inverted:

$$\mathbf{R}' \xrightarrow{\mathsf{InN}} \sum_{j=1}^{N} \mathbf{R}_j w_{2-j,0} = \mathbf{R}'$$
$$\mathbf{R}'' \xrightarrow{\mathsf{InN}} \sum_{j=1}^{N} \mathbf{R}_j v_{2-j,0} = -\mathbf{R}''$$

from which we calculate

$$\mathbf{R}' \times \mathbf{R}'' \xrightarrow{\mathsf{InN}} -\mathbf{R}' \times \mathbf{R}'' \qquad \Rightarrow \qquad \hat{\mathrm{n}} \xrightarrow{\mathsf{InN}} -\hat{\mathrm{n}} \quad .$$

In this case also the expression of the whole Cremer-Pople axes could be given: since the first atom of the numeration is unchanged, we have simply

$$\{\hat{\mathrm{l}}, \hat{\mathrm{m}}, \hat{\mathrm{n}}\} \xrightarrow{\mathsf{InN}} \{-\hat{\mathrm{l}}, \hat{\mathrm{m}}, -\hat{\mathrm{n}}\} \quad . \tag{B.13}$$

The general definitions of (q_m, ϕ_m) and $q_{N/2}$ transform as

$$q_m \cos\phi_m \xrightarrow{\mathsf{InN}} q'_m \cos\phi'_m = q_m \cos\phi_m \tag{B.14a}$$
$$q_m \sin\phi_m \xrightarrow{\mathsf{InN}} q'_m \sin\phi'_m = -q_m \sin\phi_m \tag{B.14b}$$
$$q_{N/2} \xrightarrow{\mathsf{InN}} q'_{N/2} = q_{N/2} \tag{B.14c}$$

which turns to be a slight modification for the Cremer-Pople coordinates. Indeed, the original amplitudes are unchanged

$$q'_m = q_m \ (\forall\, m) \quad , \quad q'_{N/2} = q_{N/2} \quad \Rightarrow \quad Q' = Q \tag{B.15}$$

while the phase angles are simply inverted

$$\phi'_m = -\phi_m \Leftarrow \begin{cases} \cos\phi'_m & = \cos\phi_m \\ \sin\phi'_m & = -\sin\phi_m \end{cases} \quad (\forall\, m) \quad . \tag{B.16}$$

With $N = 6$, for the three-dimensional Cartesian representations the transformation reads

$$q_x \xrightarrow{\mathsf{InN}} q'_x = q_x \tag{B.17a}$$
$$q_y \xrightarrow{\mathsf{InN}} q'_y = -q_y \tag{B.17b}$$
$$q_z \xrightarrow{\mathsf{InN}} q'_z = q_z \quad . \tag{B.17c}$$

which means the inversion of the $\hat{\mathrm{q}}_y$ direction, accordingly to the inversion of the pseudorotational angle ϕ. For the three-dimensional spherical representation, instead, we can state directly that there are no changes in the θ angle

$$\theta \xrightarrow{\mathsf{InN}} \theta' = \theta \tag{B.18}$$

since the amplitudes q_2 and q_3 are unchanged by construction.

Appendix C

Cremer-Pople coordinate inversion

By coordinate inversion we mean the process of obtaining the atomic positions of a six-membered ring, given its Cremer-Pople puckering parameters. In general, it is not possible to reconstruct all the 18 Cartesian coordinates having only the 3 puckering parameters (Q, θ, ϕ) or equivalently (q_2, ϕ_2, q_3). However, if supplementary information is given (like the set of bond lengths b_{ij} and/or bond angles β_{ijk} of the ring), the reconstruction can be be done in five steps:

1. Calculation of z_j elevations

It is the only operation that directly involves the puckering parameters, using the inversion formula

$$z_j = \frac{1}{\sqrt{3}} q_2 \cos\left[\phi_2 + \frac{2\pi(j-1)}{3}\right] + \frac{1}{\sqrt{6}} q_3 (-1)^{j-1} \tag{C.1}$$

$$= Q\left\{\frac{1}{\sqrt{3}} \sin\theta \cos\left[\phi + \frac{2\pi(j-1)}{3}\right] + \frac{1}{\sqrt{6}} \cos\theta (-1)^{j-1}\right\} \tag{C.2}$$

(see Ref. [41, 40]) which gives the elevation of the atoms with respect to the Cremer-Pople mean plane.

2. Projection of bond lengths/angles on the mean plane

The projection of bond lengths b_{ij} and the bond angles β_{ijk} permits the reconstruction of the planar "shadow" of the ring. Standard values for these extra data could be taken from ideal structures of, for example, cyclohexane or

tetrahydropyran[1]. From Ref. [40] we have the projections

$$b_{ij} \to \qquad b'_{ij} = \sqrt{b_{ij}^2 - (z_j - z_i)^2} \tag{C.3}$$

$$\beta_{ijk} \to \quad \cos\beta'_{ijk} = \frac{(z_k - z_i)^2 - (z_j - z_i)^2 - (z_k - z_j)^2 + 2b_{ij}b_{jk}\cos\beta_{ijk}}{2b'_{ij}b'_{jk}} \tag{C.4}$$

with $i, j, k = 1, \ldots, 6$.

3. Ring partition

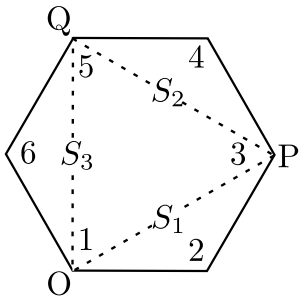


Figure C.1: Ring partition

The shadow of the ring is now divided into three parts S_n. According to Fig. C.1, in order to work with the S_n segments only few of the supplied data are needed at each segment, namely

$$\begin{aligned} S_1 &= \{1,2,3\} \quad \text{needs} \quad b'_{12}, b'_{23}, \beta'_{123} \\ S_2 &= \{3,4,5\} \quad \text{needs} \quad b'_{34}, b'_{45}, \beta'_{345} \\ S_3 &= \{5,6,1\} \quad \text{needs} \quad b'_{56}, b'_{61}, \beta'_{561} \end{aligned}$$

(all 6 bond length and 3 bond angles). For future use, we collect the lengths of the segments S_n, namely

$$|\overline{\mathsf{OP}}| = \sqrt{(b'_{12})^2 + (b'_{23})^2 - 2b'_{12}b'_{23}\cos\beta'_{123}} \tag{C.5}$$

$$|\overline{\mathsf{QP}}| = \sqrt{(b'_{34})^2 + (b'_{45})^2 - 2b'_{34}b'_{45}\cos\beta'_{345}} \tag{C.6}$$

$$|\overline{\mathsf{OQ}}| = \sqrt{(b'_{56})^2 + (b'_{61})^2 - 2b'_{56}b'_{61}\cos\beta'_{561}} \tag{C.7}$$

The segments S_n, and their endpoints O, P, and Q, are the skeleton for the ring reconstruction.

4. Coordinates calculation on ring partitions

Segments S_n can be now aligned in the arbitrary planar frame $\mathsf{O}x'y'$ (as shown in Fig. C.2) in which we can collect the coordinates $(x'_i, y'_i)_{S_n}$ of the i-th atom.

[1] bond lengths: $b_{\mathrm{CC}} = 0.154\,\mathrm{nm}$, $b_{\mathrm{CO}} = 0.143\,\mathrm{nm}$; bond angles: $\beta_{\mathrm{CCC}} = 109.5°$, $\beta_{\mathrm{CCO}} = 109.5°$, $\beta_{\mathrm{COC}} = 109.5°$

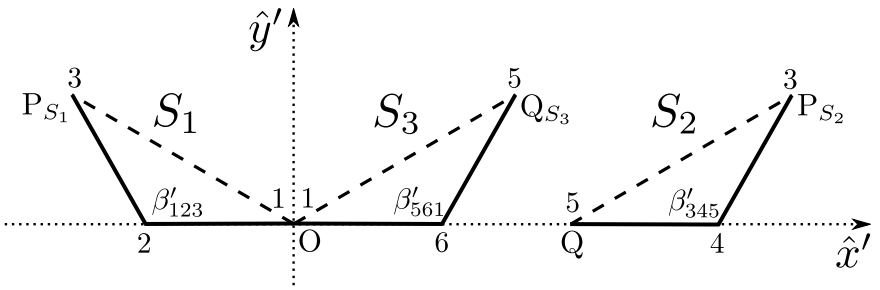


Figure C.2: Atoms coordinates

Atoms $\{1,2,3\} \in S_1$ are now located at

$$1_{S_1} \equiv \mathsf{O} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}_{S_1} \quad , \quad 2_{S_1} = \begin{pmatrix} -r'_{12} \\ 0 \end{pmatrix}_{S_1} \tag{C.8}$$

$$3_{S_1} \equiv \mathsf{P}_{S_1} = \begin{pmatrix} -r'_{12} + r'_{23}\cos\beta'_{123} \\ r'_{23}\sin\beta'_{123} \end{pmatrix}_{S_1} \tag{C.9}$$

Atoms $\{3,4,5\} \in S_2$ are now located at

$$5_{S_2} \equiv \mathsf{Q} = \begin{pmatrix} |\overline{\mathsf{OQ}}| \\ 0 \end{pmatrix}_{S_2} \quad , \quad 4_{S_2} = \begin{pmatrix} |\overline{\mathsf{OQ}}| + r'_{45} \\ 0 \end{pmatrix}_{S_2} \tag{C.10}$$

$$3_{S_2} \equiv \mathsf{P}_{S_2} = \begin{pmatrix} |\overline{\mathsf{OQ}}| + r'_{45} - r'_{34}\cos\beta'_{345} \\ r'_{34}\sin\beta'_{345} \end{pmatrix}_{S_2} \tag{C.11}$$

Atoms $\{5,6,1\} \in S_3$ are now located at

$$1_{S_3} \equiv \mathsf{O} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}_{S_3} \quad , \quad 6_{S_3} = \begin{pmatrix} r'_{61} \\ 0 \end{pmatrix}_{S_3} \tag{C.12}$$

$$5_{S_3} \equiv \mathsf{Q}_{S_3} = \begin{pmatrix} r'_{61} - r'_{56}\cos\beta'_{561} \\ r'_{56}\sin\beta'_{561} \end{pmatrix}_{S_3} \tag{C.13}$$

5. Calculation of x_j and y_j coordinates

The previous coordinates have only to be rotated properly in order to complete the reconstruction. Namely, a proper rotation matrix

$$\mathcal{R}_A(\phi) = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}$$

($R_A(\phi)$ is a counterclockwise rotation of angle ϕ around the pole A) is needed for each S_n segments. The procedure follows three steps:

i) We first need to rebuild the triangle OPQ, that is the skeleton for the ring. The vertex P of the OPQ triangle can be localized solving the system

$$\begin{cases} x^2 + y^2 = \overline{\mathsf{OP}}^2 & \text{(circle of center O and radius } \overline{\mathsf{OP}}) \\ (x - \overline{\mathsf{OQ}})^2 + y^2 = \overline{\mathsf{PQ}}^2 & \text{(circle of center Q and radius } \overline{\mathsf{PQ}}) \end{cases} \tag{C.14}$$

Eventually, taking the solution with $y > 0$, the intersection is at

$$x_{\mathsf{P}} = \frac{\overline{\mathsf{OP}}^2 + \overline{\mathsf{OQ}}^2 - \overline{\mathsf{PQ}}^2}{2\overline{\mathsf{OQ}}} \quad , \quad y_{\mathsf{P}} = \sqrt{\overline{\mathsf{OP}}^2 - \frac{(\overline{\mathsf{OP}}^2 + \overline{\mathsf{OQ}}^2 - \overline{\mathsf{PQ}}^2)^2}{4\overline{\mathsf{OQ}}^2}} \tag{C.15}$$

In Fig. C.3 points P_{S_1} and P_{S_2} has to be rotated to reach the position

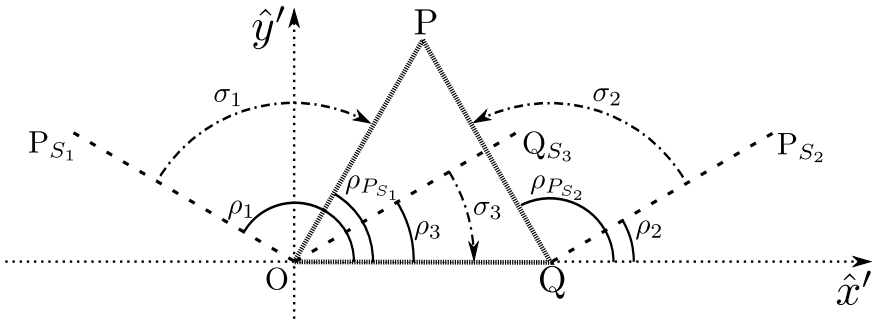


Figure C.3: P position and rotation angles

P for the reconstruction. The starting alignment of S_n segments gives the following angles measured from the $\hat{\mathrm{x}}'$ axis: for the end points of the partitions we have

$$\tan\rho_1 = \frac{y_{\mathsf{P}_{S_1}}}{x_{\mathsf{P}_{S_1}}} = \frac{r'_{23}\sin\beta'_{123}}{-r'_{12} + r'_{23}\cos\beta'_{123}} \tag{C.16a}$$

$$\tan\rho_2 = \frac{y_{\mathsf{P}_{S_2}}}{(x_{P_{S_2}} - \overline{\mathsf{OQ}})} = \frac{r'_{34}\sin\beta'_{345}}{r'_{45} - r'_{34}\cos\beta'_{345}} \tag{C.16b}$$

$$\tan\rho_3 = \frac{y_{\mathsf{Q}_{S_3}}}{x_{\mathsf{Q}_{S_3}}} = \frac{r'_{56}\sin\beta'_{561}}{r'_{61} - r'_{56}\cos\beta'_{561}} \tag{C.16c}$$

and for the real position of point P we have

$$\tan\rho_{\mathsf{P}_{S_1}} = \frac{y_{\mathsf{P}}}{x_{\mathsf{P}}} \quad , \quad \tan\rho_{\mathsf{P}_{S_2}} = \frac{y_{\mathsf{P}}}{x_{\mathsf{P}} - \overline{\mathsf{OQ}}} \tag{C.16d}$$

(see Fig. C.3 for angles definition).

ii) Now we can identify the proper rotation $\mathcal{R}_A(\phi)$ that maps the points P_{S_1} and P_{S_2} onto P and the point Q_{S_2} onto Q. The rotation needed (see Fig. C.3) are the following:

a) S_1 has to be rotated clockwise around O by angle $\sigma_1 = |\rho_1 - \rho_{\mathsf{P}_{S_1}}|$

b) S_2 has to be rotated counterclockwise around Q by angle $\sigma_2 = |\rho_{\mathsf{P}_{S_2}} - \rho_2|$

c) S_3 has to be rotated clockwise around O by angle $\sigma_3 = |\rho_3|$

In this way the coordinates $\mathbf{R}'_j = \begin{pmatrix} x_j \\ y_j \end{pmatrix}$ are

$$\mathbf{R}'_1 \equiv \mathsf{O} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{C.17a}$$

$$\mathbf{R}'_2 = \mathcal{R}_{\mathsf{O}}(-\sigma_1)2_{S_1} = \mathcal{R}_{\mathsf{O}}(-\sigma_1)\begin{pmatrix} -r'_{12} \\ 0 \end{pmatrix}_{S_1} \tag{C.17b}$$

$$\mathbf{R}'_3 \equiv \mathsf{P} = \begin{pmatrix} x_{\mathsf{P}} \\ y_{\mathsf{P}} \end{pmatrix} = \begin{pmatrix} \frac{\overline{\mathsf{OP}}^2+\overline{\mathsf{OQ}}^2-\overline{\mathsf{PQ}}^2}{2\overline{\mathsf{OQ}}} \\ \sqrt{\overline{\mathsf{OP}}^2 - \left(\frac{\overline{\mathsf{OP}}^2+\overline{\mathsf{OQ}}^2-\overline{\mathsf{PQ}}^2}{2\overline{\mathsf{OQ}}}\right)^2} \end{pmatrix} \tag{C.17c}$$

$$\mathbf{R}'_4 = \begin{pmatrix} \overline{\mathsf{OQ}} \\ 0 \end{pmatrix} + \mathcal{R}_{\mathsf{Q}}(\sigma_2)\left[\begin{pmatrix} |\overline{\mathsf{OQ}}| + r'_{45} \\ 0 \end{pmatrix}_{S_2} - \begin{pmatrix} \overline{\mathsf{OQ}} \\ 0 \end{pmatrix}_{S_2}\right] \tag{C.17d}$$

$$\mathbf{R}'_5 \equiv \mathsf{Q} = \begin{pmatrix} \overline{\mathsf{OQ}} \\ 0 \end{pmatrix} \tag{C.17e}$$

$$\mathbf{R}'_6 = \mathcal{R}_{\mathsf{O}}(-\sigma_3)6_{S_3} = \mathcal{R}_{\mathsf{O}}(-\sigma_3)\begin{pmatrix} r'_{61} \\ 0 \end{pmatrix}_{S_3} \tag{C.17f}$$

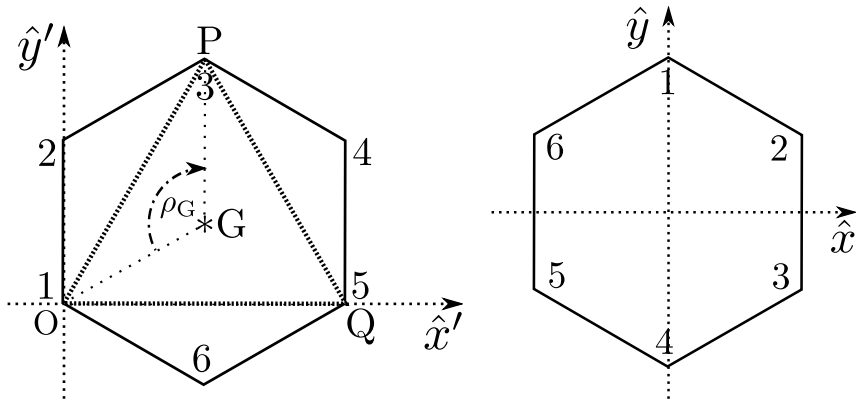
iii) Now with the $\mathbf{R}'_j$ vector we can calculate the geometric center G of the ring projection onto the mean plane

$$\mathbf{R}'_{\mathsf{G}} = \sum_{j=1}^{6} \mathbf{R}'_j \tag{C.18}$$

and the angle between the y' axes and the segment $\overline{\mathsf{OG}}$, that is

$$\tan\left[\rho_{\mathsf{G}} - \frac{\pi}{2}\right] = \frac{y_{\mathsf{G}}}{x_{\mathsf{G}}} \tag{C.19}$$

(see Fig. C.4(a) for angle definition). We are now ready to calculate the



(a) geometric center G location

(b) final ring atom positions (planar view)

Figure C.4: Final transformation

final coordinates of the atoms in the Cremer-Pople reference frame $\{\hat{\mathrm{l}}, \hat{\mathrm{m}}, \hat{\mathrm{n}}\}$ (see Section 3.2.1). This requires a global translation from O to G and a global rotation in such way that the atom 1 will be along the y' axis, namely a clockwise rotation around G by angle ρ_{G}. The transformation gives

$$\begin{pmatrix} x_j \\ y_j \end{pmatrix} = \mathcal{R}_{\mathsf{G}}(-\rho_{\mathsf{G}}) \left(\mathbf{R}'_j - \mathbf{R}'_{\mathsf{G}}\right) \tag{C.20}$$

for the final Cremer-Pople coordinates (x_j, y_j) (see Fig. C.4(b)). The full position vectors of ring atoms are now

$$\mathbf{R}_j = x_j \hat{\mathrm{l}} + y_j \hat{\mathrm{m}} + z_j \hat{\mathrm{n}} \quad .$$

Appendix D

Well-tempered metadynamics: convergence proof

During the simulation the bias will in time compensate the underlying $F(z)$ surface: the "equilibrium" probability at time t will be

$$\rho_{\text{B}}(z,t) = \frac{e^{-\beta[F(z)+V_{\text{B}}(z,t)]}}{\text{Z}(t)} \qquad . \tag{D.1}$$

For very large t the bias potential will vary slowly. As the amplitude $A(t)$ decreases in time by construction, the probability distribution will be nearly costant in time ($\dot{\rho_{\text{B}}} \simeq 0$) and then we can rewrite the hystogram $N(z,t)$ as

$$N(z,t) = \int_0^t \mathrm{d}t\delta(s - s(t)) = tP(z,t) \longrightarrow \dot{N}(z,t) = \rho_{\text{B}}(z,t) = \delta(sz - z(t)) \qquad .$$

Substituting in $\dot{V}_{\text{B}}$ we have

$$\begin{aligned}\dot{V}_G(z,t) = \omega e^{-V_{\text{B}}(z,t)/k_B\Delta T}\dot{N}(z,t) &\simeq \omega e^{-V_{\text{B}}(z,t)/k_B\Delta T}P(z,t) \\ &= \omega e^{-V_{\text{B}}(z,t)/k_B\Delta T}\frac{e^{-\beta[F(z)+V_{\text{B}}(z,t)]}}{\text{Z}(t)} \qquad .\end{aligned}$$

Hence, using $\beta' = 1/k_B\Delta T$ we have

$$\frac{\dot{V}_G(z,t)\text{Z}(t)}{\omega} = e^{-\beta' V_{\text{B}}(z,t) - \beta[F(z)+V_{\text{B}}(z,t)]} \tag{D.2}$$

from which taking the logarithm on both sides and neglecting the costant term $\ln\frac{\dot{V}_G(z,t)\text{Z}(t)}{\omega}$ we have

$$0 = -\beta' V_{\text{B}}(z,t) - \beta F(z) + \beta V_{\text{B}}(z,t) = \frac{1}{k_B T}\left[V_{\text{B}}(z,t)\frac{T+\Delta T}{\Delta T} + F(z)\right] \qquad . \tag{D.3}$$

In fact the term $\ln\frac{\dot{V}_G(z,t)\text{Z}(t)}{\omega}$ is not costant, is divergent because $\dot{V}_G \xrightarrow{t\to+\infty} 0$. Nevertheless, what we reach are free energy values that are consistently defined

up to additive costant, even if this costant is infinite, because we are only interested in free energy differences. Thus, in well-tempered metadynamics we have an exact/rigorous convergence

$$V_{\text{B}}(z, t \to +\infty) = -\frac{\Delta T}{T + \Delta T} F(z) \qquad . \tag{D.4}$$

Acknowledgements

I kindly thank my advisors (Prof. F.Pederiva and Dr. M.Sega) and all the people with whom I collaborated during my PhD (Dr. E.Chiessi, Prof G.Guella, Dr. A.Lonardi, to name a few). The work with all these people was at the same time challenging and interesting, I hope they had fun in doing it like I did.

I am glad to acknowledge the HPC facilities that made this work possible: the Wiglaf cluster (@ Department of physics, University of Trento) and the Aurora cluster (@ LISC, Trento). Part of the present work was supported by a PRIN grant from the Italian Ministry of Public Education, University and Scientific Research (PRIN 2007 project: "Proprietà fisiche di biomatrici nanostrutturate a base polimerica", A.Deriu, F.Pederiva, G.Paradossi).

Last but not least, I would like to acknowledge the thousands of individuals who have coded for the LaTeX project for free. It is due to their efforts that we can generate professionally typeset PDFs now.

Ringraziamenti

Se sei arrivato a leggere fin qui, complimenti.

Se sei saltato a leggere qui direttamente, complimenti ugualmente.

Cercherò di essere breve. "sarò breve" in genere è l'inizio di un lunghissimo sproloquio, ma spero di essere breve in verità. Anche perché, sinceramente, alla fine di questa fatica sono stanco oltre misura, con un lungo percorso alle spalle fatto e compiuto e la nebbia di fronte. Fatalmente, nell'essere breve lascerò fuori qualcuno da questi ringraziamenti. Chiunque non sia esplicitamente nominato non me ne voglia, oppure mi porti reclamo per la mancanza e provvederò diversamente.

Ringrazio (ancora ed in italiano) Francesco e Marcello, "grande" e "piccolo" capo di dottorato, rispettivamente. Ad entrambi, grazie per avermi dato qualcosa di interessante su cui lavorare per questi tre anni.

Ringrazio mamma e papà, perché nell'ultimo periodo hanno sopportato la mia monomania da tesi di dottorato con estrema pazienza. Spero che abbiate

ancora pazienza di scorta per quello che deve venire. Ringrazio il resto della famiglia (Mario e Israel e Silvana, i cognati Enrico e Marcella, i familiari allargati Paolo e Lina e Osvaldo): a tutti un "bubuli come prima". E ringrazio, anche se forse non leggeranno mai queste righe, Pietro e Luca e Francesca, i miei nipotini: lo zio ha finito, adesso può giocare con voi un po' di più.

Ringrazio Davide Campa, aspettando di sentire notizie sul suo futuro e di dargli notizie sul mio. Continueremo a vederci alla sagra di S.Croce per parlare di fisica?

Ringrazio Marco Zanatta per le chiacchierate impossibili tra prove per l'esistenza di Dio e ricette degli gnocchi (passando per il modello del *noon state*).

Ringrazio tutti i colleghi di dottorato degli ultimi anni: Diego, Enrico, Giorgia, Roberto, Elia, Anna, Paolo Armani, Gianluca,... Anche se la degenerazione delle conversazioni è all'ordine del giorno, meritate tutti il premio pazienza per le mie intemperanze canore. E ho detto tutto.

Ringrazio tutti gli amici e conoscenti informatici, per aver edotto un fisico sulle magie della programmazione. In particolare a Mitch Caceffo un grazie sperando prima o poi di fare una messa di Natale con un bel canto da "bandierone" fatto come si deve.

Ringrazio Franz e Barbara, costanti come un pendolo a tentare la difficile impresa di passare tempo assieme (ah, la mia agenda). E ringrazio le Marocchi, perché di gomitoli di lana ne abbiamo tirati tanti.

Ringrazio Cecilia e Matilde, che ci siete ancora dopo tutti questi anni. Incredibile!!!

Ringrazio Lidia ed Ema ed Elisa e Gloria e Matteo e Soma... ma basta matrimoni! Anche se ci sarebbero Cecilia e Gianna a breve...

Ringraziamenti vari vanno agli ex^n-coinquilini (dove $n = 0, 1, 2$ per le case che ho cambiato negli ultimi 3 anni), alle bariste di FBK, alle vicine di casa. Ringrazio poi Gnuplot, LaTeX (di nuovo), Inkscape, i programmi in C, gli script in Bash e in Python, e tutte le gigionate informatiche che mi hanno portato via centinaia di ore di tempo a fare... cose non meglio identificate. Mi sono divertito anche troppo con queste cose, lo ammetto. Ringrazio anche i videogiochi che mi hanno accompagnato, e come dimenticare le mie adorate serie TV. Tra prestiti e restituzioni, punti di ascolto per episodi chiave, discussioni accanite, per mia fortuna non sono state solo una maniera per isolarmi dal mondo.

Potevo scrivere meglio questi ringraziamenti, ma in fondo importa solo dire "grazie". Vorrei avere qualcosa di bello, di forte, di importante da dire in chiusura, ma non c'è molto altro in realtà da dire. Solo, grazie a tutti per esserci stati.

Se è possibile, per quanto dipende da voi,
vivete in pace con tutti gli uomini.

Lettera di S.Paolo Apostolo ai Romani 12:18

Bibliography

[1] Allen, M., and Tildesley, D. *Computer simulation of liquids*, vol. 18. Oxford university press, 1989. (See page 23).

[2] Almlöf, M., Kristensen, E. M. E., Siegbahn, H., and Aqvist, J. Molecular dynamics study of heparin based coatings. *Biomaterials 29*, 33 (Nov. 2008), 4463–9. (See page 102).

[3] Almond, A., Brass, A., and Sheehan, J. K. Deducing polymeric structure from aqueous molecular dynamics simulations of oligosaccharides: predictions from simulations of hyaluronan tetrasaccharides compared with hydrodynamic and X-ray fibre diffraction data. *Journal of molecular biology 284*, 5 (Dec. 1998), 1425–37. (See page 102).

[4] Altona, C., Francke, R., de Haan, R., Ippel, J. H., Daalmans, G. J., Hoekzema, A. J. a. W., and van Wijk, J. Empirical group electronegativities for vicinal NMR proton-proton couplings along a C-C bond: Solvent effects and reparameterization of the Haasnoot equation. *Magnetic Resonance in Chemistry 32*, 11 (Nov. 1994), 670–678. (See page 14).

[5] Angyal, S. J. The Composition and Conformation of Sugars in Solution. *Angewandte Chemie International Edition in English 8*, 3 (Mar. 1969), 157–166. (See pages 13, 14, 105, 108, 109, 115 and 116).

[6] Angyal, S. J., and Pickles, V. Equilibria between pyranoses and furanoses. II. Aldoses. *Australian Journal of Chemistry 25*, 8 (1972), 1695. (See pages 6 and 15).

[7] Appell, M., Strati, G., Willett, J. L., and Momany, F. B3LYP/6-311++G** study of alpha- and beta-D-glucopyranose and 1,5-anhydro-D-glucitol: 4C1 and 1C4 chairs, (3,O)B and B(3,O) boats, and skew-boat conformations. *Carbohydrate research 339*, 3 (Feb. 2004), 537–51. (See page 17).

[8] Autieri, E., Chiessi, E., Lonardi, A., Paradossi, G., and Sega, M. Conformation and Dynamics of Poly(N-isopropyl acrylamide) Trimers in Water: A Molecular Dynamics and Metadynamics Simulation Study. *The journal of physical chemistry. B 115*, 19 (May 2011), 5827–39. (See pages 108 and 116).

[9] Autieri, E., Faccioli, P., Sega, M., Pederiva, F., and Orland, H. Dominant reaction pathways in high-dimensional systems. *The Journal of chemical physics 130*, 6 (Feb. 2009), 064106. (See page 70).

[10] Autieri, E., Sega, M., Pederiva, F., and Guella, G. Puckering free energy of pyranoses: A NMR and metadynamics-umbrella sampling investigation. *The Journal of Chemical Physics 133*, 9 (2010), 095104. (See pages x, 15, 64, 66, 72, 93, 96, 98, 108, 109, 110, 116, 117 and 123).

[11] Autieri, E., Sega, M., Pederiva, F., and Guella, G. Erratum: "Puckering free energy of pyranoses: An NMR and metadynamics–umbrella sampling investigation" [J. Chem. Phys. 133, 095104 (2010)]. *The Journal of Chemical Physics 134*, 14 (2011), 149901. (See page 93).

[12] Babin, V., Roland, C., Darden, T. a., and Sagui, C. The free energy landscape of small peptides as obtained from metadynamics with umbrella sampling corrections. *The Journal of chemical physics 125*, 20 (Nov. 2006), 204909. (See pages 70 and 72).

[13] Babin, V., and Sagui, C. Conformational free energies of methyl-alpha-L-iduronic and methyl-beta-D-glucuronic acids in water. *The Journal of chemical physics 132*, 10 (Mar. 2010), 104108. (See page x).

[14] Barducci, A., Bonomi, M., and Parrinello, M. Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science 1*, October (Feb. 2011), n/a–n/a. (See pages x and 32).

[15] Barducci, A., Bussi, G., and Parrinello, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Physical Review Letters 100*, 2 (2008), 020603. (See page 70).

[16] Barrows, S. E., Dulles, F. J., Cramer, C. J., French, A. D., and Truhlar, D. G. Relative stability of alternative chair forms and hydroxymethyl conformations of β-d-glucopyranose. *Carbohydrate Research 276*, 2 (Oct. 1995), 219–251. (See page 17).

[17] Bartels, C., and Karplus, M. Multidimensional adaptive umbrella sampling: Applications to main chain and side chain peptide conformations. *Journal of Computational Chemistry 18*, 12 (Sept. 1997), 1450–1462. (See page x).

[18] Behler, J., Price, D. W., and Drew, M. G. B. Water structuring properties of carbohydrates, molecular dynamics studies on 1,5-anhydro-D-fructose. *Physical Chemistry Chemical Physics 3*, 4 (2001), 588–601. (See page 102).

[19] Bérces, A., Whitfield, D. M., and Nukada, T. Quantitative description of six-membered ring conformations following the IUPAC conformational nomenclature. *Tetrahedron 57*, 39 (Sept. 2001), 477–491. (See pages x, 46 and 47).

[20] Berendsen, H. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications 91*, 1-3 (Sept. 1995), 43–56. (See pages 23, 53, 64, 87 and 103).

[21] Berendsen, H., Postma, J., van Gunsteren, W., and Hermans, J. *Intermolecular Forces.* Reidel, Dordrecht, 1981, pp. 331–342. (See pages 64, 93, 94 and 103).

[22] Berendsen, H., and Van Gunsteren, W. Practical algorithms for dynamic simulations. In *Proc. Molecular dynamics simulation of statistical mechanical systems. Enrico Fermi Summer School* (1985), pp. 43–65. (See page 24).

[23] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics 81*, 8 (May 1984), 3684. (See page 26).

[24] Biarnés, X., Ardèvol, A., Planas, A., Rovira, C., Laio, A., and Parrinello, M. The conformational free energy landscape of beta-D-glucopyranose. Implications for substrate preactivation in beta-glucoside hydrolases. *Journal of the American Chemical Society 129*, 35 (Sept. 2007), 10686–93. (See pages x, 17, 86, 88 and 90).

[25] B.J.Alder, and Wainwright, T. *Molecular Dynamics by Electronic Computers, Proc. Intern. Symposium on Transport Processes in Statistical Mechanical theory of transport processes (Brussels, 1956).* Interscience, Wiley, New York, 1958, pp. 97–131. (See page 21).

[26] Boeyens, J. C. a., and Evans, D. G. Group theory of ring pucker. *Acta Crystallographica Section B Structural Science 45*, 6 (Dec. 1989), 577–581. (See pages 42, 45 and 92).

[27] Bonomi, M., Branduardi, D., Bussi, G., Camilloni, C., Provasi, D., Raiteri, P., Donadio, D., Marinelli, F., Pietrucci, F., and Broglia, R. a. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications 180*, 10 (Oct. 2009), 1961–1972. (See pages 53, 57, 61, 64, 94 and 103).

[28] Brady, J. W. Molecular dynamics simulations of alpha-D-glucose. *Journal of the American Chemical Society 108*, 26 (Dec. 1986), 8153–8160. (See page 102).

[29] Brochier-Salon, M.-C., and Morin, C. Conformational analysis of 6-deoxy-6-iodo-D-glucose in aqueous solution. *Magnetic Resonance in Chemistry 38*, 12 (Dec. 2000), 1041–1042. (See page 67).

[30] Bussi, G., Laio, A., and Parrinello, M. Equilibrium Free Energies from Nonequilibrium Metadynamics. *Physical Review Letters 96*, 9 (Mar. 2006), 10–13. (See page 36).

[31] Caffarena, E. R., and Grigera, J. Glass transition in aqueous solutions of glucose. Molecular dynamics simulation. *Carbohydrate Research 300*, 1 (May 1997), 51–57. (See page x).

[32] Camilloni, C., Provasi, D., Tiana, G., and Broglia, R. a. Exploring the protein G helix free-energy surface by solute tempering metadynamics. *Proteins 71*, 4 (June 2008), 1647–54. (See pages 53, 64, 87 and 103).

[33] Carter, E., Ciccotti, G., Hynes, J. T., and Kapral, R. Constrained reaction coordinate dynamics for the simulation of rare events. *Chemical Physics Letters 156*, 5 (Apr. 1989), 472–477. (See page 31).

[34] Cavalieri, F., Chiessi, E., Paci, M., Paradossi, G., Flaibani, A., and Cesàro, A. Conformational Dynamics of Hyaluronan in Solution. 1. A 13 C NMR Study of Oligomers. *Macromolecules 34*, 1 (Jan. 2001), 99–109. (See page 102).

[35] Chang, C., Magracheva, E., Kozlov, S., Fong, S., Tobin, G., Kotenko, S., Wlodawer, A., and Zdanov, A. Crystal structure of interleukin-19 defines a new subfamily of helical cytokines. *The Journal of biological chemistry 278*, 5 (Jan. 2003), 3308–13. (See page 16).

[36] Chiessi, E. private communication, 2009. private communication. (See pages x, 18 and 102).

[37] Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society 117*, 19 (May 1995), 5179–5197. (See page 102).

[38] Corzana, F., Motawia, M. S., Du Penhoat, C. H., Perez, S., Tschampel, S. M., Woods, R. J., and Engelsen, S. r. B. A hydration study of (1–>4) and (1–>6) linked alpha-glucans by comparative 10 ns molecular dynamics simulations and 500-MHz NMR. *Journal of computational chemistry 25*, 4 (Mar. 2004), 573–86. (See page 102).

[39] Cremer, D. A General Definition of Ring Substituent Positions. *Israel Journal of Chemistry 20* (1980), 12—-19. (See page 10).

[40] Cremer, D. Calculation of puckered rings with analytical gradients. *The Journal of Physical Chemistry 94*, 14 (July 1990), 5502–5509. (See pages 42, 46, 53, 96, 139 and 140).

[41] Cremer, D., and Pople, J. General definition of ring puckering coordinates. *Journal of the American Chemical Society 97*, 6 (Mar. 1975), 1354–1358. (See pages x, 42, 43, 44, 45 and 139).

[42] Crooks, G. Nonequilibrium Measurements of Free Energy Differences for Microscopically Reversible Markovian Systems. *Journal of Statistical Physics 90*, 5 (Mar. 1998), 1481–1487. (See page 31).

[43] Damm, W., Frontera, A., Tirado-Rives, J., and Jorgensen, W. OPLS all-atom force field for carbohydrates. *Journal of computational chemistry 18*, 16 (Dec. 1997), 1955–1970. (See pages 18 and 102).

[44] Darve, E., and Pohorille, A. Calculating free energies using average force. *The Journal of Chemical Physics 115*, 20 (2001), 9169. (See pages 31 and 32).

[45] Dellago, C., Bolhuis, P., and Geissler, P. Transition path sampling. *Advances in chemical physics* (2002), 1–78. (See page 32).

[46] Dellago, C., Bolhuis, P. G., Csajka, F. S., and Chandler, D. Transition path sampling and the calculation of rate constants. *The Journal of Chemical Physics 108*, 5 (1998), 1964. (See page 32).

[47] Dixon, B. F., Jeannin, Y., Loening, K. L., and Moss, G. P. Conformational Nomenclature for Five and Six-Membered Ring Forms of Monosaccharides and Their Derivatives. Recommendations 1980. *European Journal of Biochemistry 111*, 2 (Oct. 1980), 295–298. (See pages 8, 9, 40 and 46).

[48] Edward, J. Stability of glycosides to acid hydrolysis. *Chem. Ind.(London) 1102* (1955). (See page 13).

[49] Eklund, R., and Widmalm, G. Molecular dynamics simulations of an oligosaccharide using a force field modified for carbohydrates. *Carbohydrate Research 338*, 5 (Feb. 2003), 393–398. (See page 102).

[50] Essén, H., and Cremer, D. On the relationship between the mean plane and the least-squares plane of an N -membered puckered ring. *Acta Crystallographica Section B Structural Science 40*, 4 (Aug. 1984), 418–420. (See page 42).

[51] Evans, D. G., and Boeyens, J. C. a. Conformational analysis of ring pucker. *Acta Crystallographica Section B Structural Science 45*, 6 (Dec. 1989), 581–590. (See pages 42 and 45).

[52] Feller, S. E., Zhang, Y., Pastor, R. W., and Brooks, B. R. Constant pressure molecular dynamics simulation: The Langevin piston method. *The Journal of Chemical Physics 103*, 11 (1995), 4613. (See page 61).

[53] Ferrenberg, A., and Swendsen, R. New Monte Carlo technique for studying phase transitions. *Physical Review Letters 61*, 23 (Dec. 1988), 2635–2638. (See page 31).

[54] Figueiras, A., Sarraguça, J. M. G., Carvalho, R. A., Pais, A. A. C. C., and Veiga, F. J. B. Interaction of omeprazole with a methylated derivative of beta-cyclodextrin: phase solubility, NMR spectroscopy and molecular simulation. *Pharmaceutical research 24*, 2 (Feb. 2007), 377–89. (See page 102).

[55] Finelli, I., Chiessi, E., Galesso, D., Renier, D., and Paradossi, G. Gel-like structure of a hexadecyl derivative of hyaluronic acid for the treatment of osteoarthritis. *Macromolecular bioscience 9*, 7 (July 2009), 646–53. (See page ix).

[56] Frenkel, D., and Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*, 1st ed. Academic Press, Sept. 1996. (See pages 23, 24 and 26).

[57] Gandhi, N. S., and Mancera, R. L. Can current force fields reproduce ring puckering in 2-O-sulfo-alpha-L-iduronic acid? A molecular dynamics simulation study. *Carbohydrate research 345*, 5 (Mar. 2010), 689–95. (See page 18).

[58] Garrett, E. C., and Serianni, A. S. Ab initio molecular orbital calculations on furanose sugars: a study with the 6–31G basis set. *Carbohydrate Research 206*, 2 (Oct. 1990), 183–191. (See page 17).

[59] Garrett, E. C., and Serianni, A. S. *Computer Modeling of Carbohydrate Molecules*, vol. 430 of *ACS Symposium Series*. American Chemical Society, Washington, DC, July 1990. (See page 17).

[60] González-Outeiriño, J., Kadirvelraj, R., and Woods, R. J. Structural elucidation of type III group B Streptococcus capsular polysaccharide using molecular dynamics simulations: the role of sialic acid. *Carbohydrate research 340*, 5 (Apr. 2005), 1007–18. (See page 102).

[61] Guvench, O., Greene, S. N., Kamath, G., Brady, J. W., Venable, R. M., Pastor, R. W., and Mackerell, A. D. Additive empirical force field for hexopyranose monosaccharides. *Journal of computational chemistry 29*, 15 (Nov. 2008), 2543–64. (See pages 18, 61 and 123).

[62] Guvench, O., Hatcher, E. R., Venable, R. M., Pastor, R. W., and Mackerell, A. D. CHARMM Additive All-Atom Force Field for Glycosidic Linkages between Hexopyranoses. *Journal of chemical theory and computation 5*, 9 (Aug. 2009), 2353–2370. (See page 18).

[63] Haasnoot, C. a. G. The relationship between proton-proton NMR coupling constants and substituent electronegativities—I An empirical generalization of the karplus equation. *Tetrahedron 36*, 19 (1980), 2783–2792. (See page 14).

[64] Haasnoot, C. a. G. The conformation of six-membered rings described by puckering coordinates derived from endocyclic torsion angles. *Journal of the American Chemical Society 114*, 3 (Jan. 1992), 882–887. (See pages x, 46 and 47).

[65] Hansen, H. S., and Hünenberger, P. H. A reoptimized GROMOS force field for hexopyranose-based carbohydrates accounting for the relative free energies of ring conformers, anomers, epimers, hydroxymethyl rotamers, and glycosidic linkage conformers. *Journal of computational chemistry 32*, 6 (Nov. 2010), 998–1032. (See pages 93, 107 and 114).

[66] Hansen, H. S., and Hünenberger, P. H. Using the local elevation method to construct optimized umbrella sampling potentials: calculation of the relative free energies and interconversion barriers of glucopyranose ring conformers in water. *Journal of computational chemistry 31*, 1 (Jan. 2010), 1–23. (See pages x, 18, 32, 92, 95, 97, 102, 107 and 113).

[67] Hardy, B., and Sarko, A. Molecular dynamics simulation of cellobiose in water. *Journal of Computational Chemistry 14*, 7 (July 1993), 848–857. (See page 102).

[68] Hardy, B. J., and Sarko, A. Molecular dynamics simulations and diffraction-based analysis of the native cellulose fibre: structural modelling of the I-α and I-β phases and their interconversion. *Polymer 37*, 10 (May 1996), 1833–1839. (See page 102).

[69] Hassel, O., and Ottar, B. The structure of molecules containing cyclohexane or pyranose rings. *Acta Chemica Scandinavica 1* (1947), 929—-942. (See pages 12 and 13).

[70] Haworth, W. N. *The constitution of sugars.* Edward Arnold & Co, London, 1929. (See pages 7 and 12).

[71] Heymann, B. 'Chair–boat' transitions and side groups affect the stiffness of polysaccharides. *Chemical Physics Letters 305*, 3-4 (May 1999), 202–208. (See page 102).

[72] Hill, A. D., and Reilly, P. J. Puckering coordinates of monocyclic rings by triangular decomposition. *Journal of chemical information and modeling 47*, 3 (2007), 1031–5. (See pages 46 and 51).

[73] Hockney, R., Goel, S., and Eastwood, J. Quiet high-resolution computer models of a plasma. *Journal of Computational Physics 14*, 2 (Feb. 1974), 148–158. (See page 24).

[74] Hoover, W. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A 31*, 3 (Mar. 1985), 1695–1697. (See pages 25, 64, 94 and 103).

[75] Hoover, W. Constant-pressure equations of motion. *Physical Review A 34*, 3 (Sept. 1986), 2499–2500. (See page 25).

[76] Howard, E. I., and Raul Grigera, J. On the sweetness properties of aldoses: characterization of molecular active sites by computer simulation. *Carbohydrate Research 282*, 1 (Feb. 1996), 25–40. (See page x).

[77] Ikemizu, S., Sparks, L. M., van der Merwe, P. a., Harlos, K., Stuart, D. I., Jones, E. Y., and Davis, S. J. Crystal structure of the CD2-binding domain of CD58 (lymphocyte function-associated antigen 3) at 1.8-A resolution. *Proceedings of the National Academy of Sciences of the United States of America 96*, 8 (Apr. 1999), 4289–94. (See page 16).

[78] Immel, S., Fujita, K., and Lichtenthaler, F. W. Solution Geometries and Lipophilicity Patterns ofα-Cycloaltrin. *Chemistry - A European Journal 5*, 11 (Nov. 1999), 3185–3192. (See pages 16 and 118).

[79] Isbell, H. S. OXIDATION OF THE ALPHA AND BETA FORMS OF THE SUGARS. *Journal of the American Chemical Society 54*, 4 (1932), 1692–1693. (See page 12).

[80] Isbell, H. S. Chemistry of the Carbohydrates and Glycosides, June 1940. (See page 12).

[81] Isbell, H. S., and Pigman, W. W. A STUDY OF THE α- AND β-ALDOSES AND THEIR SOLUTIONS BY BROMINE OXIDATION AND MUTAROTATION MEASUREMENTS. *The Journal of Organic Chemistry 1*, 6 (1937), 505–539. (See page 12).

[82] Jarzynski, C. Nonequilibrium Equality for Free Energy Differences. *Physical Review Letters 78*, 14 (Apr. 1997), 2690–2693. (See page 31).

[83] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics 79*, 2 (1983), 926. (See page 61).

[84] Joshi, N. V., and Rao, V. R. Flexibility of the pyranose ring in \alpha- and \beta-D-glucoses. *Biopolymers 18*, 12 (Dec. 1979), 2993–3004. (See page 13).

[85] Karplus, M. Contact Electron-Spin Coupling of Nuclear Magnetic Moments. *The Journal of Chemical Physics 30*, 1 (1959), 11. (See page 14).

[86] Karplus, M. Vicinal Proton Coupling in Nuclear Magnetic Resonance. *Journal of the American Chemical Society 85*, 18 (1963), 2870–2871. (See page 14).

[87] Karplus, M., and Petsko, G. A. Molecular dynamics simulations in biology. *Nature 347*, 6294 (Oct. 1990), 631–9. (See page 16).

[88] Kelly, R. A Relationship Between the Conformations of Cyclohexane Derivatives and Their Physical Properties. *Canadian Journal of Chemistry 35* (1957), 149—-155. (See page 13).

[89] Kilpatrick, J. E., Pitzer, K. S., and Spitzer, R. The Thermodynamics and Molecular Structure of Cyclopentane. *Journal of the American Chemical Society 69*, 10 (1947), 2483–2488. (See page 40).

[90] Kim, Y., and Sung, W. Membrane curvature induced by polymer adsorption. *Physical Review E 63*, 4 (Mar. 2001). (See page 102).

[91] Kirschner, K. N., Yongye, A. B., Tschampel, S. M., González-Outeiriño, J., Daniels, C. R., Foley, B. L., and Woods, R. J. GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *Journal of computational chemistry 29*, 4 (Mar. 2008), 622–55. (See pages 18 and 123).

[92] Koehler, J. E. H., Saenger, W., and Gunsteren, W. F. A molecular dynamics simulation of crystalline alpha-cyclodextrin hexahydrate. *European Biophysics Journal 15*, 4 (1987), 197–210. (See page 102).

[93] Kony, D. B., Damm, W., Stoll, S., van Gunsteren, W. F., and Hünenberger, P. H. Explicit-solvent molecular dynamics simulations of the polysaccharide schizophyllan in water. *Biophysical journal 93*, 2 (July 2007), 442–55. (See page 102).

[94] Kouwizjer, M. L. C. E., and Grootenhuis, P. D. J. Parametrization and application of CHEAT95, and extended atom force field for hydrated oligosaccharides. *The Journal of Physical Chemistry 99*, 36 (Sept. 1995), 13426–13436. (See page 102).

[95] Kräutler, V., Müller, M., and Hünenberger, P. H. Conformation, dynamics, solvation and relative stabilities of selected beta-hexopyranoses in water: a molecular dynamics study with the GROMOS 45A4 force field. *Carbohydrate research 342*, 14 (Oct. 2007), 2097–124. (See pages 102, 110 and 111).

[96] Kroon-Batenburg, L. M. J., Kruiskamp, P. H., Vliegenthart, J. F. G., and Kroon, J. Estimation of the Persistence Length of Polymers by MD Simulations on Small Fragments in Solution. Application to Cellulose. *The Journal of Physical Chemistry B 101*, 42 (Oct. 1997), 8454–8459. (See pages 18 and 102).

[97] Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., and Kollman, P. A. Multidimensional free-energy calculations using the weighted histogram analysis method. *Journal of Computational Chemistry 16*, 11 (Nov. 1995), 1339–1350. (See page 31).

[98] Laine, R. a. Information capacity of the carbohydrate code. *Pure and Applied Chemistry 69*, 9 (1997), 1867–1874. (See pages 4 and 11).

[99] Laio, A., and Gervasio, F. L. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics 71*, 12 (Dec. 2008), 126601. (See pages x, 32, 33, 34, 35, 36, 37, 61 and 65).

[100] Laio, A., and Parrinello, M. Escaping free-energy minima. *Proceedings of the National Academy of Sciences of the United States of America 99*, 20 (Oct. 2002), 12562–6. (See pages x and 32).

[101] Laio, A., Rodriguez-Fortea, A., Gervasio, F. L., Ceccarelli, M., and Parrinello, M. Assessing the accuracy of metadynamics. *The journal of physical chemistry. B 109*, 14 (Apr. 2005), 6714–21. (See page 32).

[102] Lawtrakul, L. Molecular dynamics simulations of β-cyclodextrin in aqueous solution. *International Journal of Pharmaceutics 256*, 1-2 (Apr. 2003), 33–41. (See page 102).

[103] Lee, G., Nowak, W., Jaroniec, J., Zhang, Q., and Marszalek, P. E. Nanomechanical control of glucopyranose rotamers. *Journal of the American Chemical Society 126*, 20 (May 2004), 6218–9. (See pages 16 and 102).

[104] Lee, K.-H., Benson, D. R., and Kuczera, K. Transitions from α to π Helix Observed in Molecular Dynamics Simulations of Synthetic Peptides †. *Biochemistry 39*, 45 (Nov. 2000), 13737–13747. (See page 102).

[105] Lelièvre, T., Rousset, M., and Stoltz, G. *Free Energy Computations: a mathematical perspective*, 1 ed. Imperial College Press, 2010. (See pages 24, 25, 27 and 28).

[106] Lemieux, R. U., Kullnig, R. K., Bernstein, H. J., and Schneider, W. G. Configurational Effects on the Proton Magnetic Resonance Spectra of Six-membered Ring Compounds1. *Journal of the American Chemical Society 80*, 22 (1958), 6098–6105. (See pages 13 and 14).

[107] Li, H. Single-molecule force spectroscopy on polysaccharides by AFM – nanomechanical fingerprint of α-(1,4)-linked polysaccharides. *Chemical Physics Letters 305*, 3-4 (May 1999), 197–201. (See page 16).

[108] Lichtenthaler, F. W. 4,6-Di-O-benzoyl-3-O-benzyl-alfa-D-arabino-hexo-pyranos-2-ulosyl bromide: A conveniently accessible glycosyl donor for the expedient construction of diantennary beta-D-mannosides branched at O-3 and O-6. *Carbohydrate Research 305*, 2 (Dec. 1997), 293–303. (See page 16).

[109] Limbach, H. J., and Ubbink, J. Structure and dynamics of maltooligomer–water solutions and glasses. *Soft Matter 4*, 9 (July 2008), 1887. (See pages 18 and 102).

[110] Lindahl, E., and Edholm, O. Spatial and energetic-entropic decomposition of surface tension in lipid bilayers from molecular dynamics simulations. *The Journal of Chemical Physics 113*, 9 (2000), 3882. (See page 61).

[111] Lindahl, E., Hess, B., and Van Der Spoel, D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Journal of Molecular Modeling* (2001), 306–317. (See pages 23, 53, 64, 87 and 103).

[112] Lins, R. D., and Hünenberger, P. H. A new GROMOS force field for hexopyranose-based carbohydrates. *Journal of computational chemistry 26*, 13 (Oct. 2005), 1400–12. (See pages ix, x, 18, 67, 86, 93, 102 and 104).

[113] Mackerell, A. D. Empirical force fields for biological macromolecules: overview and issues. *Journal of computational chemistry 25*, 13 (Oct. 2004), 1584–604. (See pages 17 and 18).

[114] Mamonova, T., and Kurnikova, M. Structure and energetics of channel-forming protein-polysaccharide complexes inferred via computational statistical thermodynamics. *The journal of physical chemistry. B 110*, 49 (Dec. 2006), 25091–100. (See page 102).

[115] Marszalek, P. E., Li, H., Oberhauser, A. F., and Fernandez, J. M. Chair-boat transitions in single polysaccharide molecules observed with force-ramp AFM. *Proceedings of the National Academy of Sciences of the United States of America 99*, 7 (Apr. 2002), 4278–83. (See page 16).

[116] Marszalek, P. E., Oberhauser, A. F., Pang, Y. P., and Fernandez, J. M. Polysaccharide elasticity governed by chair-boat transitions of the glucopyranose ring. *Nature 396*, 6712 (Dec. 1998), 661–4. (See page 16).

[117] Martyna, G. J., Tobias, D. J., and Klein, M. L. Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics 101*, 5 (Nov. 1994), 4177. (See page 61).

[118] Maveyraud, L., Niwa, H., Guillet, V., Svergun, D. I., Konarev, P. V., Palmer, R. a., Peumans, W. J., Rougé, P., Van Damme, E. J. M., Reynolds, C. D., and Mourey, L. Structural basis for sugar recognition, including the Tn carcinoma antigen, by the lectin SNA-II from Sambucus nigra. *Proteins 75*, 1 (Apr. 2009), 89–103. (See page 16).

[119] McNaught, A. D. Nomenclature of carbohydrates (IUPAC Recommendations 1996). *Pure and Applied Chemistry 68*, 10 (1996), 1919–2008. (See pages 5, 7, 8 and 46).

[120] Merlitz, H., and Wenzel, W. Comparison of stochastic optimization methods for receptor–ligand docking. *Chemical Physics Letters 362*, 3-4 (Aug. 2002), 271–277. (See page 32).

[121] Mezei, M. Adaptive umbrella sampling: Self-consistent determination of the non-Boltzmann bias. *Journal of Computational Physics 68*, 1 (Jan. 1987), 237–248. (See page 72).

[122] Miyamoto, S., and Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of Computational Chemistry 13*, 8 (Oct. 1992), 952–962. (See page 103).

[123] Momany, F. A., and Willett, J. L. Molecular dynamics calculations on amylose fragments. I. Glass transition temperatures of maltodecaose at 1, 5, 10, and 15.8% hydration. *Biopolymers 63*, 2 (Mar. 2002), 99–110. (See page 102).

[124] Naidoo, K. J., and Kuttel, M. Water structure about the dimer and hexamer repeat units of amylose from molecular dynamics computer simulations. *Journal of Computational Chemistry 22*, 4 (Mar. 2001), 445–456. (See page 102).

[125] Neelov, I. M., Adolf, D. B., McLeish, T. C. B., and Paci, E. Molecular dynamics simulation of dextran extension by constant force in single molecule AFM. *Biophysical journal 91*, 10 (Nov. 2006), 3579–88. (See page 102).

[126] Nishida, Y., Hori, H., Ohrui, H., and Meguro, H. 1 H NMR Analyses of Rotameric Distribution of C5-C6 bonds of D-Glucopyranoses in Solution. *Journal of Carbohydrate Chemistry 7*, 1 (Mar. 1988), 239–250. (See page 67).

[127] Nishida, Y., Ohrui, H., and Meguro, H. 1H-NMR studies of (6r)- and (6s)-deuterated d-hexoses: assignment of the preferred rotamers about C5-C6 bond of D-glucose and D-galactose derivatives in solutions. *Tetrahedron Letters 25*, 15 (Jan. 1984), 1575–1578. (See page 67).

[128] Nogami, Y., Nasu, K., Koga, T., Ohta, K., Fujita, K., Immel, S., Lindner, H. J., Schmitt, G. E., and Lichtenthaler, F. W. Synthesis, Structure, and Conformational Features ofα-Cycloaltrin: A

Cycloeligosaccharide with Alternating4C1/1C4 Pyranoid Chairs. *Angewandte Chemie International Edition in English 36*, 17 (Sept. 1997), 1899–1902. (See page 16).

[129] Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics 81*, 1 (1984), 511. (See pages 25, 64, 94 and 103).

[130] Nosé, S. A molecular dynamics method for simulations in the canonical ensemble. *Molecular Physics 100*, 1 (Jan. 2002), 191–198. (See page 25).

[131] Nosé, S., and Klein, M. Constant pressure molecular dynamics for molecular systems. *Molecular Physics 50*, 5 (Dec. 1983), 1055–1076. (See page 25).

[132] Ohrui, H., Nishida, Y., Higuchi, H., Hori, H., and Meguro, H. The preferred rotamer about the C 5-C 6 bond of D -galactopyranoses and the stereochemistry of dehydrogenation by D -galactose oxidase. *Canadian Journal of Chemistry 65*, 6 (June 1987), 1145–1153. (See page 67).

[133] Oostenbrink, C., Villa, A., Mark, A. E., and van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *Journal of computational chemistry 25*, 13 (Oct. 2004), 1656–76. (See page 18).

[134] Ott, K.-H., and Meyer, B. Parametrization of GROMOS force field for oligosaccharides and assessment of efficiency of molecular dynamics simulations. *Journal of Computational Chemistry 17*, 8 (June 1996), 1068–1084. (See page 102).

[135] Palleschi, A., Bocchinfuso, G., Coviello, T., and Alhaique, F. Molecular dynamics investigations of the polysaccharide scleroglucan: first study on the triple helix structure. *Carbohydrate research 340*, 13 (Sept. 2005), 2154–62. (See page 102).

[136] Paradossi, G., Cavalieri, F., and Chiessi, E. A Conformational Study on the Algal Polysaccharide Ulvan. *Macromolecules 35*, 16 (July 2002), 6404–6411. (See page 102).

[137] Paradossi, G., Chiessi, E., Barbiroli, A., and Fessas, D. Xanthan and Glucomannan Mixtures: Synergistic Interactions and Gelation. *Biomacromolecules 3*, 3 (May 2002), 498–504. (See page 102).

[138] Parrinello, M., and Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics 52*, 12 (1981), 7182. (See pages 27, 64, 94 and 103).

[139] Patey, G. N., and Valleau, J. P. A Monte Carlo method for obtaining the interionic potential of mean force in ionic solution. *The Journal of Chemical Physics 63*, 6 (1975), 2334. (See page 31).

[140] Pérez, S. A comparison and chemometric analysis of several molecular mechanics force fields and parameter sets applied to carbohydrates. *Carbohydrate Research 314*, 3-4 (Dec. 1998), 141–155. (See pages 18 and 102).

[141] Peters, B., and Trout, B. L. Obtaining reaction coordinates by likelihood maximization. *The Journal of chemical physics 125*, 5 (Aug. 2006), 054108. (See page 32).

[142] Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., and Schulten, K. Scalable molecular dynamics with NAMD. *Journal of computational chemistry 26*, 16 (Dec. 2005), 1781–802. (See pages 53 and 61).

[143] Pickett, H. M. Symmetry and Conformation of the Cycloalkanes. *The Journal of Chemical Physics 55*, 1 (1971), 324. (See pages 41 and 45).

[144] Polavarapu, P., and Ewig, C. Ab Initio computed molecular structures and energies of the conformers of glucose. *Journal of Computational Chemistry 13*, 10 (Dec. 1992), 1255–1261. (See page 17).

[145] Rao, V. R. Theoretical studies on the conformation of aldohexopyranoses. *Carbohydrate Research 17*, 2 (Apr. 1971), 341–352. (See page 13).

[146] Rao, V. R., Qasba, P. K., Balaji, P. V., and Chandrasekhar, R. *Conformation of Carbohydrates*, 1 ed. harwood academic publisher, 1998. (See pages 4, 5, 6, 10, 14, 15 and 93).

[147] Reeves, R. E. The Shape of Pyranoside Rings. *Journal of the American Chemical Society 72*, 4 (1950), 1499–1506. (See pages 12 and 13).

[148] Reeves, R. E. Cuprammonium-glycoside complexes. *Advances in Carbohydrate Chemistry 6* (Jan. 1951), 107–134. (See pages 12, 13 and 14).

[149] Reeves, R. E. Cuprammonium-Glycoside Complexes. VII. Glucopyranoside Ring Conformations in Amylose. *Journal of the American Chemical Society 76*, 18 (1954), 4595–4598. (See page 14).

[150] Roberts, C. J., Debenedetti, P. G., and Stillinger, F. H. Equation of State of the Energy Landscape of SPC/E Water. *The Journal of Physical Chemistry B 103*, 46 (Nov. 1999), 10258–10265. (See page x).

[151] Rodriguez-Gomez, D., Darve, E., and Pohorille, A. Assessing the efficiency of free energy calculation methods. *The Journal of chemical physics 120*, 8 (Mar. 2004), 3563–78. (See pages 31 and 32).

[152] Roux, B. The calculation of the potential of mean force using computer simulations. *Computer Physics Communications 91*, 1-3 (Sept. 1995), 275–282. (See page 31).

[153] Ryckaert, J., Ciccotti, G., and Berendsen, H. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics 23*, 3 (Mar. 1977), 327–341. (See pages 57, 61, 87, 94 and 103).

[154] Sachse, H. Ueber die geometrischen Isomerien der Hexamethylenderivate. *Berichte der deutschen chemischen Gesellschaft 23*, 1 (Jan. 1890), 1363–1370. (See pages x, 12 and 40).

[155] Sachse, H. Ueber die konfigurationen der polymethylenringe. *Zeitschrift fuer physikalische Chemie 10* (1892), 203–41. (See pages 12 and 40).

[156] Sandoval, C., Castro, C., Gargallo, L., Radic, D., and Freire, J. Specific interactions in blends containing Chitosan and functionalized polymers. Molecular dynamics simulations. *Polymer 46*, 23 (Nov. 2005), 10437–10442. (See page 102).

[157] Schnupf, U., Willett, J. L., and Momany, F. DFTMD studies of glucose and epimers: anomeric ratios, rotamer populations, and hydration energies. *Carbohydrate research 345*, 4 (Feb. 2010), 503–11. (See page 17).

[158] Schwarz, J. C. P. Rules for conformation nomenclature for five- and six-membered rings in monosaccharides and their derivatives. *Journal of the Chemical Society, Perkin Transactions 1* (1973), X002. (See pages 8 and 46).

[159] Sega, M., Autieri, E., and Pederiva, F. On the calculation of puckering free energy surfaces. *The Journal of chemical physics 130*, 22 (June 2009), 225102. (See pages x, 17 and 53).

[160] Serianni, A. S., and Podlasek, C. A. 13C-1H spin-coupling constants in carbohydrates : magnitude and sign determinations via 2D NMR methods. *Carbohydrate research 259*, 2 (1994), 277–282. (See page 16).

[161] Sharon, N., and Lis, H. History of lectins: from hemagglutinins to biological recognition molecules. *Glycobiology 14*, 11 (Nov. 2004), 53R–62R. (See page 11).

[162] Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., Bank, J. a., Jumper, J. M., Salmon, J. K., Shan, Y., and Wriggers, W. Atomic-level characterization of the structural dynamics of proteins. *Science (New York, N.Y.) 330*, 6002 (Oct. 2010), 341–6. (See page 21).

[163] Snyder, J. R., and Serianni, A. S. D-Idose: a one- and two-dimensional NMR investigation of solution composition and conformation. *The Journal of Organic Chemistry 51*, 14 (July 1986), 2694–2702. (See pages x, 108, 109, 116 and 118).

[164] Sonne, J., Hansen, F. Y., and Peters, G. H. Methodological problems in pressure profile calculations for lipid bilayers. *The Journal of chemical physics 122*, 12 (Mar. 2005), 124903. (See page 61).

[165] Spieser, S. Improved carbohydrate force field for GROMOS: ring and hydroxymethyl group conformations and exo-anomeric effect. *Carbohydrate Research 322*, 3-4 (Dec. 1999), 264–273. (See pages 18, 102 and 112).

[166] Spiwok, V., Králová, B., and Tvaroska, I. Modelling of beta-D-glucopyranose ring distortion in different force fields: a metadynamics study. *Carbohydrate research 345*, 4 (Feb. 2010), 530–7. (See pages x, 18, 102 and 107).

[167] Sponsler, O., and Dore, W. The structure of ramie cellulose as derived from x-ray data. In *Fourth Colloid Symposium Monograph* (1926), vol. 41, pp. 174–202. (See pages 7 and 12).

[168] Sprik, M., and Ciccotti, G. Free energy from constrained molecular dynamics. *The Journal of Chemical Physics 109*, 18 (1998), 7737. (See page 31).

[169] Stoddart, J. F. *Stereochemistry of carbohydrates*. Wiley-Interscience, New York, 1971. (See page 86).

[170] Stortz, C. a., Johnson, G. P., French, A. D., and Csonka, G. I. Comparison of different force fields for the study of disaccharides. *Carbohydrate research 344*, 16 (Nov. 2009), 2217–28. (See page 18).

[171] Strauss, H. L., and Pickett, H. M. Conformational structure, energy, and inversion rates of cyclohexane and some related oxanes. *Journal of the American Chemical Society 92*, 25 (Dec. 1970), 7281–7290. (See pages 40 and 41).

[172] Sugita, Y. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters 314*, 1-2 (Nov. 1999), 141–151. (See page 32).

[173] Sundararajan, P. Theoretical studies on the conformation of aldopyranoses. *Tetrahedron 24*, 1 (Apr. 1968), 289–295. (See page 13).

[174] Thibaudeau, C., Stenutz, R., Hertz, B., Klepach, T., Zhao, S., Wu, Q., Carmichael, I., and Serianni, A. S. Correlated C-C and C-O bond conformations in saccharide hydroxymethyl groups: parametrization and application of redundant 1H-1H, 13C-1H, and 13C-13C NMR J-couplings. *Journal of the American Chemical Society 126*, 48 (Dec. 2004), 15668–85. (See page 67).

[175] Torrie, G., and Valleau, J. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics 23*, 2 (Feb. 1977), 187–199. (See pages x and 72).

[176] Ueda, K., Imamura, A., and Brady, J. W. Molecular Dynamics Simulation of a Double-Helical β-Carrageenan Hexamer Fragment in Water. *The Journal of Physical Chemistry A 102*, 17 (Apr. 1998), 2749–2758. (See page 102).

[177] Umemura, M., and Yuguchi, Y. Solvation of xyloglucan in water/alcohol systems by molecular dynamics simulation. *Cellulose 16*, 3 (Feb. 2009), 361–371. (See page 102).

[178] Umemura, M., Yuguchi, Y., and Hirotsu, T. Interaction between Cellooligosaccharides in Aqueous Solution from Molecular Dynamics Simulation: Comparison of Cellotetraose, Cellopentaose, and Cellohexaose. *The Journal of Physical Chemistry A 108*, 34 (Aug. 2004), 7063–7070. (See page 102).

[179] Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. C. GROMACS: fast, flexible, and free. *Journal of computational chemistry 26*, 16 (Dec. 2005), 1701–18. (See pages 53, 57, 64 and 94).

[180] van Gunsteren, W., Billeter, S., Eising, A., Huenenberger, P., Krueger, P., Mark, A., Scott, W., and I.G.Tironi. *The GROMOS96 Manual and User Guide*. Zurich, Switzerland, 1996. (See page 102).

[181] Venkatarangan, P., and Hopfinger, A. J. Prediction of ligand-receptor binding thermodynamics by free energy force field three-dimensional quantitative structure-activity relationship analysis: applications to a set of glucose analogue inhibitors of glycogen phosphorylase. *Journal of medicinal chemistry 42*, 12 (June 1999), 2169–79. (See page 102).

[182] Verli, H., and Guimarães, J. A. Molecular dynamics simulation of a decasaccharide fragment of heparin in aqueous solution. *Carbohydrate Research 339*, 2 (Jan. 2004), 281–290. (See page 102).

[183] Vijayalakshmi, K., and Rao, V. R. Theoretical studies on the conformation of aldopyranoses. *Carbohydrate Research 22*, 2 (May 1972), 413–424. (See pages 13, 14, 105, 108, 109, 115 and 116).

[184] Wolfe, S. Gauche effect. Stereochemical consequences of adjacent electron pairs and polar bonds. *Accounts of Chemical Research 5*, 3 (Mar. 1972), 102–111. (See page 17).

[185] Wolfe, S. On the magnitudes and origins of the "anomeric effects", "exo-anomeric effects", "reverse anomeric effects", and C_X and C_Y bond - lengths in XCH2YH molecules. *Carbohydrate Research 69*, 1 (Mar. 1979), 1–26. (See page 17).

[186] Woods, R. J., Dwek, R. A., Edge, C. J., and Fraser-Reid, B. Molecular Mechanical and Molecular Dynamic Simulations of Glycoproteins and Oligosaccharides. 1. GLYCAM_93 Parameter Development. *The Journal of Physical Chemistry 99*, 11 (Mar. 1995), 3832–3846. (See page 102).

[187] XIE, Y., and SOH, A. Investigation of non-covalent association of single-walled carbon nanotube with amylose by molecular dynamics simulation. *Materials Letters 59*, 8-9 (Apr. 2005), 971–975. (See page 102).

[188] Xu, R., McBride, R., Paulson, J. C., Basler, C. F., and Wilson, I. a. Structure, receptor binding, and antigenicity of influenza virus hemagglutinins from the 1957 H2N2 pandemic. *Journal of virology 84*, 4 (Feb. 2010), 1715–21. (See page 16).

[189] Yoshida, Y., Isogai, A., and Tsujii, Y. Structural analysis of polymer-brush-type cellulose β-ketoesters by molecular dynamics simulation. *Cellulose 15*, 5 (May 2008), 651–658. (See page 102).

[190] Yu, H., Amann, M., Hansson, T., Köhler, J., Wich, G., and van Gunsteren, W. F. Effect of methylation on the stability and solvation free energy of amylose and cellulose fragments: a molecular dynamics study. *Carbohydrate research 339*, 10 (July 2004), 1697–709. (See page 102).

[191] Zefirov, N., Palyulin, V., and Dashevskaya, E. Stereochemical studies. XXXIV. Quantitative description of ring puckering via torsional angles. The case of six-membered rings. *Journal of Physical Organic Chemistry 3*, 3 (Mar. 1990), 147–158. (See pages x, 46, 47 and 48).

[192] Zhang, Q., Jaroniec, J., Lee, G., and Marszalek, P. E. Direct detection of inter-residue hydrogen bonds in polysaccharides by single-molecule force spectroscopy. *Angewandte Chemie (International ed. in English) 44*, 18 (Apr. 2005), 2723–7. (See page 16).