Chapter 2

Molecular Analysis and Visualization

Molecular analysis is concerned with gaining insight into the behavior of molecules and molecular systems. Although we could perform molecular analysis completely without any molecular visualization, molecular analysis and visualization have been closely coupled from the early days of molecular modeling. One reason for this lies in the human visual system, which is possibly the best trained sense of the human senses. Everyday the human visual system is faced with the task of distinguishing between more important and less important information. As soon as we open our eyes in the morning, our visual system perceives images of our surrounding and tries to match them with patterns stored in our brain. This process, sometimes called human pattern recognition, can also be described as deriving a probable interpretation from incomplete data. The visual system is surprisingly good at this. Yet, for a deep understanding of the behavior of molecular systems, more is needed than molecular visualization.

> *"It is the interaction between molecular graphics and the underlying theoretical methods that has enhanced the accessibility of molecular methods and assisted the analysis and interpretation of such calculations."* [107]

This thesis is more concerned with molecular analysis than with molecular visualization, yet molecular visualization also plays an important role for at least two reasons. The more obvious one is that molecular visualization has been used throughout the thesis to illustrate the algorithms and their results. The second – less obvious but more important – reason is that a biochemist using the molecular surface alignment algorithm proposed here can already gain much inside into the molecular system, i.e. a set of active ligands, without any further analysis. But even when applying analysis methods, such as *structure activity relationship* (SAR) methods, we cannot apply these methods blindly. There will always be a need for visual inspection and chemical intuition.

In this chapter we present useful concepts of molecular analysis and visualization that will be used in subsequent chapters. We start in Section 2.1 with an overview of molecular representations that are useful both for the analysis of molecular structure and for visualization. In particular, we will concentrate on those molecular representations that

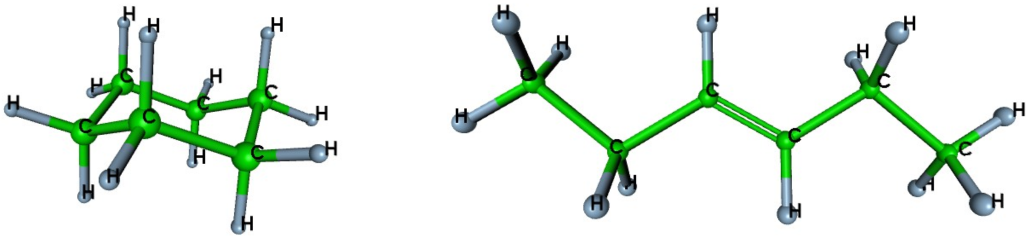


Figure 2.1: *Left:* Cyclohexane. *Right:* 2-hexene.

are commonly used for aligning molecules: atomic representations, molecular surfaces, and molecular fields. In Section 2.2 we take a look at molecular interactions, especially, physico-chemical interactions that play a major role in the binding process of a ligand to a receptor. Physico-chemical properties responsible for these interactions need to be considered in the alignment process in order to yield meaningful alignments that allow to generate a pharmacophore. Section 2.3 gives an introduction to molecular visualization. Here, we describe how the molecular representations introduced in Section 2.1 can be visualized. The first three sections only deal with the static structure of a molecule. But molecules are persistently in motion and constantly change their form. Most often these changes are small, but at certain occasions more drastic changes occur. It is these changes and their corresponding forms that conformation analysis is concerned with. The field of conformation analysis will be the subject to Section 2.4. In the absence of active forms of the chemical compounds of interest, we need to apply conformation analysis to generate feasible forms that can be used for molecular alignment. Recently, much progress has been made in determining metastable conformations, which represent subsets of the configurational space in which a molecule resides for a longer period of time. Metastable conformations and their visualization will also be described in Section 2.4.

2.1 Molecular Representations

The most simple molecular representation is the *chemical formula*. A chemical formula gives precise information about the type and number of atoms constituting a chemical compound. The chemical formula of *methane*, e.g., is CH_4, stating that methane consists of a single carbon (C) atom and four hydrogen (H) atoms. While in the case of methane it is fairly obvious that the four hydrogen atoms are bonded to the carbon atom by single bonds, it is not always that obvious. The chemical formula C_6H_{12}, e.g., denotes *hexene* as well as *cyclohexane*, which are completely different compounds as can be seen in Figure 2.1. In the following sections we give examples of molecular representations that more clearly describe a chemical compound.

2.1.1 Molecular Graph

The relative positions of a molecule's atoms to each other are determined by the interactions of each atom with its surrounding atoms. Some of the atoms within a molecule have particularly strong interactions with each other. These strong interactions are called *covalent bonds*. It is these covalent bonds that hold the molecule together and mainly constrain the molecule in adopting different shapes. Hence, if – in addition to the chemical formula – we know which atoms are covalently bonded, we can uniquely identify the molecule in question. The chemical formula plus the molecule's covalent bonds are often referred to as the molecule's *topology*. However, *"topology is one of the most loosely used words in the contemporary chemical literature and is often confused with graph theory. Topology deals with continuum problems, graph theory with discrete problems"* [126]. Thus, in order to make a clear distinction between topology in the strict mathematical sense and the atom and bonding information of a molecule, we will use the term *molecular graph* instead.

Definition 2.1.1 (Undirected Graph)**.** An *undirected unlabeled graph* G is an ordered pair $G = (V, E)$ of sets V and $E \subset V \times V$, where V is called the set of *vertices*, and E is called the set of *edges*. An edge e of an undirected graph is an unsorted pair of vertices, i.e. $e = (u, v) = (v, u)$, $\forall u, v \in V$. If labels are associated with both the set of vertices and the set of edges we call this graph an *undirected labeled graph* denoted by $G = (V, E, L(V), L(E))$, where $L(V)$ and $L(E)$ denote the set of vertex labels and the set of edge labels, respectively. Note, that the labels do not necessarily have to be unique. For V and E we will sometimes also write $V(G)$ and $E(G)$, respectively.

In this thesis we do not consider directed graphs, hence we will usually use the term graph instead of undirected graph. With the above definition we can define a molecular graph as follows.

Definition 2.1.2 (Molecular Graph)**.** A *molecular graph* of a molecule M is an undirected labeled graph $G = (V, E, L(V), L(E))$, where V represents the set of atoms of M, E the set of covalent bonds, $L(V)$ the atom types, and $L(E)$ the bond order. We denote the molecular graph of molecule M by $G(M)$.

2.1.2 3-Dimensional Molecular Structure

The molecular graph of some chemical compound gives us more information than the chemical formula, but it does not give us the positions of the atoms in 3-dimensional space. There are two common ways to represent the positions of all atoms in a molecule. The first representation uses *Cartesian coordinates*, given by the $x, y,$ and z values for each atom. The second representation uses *internal coordinates*, which are given by the *bond lengths* for every two covalently bonded atoms, the *bond angles* for every three consecutive atoms in the molecular graph, and the *dihedral angles* for every four consecutive atoms in the molecular graph. It is straightforward to convert internal coordinates into Cartesian coordinates and vice versa. For a specific purpose, one of the two types of coordinates is usually preferred. For example, for quantum mechanics calculations, internal coordinates are usually preferred, whereas for molecular mechanics simulations, Cartesian coordinates are favored. Alignment of molecules is also usually done using Cartesian coordinates.

In order to account for the volumetric extension of the molecule, the atoms are often represented by spheres rather than zero-dimensional points. But what radius should be assigned to each atom sphere? Generally accepted is the use of the *van der Waals radius*, which is defined with the help of the *van der Waals contact distance* (cf. Section 2.2.3).

Representing a molecule as a set of connected points or spheres with certain properties is a rather abstract way of looking at a molecule, yet it has proved very helpful and indeed is sufficient for many applications. Nevertheless, it is more realistic to view a molecule as some kind of density.

2.1.3 Electron Density

A molecule consists of atomic nuclei and electrons. A nucleus is positively charged and consists of both protons and neutrons. The only exception to this is hydrogen, whose nucleus consists of only a single proton. The electrons, which are in continuous movement around the nuclei, carry a single negative charge. The *electron density* gives for each position in space the probability that an electron will be present at this position at any time. While the nuclei carry almost all of the molecule's mass, the electrons are responsible for chemical bonding and interactions with the molecule's surrounding. For this reason the electron density is of great interest in molecular modeling. The electron density can be determined experimentally by X-ray diffraction scans or computed, e.g., by *ab initio* or *density functional methods* [107]. While ab initio methods are more exact, they are limited to rather small molecules.

2.2 Physico-Chemical Interactions

In this section we introduce some physico-chemical interactions that are important for the binding of a ligand to a receptor, and, hence, should play an important role in the design of an alignment method used for elucidating the pharmacophore from a set of active compounds.

The binding of a ligand to a receptor is driven by the *change of the Gibbs free energy*, ΔG, of the system. The change of the Gibbs free energy is given by the free-energy function

$$\Delta G = \Delta H - T\Delta S \ ,$$

which was introduced by Josiah Willard Gibbs in 1878 [161]. The free-energy function combines the first and second laws of thermodynamics. The *enthalpic* term, which "represents" the first law of thermodynamics, is given by the *enthalpy change*

$$\Delta H = \Delta E + P\Delta V \ ,$$

which is the sum of the *internal energy change* ΔE, and the product of the *pressure* P and the *change of volume* ΔV, where P is assumed to be constant.

The *entropic* term, "representing" the second law of thermodynamics, is given by the product of the *temperature* T and the *change of the entropy* ΔS, which measures to what extend the degree of randomness or disorder of the system has changed. The *entropy* S increases when the system becomes more disordered, in which case ΔS is positive.

Since the volume change is very small for almost all biochemical reactions, it is generally ignored and we get the following approximation for the free energy change:

$$\Delta G \cong \Delta E - T\Delta S \ .$$

A ligand binds to a receptor due to multiple weak forces, which can be classified as either being enthalpic or entropic. Forces with enthalpic origin are hydrogen bonds, electrostatic interactions, and van der Waals interactions, while hydrophobic effects are of entropic origin [161]. Van der Waals forces can occur across the whole molecule, since they emerge whenever atoms get into a distance of 3 to 4 Å. In contrast to this, the other three forces that were mentioned above generally vary across the whole molecule according to the local molecular structure. All of these forces will be described in the following subsections.

2.2.1 Hydrogen Bonds

A hydrogen bond can be considered as a particularly strong form of a dipole-dipole attraction. Even though its strength is much weaker than that of a covalent bond, hydrogen bonds contribute the largest part to the change of enthalpy. In organic compounds, a hydrogen can participate in hydrogen bonding, if it is bonded to an oxygen or nitrogen atom. These atom pairs, i.e. OH and NH, are strongly polarized, resulting in the hydrogen atom having a partial positive charge. Due to this polarization, the hydrogen atom has a strong affinity to non-bonding electrons found in oxygen and nitrogen atoms [167]. As a result, the hydrogen atom is shared between the two atoms. The atom which the hydrogen atom is more tightly linked to is called the *hydrogen donor*, whereas the second atom participating in the hydrogen bond is called the *hydrogen acceptor*. A hydrogen bond can be considered as an intermediate in the transfer of a proton between an acid and a base [161]. Hydrogen bonds are highly directional, which means that the three atoms forming a hydrogen bond prefer a co-linear orientation. Consequently, hydrogen bonds are strongest if donor, hydrogen and acceptor lie on a single line. In addition to a preferred orientation, hydrogen bonds also have a preferred length which is somewhere in between the van der Waals contact distance and the length of covalent bonds. Hydrogen bonds are not only important for inter-molecular interaction, such as between a receptor and a ligand, but also for intra-molecular interactions, such as in proteins, where they are mainly responsible for building secondary structures like α-helices and β-sheets.

Since hydrogen bonds contribute largely to the free-energy change of the receptor-ligand complex, they are of particular importance for the alignment of molecular surfaces. In many cases, "good" alignments will overlay regions having a similar potential for building hydrogen bonds.

2.2.2 Electrostatic Interactions

Attracting electrostatic forces occur between oppositely charged regions of two molecules, e.g. a receptor and a ligand. The *force* F between two point charges q_1 and q_2 is given by Coulomb's law:

$$F = \frac{q_1 q_2}{\varepsilon_0\, 4\pi\, r^2} \ ,$$

where r is the distance between q_1 and q_2, and ε_0 is the *dielectric constant in vacuum.* If there exists a medium between q_1 and q_2, the force F reduces to

$$F = \frac{q_1 q_2}{\varepsilon_r \varepsilon_0\, 4\pi\, r^2} \ ,$$

where ε_r is the medium's *relative dielectric constant.* In general, *work* is defined as the integral of force over distance. If a point charge q_1 is carried in vacuum towards a point charge q_2 from infinity up to a distance r, the work W needed to achieve this is

$$W = -\int_{\infty}^{r} F\,\mathrm{d}r = -\frac{q_1 q_2}{4\pi\,\varepsilon_0} \int_{\infty}^{r} \frac{1}{r^2}\,\mathrm{d}r = \frac{q_1 q_2}{4\pi\,\varepsilon_0\, r} \ .$$

If this work is carried out, the *potential energy* E_{pot} increases by W, i.e.,

$$\Delta E_{pot} = W = q_1\,\phi \text{ , where}$$

$$\phi = \frac{q_2}{4\pi\,\varepsilon_0\, r}$$

is called the *electrostatic potential* generated by q_2. In a homogeneous medium with relative dielectric constant ε_r, this potential reduces by the factor ε_r. For more information, we refer the interested reader to, e.g., [9].

The electrostatic potential is of particular interest in the study of molecular interactions for two reasons. First, electrostatic forces contribute a similar amount to the free energy change as hydrogen bonds do. This amount of energy is much larger than that of van der Waals interactions. Second, electrostatic interactions are long range forces in contrast to hydrogen bonds and van der Waals interactions, which only have a short range. Hence, electrostatic interactions strongly influence the orientation of the ligand when binding to the receptor.

2.2.3 Van der Waals Interactions

The *van der Waals interaction* is a force, which comes into play when two atoms are a distance of 3 to 4 Å apart [161]. The attraction of two atoms caused by van der Waals interaction is due to the dynamic change of the electronic charge distribution around an atom, which, at any time, is not perfectly symmetric. By this asymmetry, a similar asymmetry is encouraged around neighboring atoms, resulting in an attractive force between the atoms. The attraction becomes stronger as the atoms get closer to each other until the *van der Waals contact distance* is reached. At a shorter distance a strong repulsive force becomes dominant, which is due to the overlapping electron clouds [161] (cf. Figure 2.2).

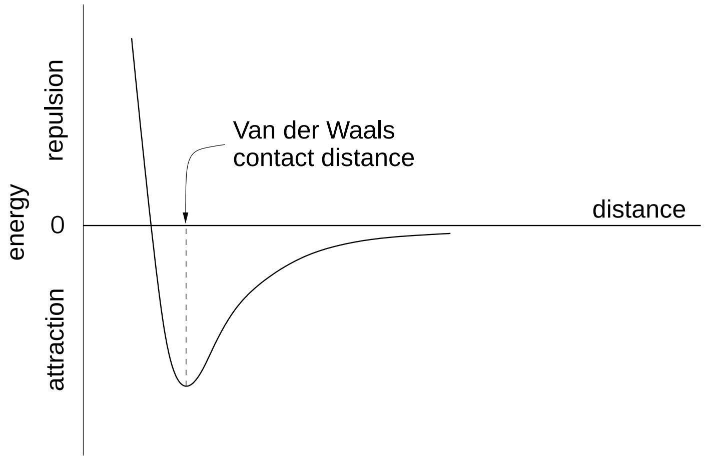


Figure 2.2: Energy of a van der Waals interaction as a function of the distance between two atoms [161].

Based on the concept of the van der Waals contact distance, the van der Waals radius of each atom can be defined. Since the van der Waals contact distance is specific to a particular atom pair, the van der Waals radius needs to be averaged over all chemical elements, including the element equivalent to the atom. By construction, it is clear, that the van der Waals radius is not a specific physical attribute of each atom, but a useful interpretation of atomic extension. With this construct we can now define the *van der Waals sphere* of an atom as the sphere centered at the position of the atom's nucleus with radius equal to the van der Waals radius of the atom.

2.2.4 Hydrophobic Attractions

Each water molecule tends to form hydrogen bonds with its neighboring water molecules. In liquid water, a water molecule forms 3.4 hydrogen bonds on average [161]. If a non-polar molecule is introduced to water, it builds a hydrogen bond cavity, since the water is not able to form hydrogen bonds with the non-polar molecule. Yet, in order to form as many hydrogen bonds as possible, the water molecules have to rearrange around the non-polar molecule. This leads to a more ordered structure of the water molecules around the non-polar molecule than in other regions, resulting in a decrease of entropy. If a second non-polar molecule builds an association with the first, some of the previously more ordered water molecules will be released, since less water molecules are needed to envelop the clustered molecules than to envelop each non-polar molecule separately. This gained freedom of the released water molecules leads to an increase of entropy which is favorable. Hence, non-polar molecules are driven together in water not because of a high affinity for each other, but because clustered non-polar molecules disrupt the hydrogen bond structure less than non-clustered ones.

This effect is commonly termed *hydrophobic attraction*, and it applies not only to completely non-polar molecules, but also to *regions* of non-polarity. Hydrophobic attractions

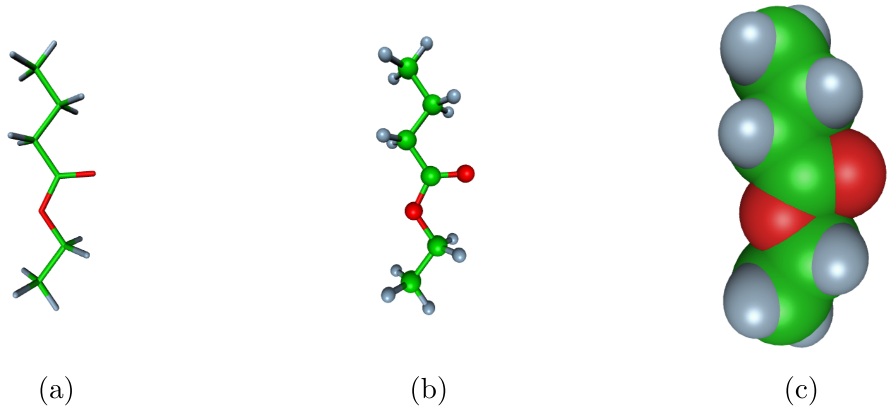


Figure 2.3: (a) Stick model. (b) Ball-and-stick model. (c) CPK model.

are also a major driving force in protein folding, which tends to bury non-polar amino acids in the interior of the protein.

In general, hydrophobic effects play a less important role in the binding than the enthalpic forces introduced earlier in this section. However, in absence of some of these forces, hydrophobic attractions might be of importance.

2.3 Visualization

Throughout this thesis, many visualization techniques will be applied to illustrate the work that has been done. The following sections give a short introduction to these techniques, all of which have been implemented in the visualization and analysis tool Amira [2] and its molecular extension AmiraMol [3].

2.3.1 3-Dimensional Molecular Structure

For the visualization of the molecule's 3-dimensional structure, three representations are commonly used. The first only displays the covalent bonds of a molecule by either lines (wire-frame model) or cylinders (stick model, cf. Figure 2.3(a)). The lines or cylinders can be colored according to properties, such as atomic number, charge, etc., of the atoms they connect. The second representation, known as ball-and-stick model (cf. Figure 2.3(b)), displays both cylinders representing the covalent bonds and balls representing the atoms. The third representation displays only the atoms of a molecule. This model is known as the space-filling or CPK model (cf. Figure 2.3(c)), named after the Corey-Pauling-Koltun (CPK) mechanical models. The first space-filling models date back to Stuart in 1927 [162].

For proteins it is often desired not to display all atoms of the molecule, but to use a more abstract representation by displaying the molecule's secondary structures (α-helices and β-sheets) together with a tube connecting the backbone atoms not belonging to the α-helices and β-sheets. The secondary structures can be displayed using, e.g., cartoons or ribbons approximating the backbone atoms.

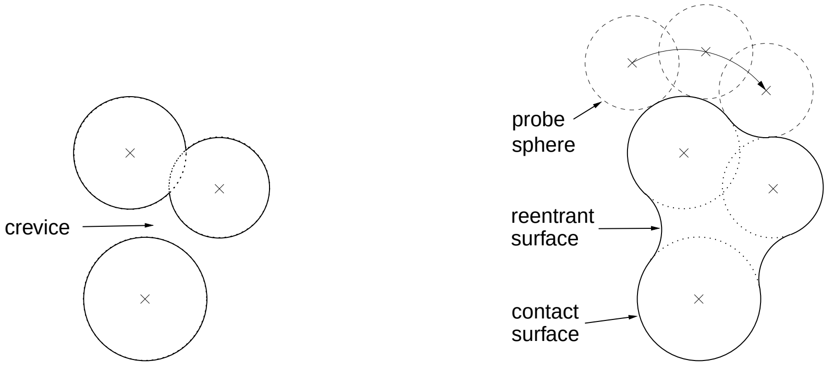


Figure 2.4: *Left:* Van der Waals surface. *Right:* Solvent excluded surface.

2.3.2 Solvent Excluded Surface

Non-covalent interactions between molecules occur where the van der Waals spheres of the molecules get close to each other. Therefore, the separation surface between the van der Waals spheres of a molecule and its surrounding is of particular interest for the analysis of molecular interactions. The surface enclosing all van der Waals spheres of a molecule is known as the *van der Waals surface* (cf. Figure 2.4, left). While it is clear that this surface separates the van der Waals spheres from their surrounding, there may be many parts of this surface which are not accessible to other molecules, not even the smallest molecules of interest – water molecules. This is due to the fact that the van der Waals surface builds many thin crevices which cannot be reached by other molecules (cf. Figure 2.4, left). Hence, the van der Waals surface is not very appropriate for the study of molecular interactions.

In 1977, Richards gave a definition of a molecular surface suitable for the study of molecular interactions [148]. He defined the molecular surface with the help of a probe sphere approximating a water molecule. According to him the molecular surface is defined as a smooth envelope partially touching the van der Waals surface as the probe sphere is rolled over the van der Waals spheres of the molecule. This surface divides the 3-dimensional space into two parts, one of which is accessible to water molecules, the other one being that part which water molecules are excluded from. Therefore, this surface is both named *solvent accessible surface* and *solvent excluded surface*. In this thesis we will use the term solvent excluded surface (SES), or, where it is clear from context, molecular surface only. According to Richards [148], two parts of the SES can be distinguished: the van der Waals *contact surface* and the *reentrant surface* (cf. Figure 2.4, right). As the name suggests, the former is part of the van der Waals surface and it includes all those parts of the van der Waals surface which the probe sphere can be in direct contact with. The latter part, the reentrant surface, is formed by the inward facing parts of the probe sphere while simultaneously being in contact with two or more van der Waals spheres. The reentrant surface decomposes into two different surface patch types: the *toroidal* or *saddle-shaped* patches, which are generated while the probe sphere rolls over two atoms,

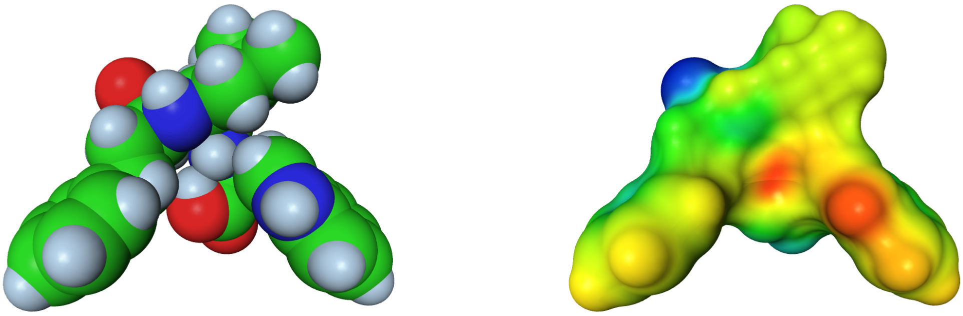


Figure 2.5: Thermolysin inhibitor from PDB-entry 1TMN. *Left:* Van der Waals surface with coloring by atomic number. *Right:* Solvent excluded surface with electrostatic potential.

and the *spherical reentrant* patches, which are generated when the probe sphere rests on three atoms simultaneously.

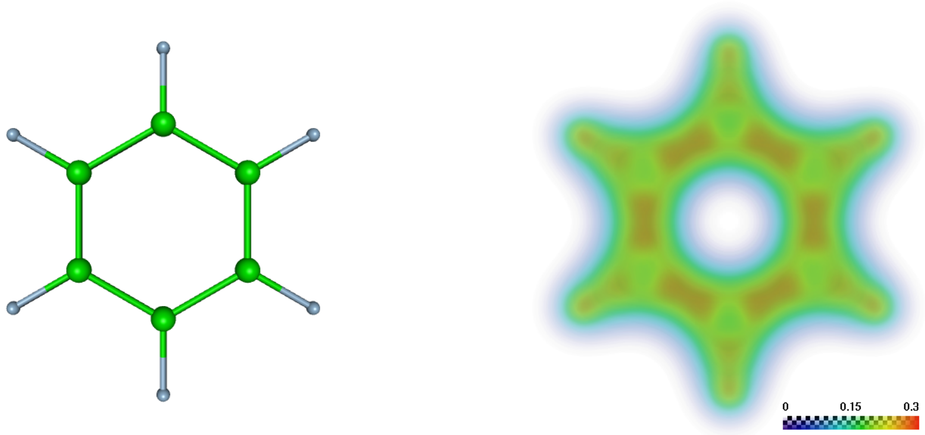
From the above description it follows that the solvent excluded surface is made up of parts of different tori and parts of different spheres – van der Waals spheres and the probe sphere at different positions. Hence, we can analytically describe the SES by the parameters of a set of objects, namely spheres and tori, and circular arcs determining which parts of the objects' surfaces belong to the SES. This analytical surface description can be efficiently computed using the *contour-buildup algorithm* [164] proposed by Totrov and Abagyan. In order to display the SES, however, one needs to compute a triangular mesh from the analytical surface description [17]. The triangles of the mesh can be colored according to certain properties, such as atomic number, charge, etc., of the atoms they are assigned to, but we can also map other molecular properties, such as electrostatic potential or hydrophobicity, onto the triangle mesh by color-coding the triangles according to some scalar property. Figure 2.5 shows the comparison of the van der Waals surface and the solvent excluded surface of the same molecule.

2.3.3 Volume Visualization

The representation of molecules by hard spheres or surfaces is a very abstract yet useful simplification. More appropriately, however, molecules are represented by their electron densities. There exist several methods for displaying molecular densities. Two of these methods, the most common ones, shall be described in the following paragraphs, namely *direct volume rendering* and *isodensity surfaces.*

Direct Volume Rendering

As the name suggests, *direct volume rendering* is a technique for *directly* visualizing a volume given by scalar values on a 3-dimensional grid. In direct volume rendering, the scalar field, e.g. the electron density of some molecule, is considered as a luminous object. At each point in space the scalar value, e.g. the probability of an electron being present, is considered as the amount of light emitted from this point. The 3-dimensional density



(a) (b)

Figure 2.6: (a) Ball-and-stick model of benzene. (b) Direct volume rendering of benzene's electron density. Each point in space emits light of a certain color and intensity according to its scalar value.

is projected onto the 2-dimensional screen by accumulating the emitted light back to front in the direction opposite to the view direction. Apart from emitting light, each point also absorbs light according to its scalar value. Direct volume rendering can be efficiently implemented using a stack of 2D textures, or 3D textures, which have become readily available in the recent past. An example is shown in Figure 2.6. For an excellent introductory text to volume rendering, see the article by Hege et al. [77].

Isodensity Surface

A second technique for visualizing volumes or densities are *isodensity surfaces*. An isodensity surface displays all points where some constant density value, the *isodensity value*, is present. In contrast to direct volume rendering, isodensity surfaces do not visualize the whole volume but focus on some specific part of the volume. Since we are more used to looking at opaque objects rather than transparent ones, isodensity surfaces allow for an easier perception of the "shape" of the volume. Isodensity surfaces are not necessarily connected but often decompose into several closed surfaces enclosing minima or maxima of the volume. Three isodensity surfaces of the same scalar field with varying isodensity values are shown in Figure 2.7.

2.4 Conformation Analysis

A molecule is not a static but a dynamic object. At any time it is in motion, whereby it changes both its global position and its geometry. Most of the time the changes in geometry will be small, if the considered time span is small too. However, at some time the geometry can change more drastically, which might be a spontaneous action but it might also be driven by the interaction with other molecules. In the following definition

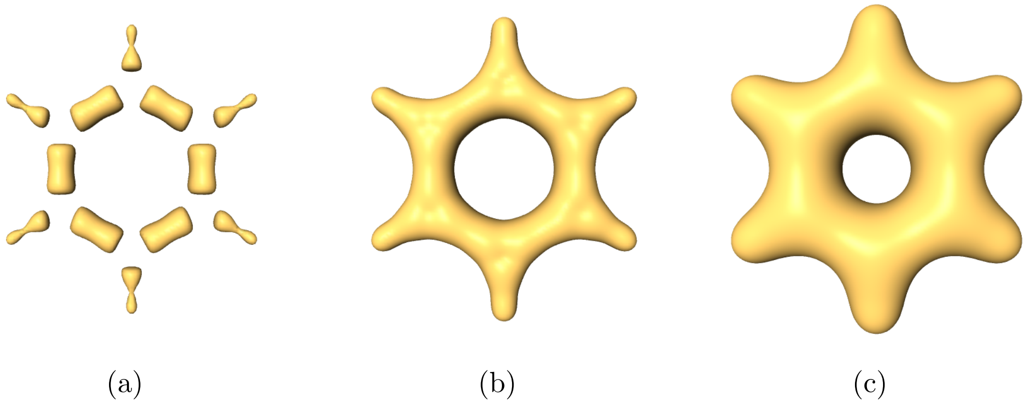


Figure 2.7: Isodensity surfaces of the electron density of benzene, calculated with an ab initio method. The isodensity surfaces are displayed with different isodensity values: (a) 0.25, (b) 0.2, and (c) 0.1. Data courtesy of Bernd Kallies (ZIB).

we specify both what we mean by the "geometry of a molecule" and the corresponding space the geometry resides in.

Definition 2.4.1 (Conformer and Configurational Space)**.** Let M be a molecule defined by its molecular graph $G(M) = (V, E, L(V), L(E))$ (cf. Definition 2.1.2).
We define a *conformer* as a geometrical instance of M given by its internal coordinates. The subset $\Omega \subset \mathbb{R}^{3n-6}$ of all conformers is called *configurational space*, where $n = |V|$ is the number of atoms.

Of most interest to us are those conformers with which molecule M can bind to the active site of some receptor R.

Definition 2.4.2 (Active Conformer)**.** We call a conformer $q \in \Omega$ an *active conformer* of some receptor R, if M is able to bind to the active site of R with the geometry given by q.

If the active conformer of molecule M is not known, we need to consider all feasible conformers of M. This is not possible, since the number of conformers is infinite. Therefore, in general, only conformers with a local minimum on the potential energy surface are used. Beusen et al. justify this approach by arguing, that *"while interaction* [of the ligand] *with the receptor will certainly perturb the conformational energy surface of a flexible ligand, high affinity would suggest that the ligand binds in a conformation*[1] *that is not exceptionally different from one of its low-energy minima."* [24] That is to say, if large geometrical changes from a local minimum are needed, the ligand will probably not have a high affinity to the receptor. Hence, the closer an active conformer is to a local energy minimum, the more stable the receptor-ligand complex will be. But the stability of a receptor-ligand complex is also influenced by the *stability* of the active conformer.

[1] In the literature, the word conformation is sometimes used to denote a specific geometrical instance of some molecule. We will always use conformer instead and reserve the term conformation for subsets of the configurational space, namely metastable conformations.

This leads us to the concept of *metastable conformations*, which will be described in Section 2.4.2. Before we look at the concept of metastable conformations, however, we will first look at different methods for sampling the configurational space.

2.4.1 Exploring Configurational Space

There exists a large number of methods for exploring the configurational space of some molecule. A good overview can be found in [107]. Some of these methods shall be shortly explained in the following paragraphs.

As mentioned in Section 2.1.2, the 3-dimensional structure of a molecule can be described using Cartesian coordinates of its atoms as well as internal coordinates, consisting of bond lengths, bond angles, and dihedral angles. Depending on the search method, different representations of the molecule are preferred.

Random Search Strategies

The most simple approach to exploring configurational space is by applying a random strategy. Random methods can be distinguished according to the coordinates they work on. Cartesian methods, such as [153], work on the Cartesian coordinates of the atoms and randomly perturb the x, y and z coordinates of each atom. After modification of all atomic coordinates, energy minimization is applied to the modified structure, keeping the minimum as new conformer, if its energy is small enough. Otherwise, the conformer will be rejected.

When applying a random strategy, it is often favorable to work on the internal coordinates [33] of the molecules, since these are fewer, especially if the search is restricted to dihedral angles, keeping bond lengths and angles fixed. Problematic for this approach are rings, which are usually broken for the randomization step. After performing a random modification of the dihedral angles, for the previously broken dihedral angles it needs to be checked whether their values are in the allowed range. Again, the randomized structure is minimized and the energy minimum is kept as new conformer. While the random search strategy is very simple in nature, there is one problem inherent to this approach. Since the configurational space is not sampled systematically, it is not clear when to stop, i.e. when the space has been sampled long enough such that all feasible areas of the configurational space have been visited once. Usually, sampling is carried out as long as no new conformer can be found, which results in many conformers being generated multiple times. This leads to long run times.

Systematic Search Strategies

Systematic methods [107, 25] are similar to random methods in that they first generate initial geometries which are then energy-minimized. However, they differ from random methods in the generation of the initial structures, which, as the name suggests, are generated in a systematic way. Systematic methods work on the molecule's internal coordinates, whereby, in general, only dihedral angles are considered. Systematic strategies are problematic because of combinatorial explosion. For example, if we consider a molecule with five dihedral angles and vary each dihedral angle with an increment of 30°, $12^5 = 248832$ initial structures are generated, and, hence, need to be minimized.

We can think of all initial structures as the leaves of a search tree with depth equal to the number of dihedral angles. Each level represents one dihedral angle. The number of children of each internal node depends on the increment used for that particular dihedral. In general, the search tree can be pruned drastically, because when traversing the tree from the root to the leaves, it often becomes apparent at an early stage that the path leads to structures violating some energetic or geometric criterion.

Distance Geometry

Crippen [36] applied distance geometry to sampling the configurational space. In distance geometry, the molecule is neither represented by Cartesian coordinates nor by internal coordinates, but by the pairwise distances of all atoms. These distances are stored in an $n \times n$-matrix, where n is the number of atoms. The entries (i, j) and (j, i) are identical and contain the distance between the atoms i and j, hence the matrix is symmetric. Instead of randomly perturbing the Cartesian or internal coordinates of the molecule, as in the random search, in distance geometry the inter-atomic distances are randomly modified. However, lower and upper bounds of many inter-atomic distances can be determined from simple chemical principles [107]. These lower and upper bounds are used to generate only geometrically feasible conformers which are then energy-minimized to arrive at low-energy conformers.

Molecular Dynamics Simulations

Another way of exploring configurational space is by simulating the dynamics of a molecule using classical mechanics. Molecular dynamics (MD) simulations generate conformers of the system by computing the motion of the molecule's atoms according to Newton's equations of motion. That is to say, changes in geometry of consecutively generated conformers will be very small. This is in contrast to the sampling methods considered earlier, where the differences in geometry could be arbitrarily large between two consecutive conformers.

MD simulations start with the initialization of the system, which is done by selecting initial positions and initial velocities of the molecule's atoms. During an MD run, the positions and velocities are propagated according to internal and external forces acting on the atoms. These forces need to be computed at the beginning of each propagation step. The way the forces are modified depends on the *molecular force field* used. A good introduction to molecular force fields can be found, e.g., in [107].

In contrast to macroscopic classical mechanics, in thermodynamical applications for microscopic systems *"the aim of an MD simulation is* not *to predict precisely what will happen to a system that has been prepared in a precisely known initial condition: we are always interested in statistical predictions."* [59] The reason for using MD simulations instead of the aforementioned methods for sampling the configurational space, is that in addition to the feasible conformers of the molecule, we are also interested in the correct distribution of the conformers. However, even though MD simulations are used to make statistical predictions, it is not guaranteed, that MD runs generate a thermodynamically correct distribution, the *Boltzmann distribution* [7]. It is for this reason, that MD simulations have been coupled with *Monte Carlo Simulations*, which will be shortly described next.

Monte Carlo Simulations

Monte Carlo (MC) methods constitute a stochastic approach to exploring configurational space. They consist of two steps, the *proposal step* and the *acceptance step.* In the proposal step, a more or less random change of the atom positions is proposed. The proposed new conformer is then passed to the acceptance step, which decides whether the new conformer is accepted according to some probability criterion derived from the *Boltzmann distribution.* For details see, e.g., [59]. While MC methods guarantee a correct generation of the Boltzmann distribution, it is difficult to generate proposals that are accepted with a high probability. To overcome this problem, MC can be coupled with MD simulations, yielding so called *Hybrid Monte Carlo* (HMC) methods.

Hybrid Monte Carlo is a method that tries to combine the advantages of both MC methods and MD simulations. The disadvantage of MC methods is that the acceptance rate might be rather poor. Here, MD simulations come into play. They allow to generate trial moves with a much higher acceptance rate. In HMC, the Monte Carlo step does not change the atom positions, but only the initial velocities for MD. These velocities are generated according to the Boltzmann distribution of the *momenta.* Then, a short MD run is carried out and the last conformer of this run is passed to the acceptance step. As in a common MC method, this conformer is accepted according to the Boltzmann distribution of the *positions.*

2.4.2 Metastable Molecular Conformations

The potential energy of some conformer $q \in \Omega$ does neither tell us anything about the probability that the molecule will adopt this conformer or a conformer close to q, nor about the probability that the molecule will leave this conformation. Although it is more likely that the molecule will adopt some low-energy conformer, it depends on the potential energy surface, how long the molecule will stay close to some conformer. This leads us to the concept of *metastable molecular conformations* introduced by Deuflhard and Schütte et al. [43, 44].

Definition 2.4.3 (Metastable Molecular Conformation)**.** A subset $C \subset \Omega$ of the configurational space $\Omega \subset \mathbb{R}^{3n-6}$ of some molecule M is called *metastable molecular conformation*, if the probability of M to stay in C is almost 1. Since, in general, there will be transitions between C and conformations $\tilde{C} \subset \Omega$, this probability is smaller than 1, and, hence, C is called *metastable* molecular conformation.

For simplicity we shall use the term “metastable conformation” instead of the longer term “metastable molecular conformation”.

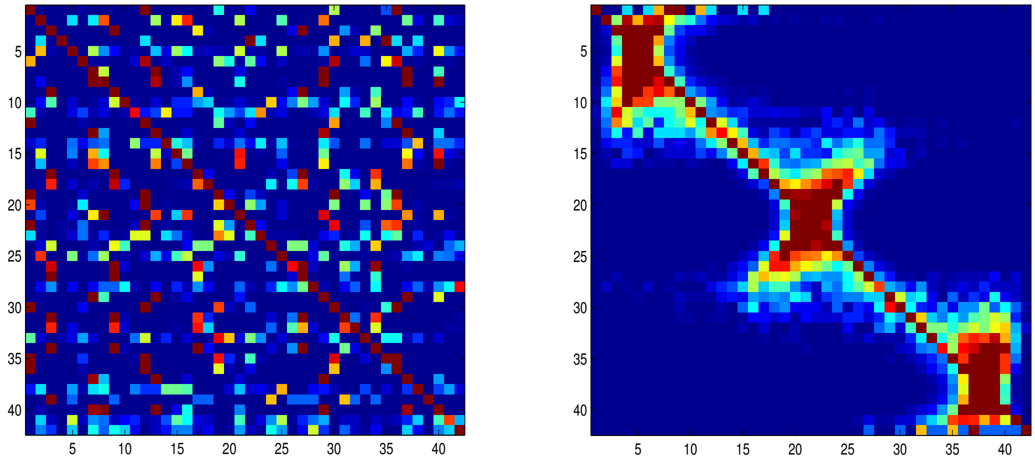


Figure 2.8: Transition matrices of butane. The left image shows the unsorted transition matrix. In this image the block-diagonal matrix is not visible. The right image depicts the sorted transition matrix according to the metastability analysis. It can clearly be seen that butane has three metastable conformations. Image courtesy of Susanna Kube (ZIB).

Identification of Metastable Conformations

Metastable conformations can be identified by means of a transition matrix $T = (T_{ij})$, see, e.g., [42]. To construct the transition matrix, the configurational space $\Omega \subset \mathbb{R}^{3n-6}$ is first discretized in terms of boxes $B_1, \ldots, B_N$. Now, given a realization, $q_1, \ldots, q_m$, of some *reversible Markov chain*, obtained, e.g., by an HMC run, the entries of the transition matrix are given by

$$T_{ij} = \frac{\#\{q_{k+1} \in B_j \wedge q_k \in B_i\}}{\#\{q_k \in B_i\}} \; .$$

If the Markov chain represents metastable conformations, the transition matrix T has an almost block diagonal structure, which might be hidden (cf. Figure 2.8, left). Since the Markov chain is reversible, the transition matrix is generalized symmetric. In order to determine the hidden almost block diagonal structure of the transition matrix, Deuflhard et al. applied *Perron Cluster Analysis* (PCCA) [44]. PCCA not only identifies the metastable conformations but also gives their life times and decay patterns [42]. Recently, Deuflhard and Weber [45, 170] improved PCCA, which led to the *Robust Perron Cluster Analysis* (PCCA+). Here, the characteristic functions are replaced by almost characteristic functions, which can be written as linear combinations of the dominant eigenvectors of T.

2.4.3 Visualization of Metastable Conformations

The metastable molecular conformations visualized in this section were computed with the program ConfJump [169]. ConfJump used a pre-discretization generated with the program ConFlow [125]. Both programs were developed at the Zuse Institute Berlin.

The Alignment Problem

A crucial difficulty for the visualization of molecular conformations lies in the choice of Cartesian coordinates. All geometric comparisons and distance measurements between molecular geometries, as they are performed during the computation of metastable molecular conformations, rely on internal coordinates, i.e. bond lengths, bond angles, and dihedral angles. In other words, molecular conformations are independent of any rigid transformations that are previously applied to single geometries. In contrast to this, the visualization of molecular conformations takes place in Cartesian coordinates. Therefore it is necessary to assign global positions and orientations to the geometries and thereby to define a relative alignment between all geometries of a metastable molecular conformation. The determination of the optimal relative alignments of all geometries to each other shall be described in this section. Except for the last paragraph, this section is the summary of an earlier publication [155].

Let M be a molecule with n atoms and let $\Omega \subset \mathbb{R}^{3n-6}$ be the configurational space of M. For some geometry $q_i \in \Omega$, let $\mathbf{x}(q_{i1}), \ldots, \mathbf{x}(q_{in})$ denote the Cartesian coordinates of the atomic nuclei of molecule M, as they are given by, e.g., an HMC run. We define the distance between two geometries q_1 and q_2 as the sum of squared distances between corresponding atoms

$$d_{\text{Cart}}(q_1, q_2) := \sum_{i=1}^{n} (\mathbf{x}(q_{1i}) - \mathbf{x}(q_{2i}))^2 \ ,$$

and the optimal rigid transformation T_{opt} as the transformation minimizing the distance between q_1 and q_2, i.e.

$$T_{\text{opt}} := \arg\min_{T} d_{\text{Cart}}(q_1, T(q_2)) = \arg\min_{T} \sum_{i=1}^{n} (\mathbf{x}(q_{1i}) - T(\mathbf{x}(q_{2i})))^2 \ .$$

According to Kabsch [88, 89], T_{opt} can be computed in two steps. In the optimal alignment of q_1 and q_2, the barycenters of the coordinates of q_1 and q_2 coincide. Thus, the first step is to translate q_2 such that the barycenters of q_1 and q_2 coincide. The remaining optimization problem merely deals with rotations around the common barycenter. The optimal rotation for q_2 can be found via solving the eigenproblem of the symmetric, positive definite 3×3 matrix $\mathcal{M}^T \mathcal{M}$, where

$$\mathcal{M} = \left(\sum_{i=1}^{n} \mathbf{x}(q_{1i}) \otimes \mathbf{x}(q_{2i}) \right) \ .$$

Here $(x \otimes y)_{st} \equiv x_s y_t$, where x_s and y_t denote components of the spatial coordinate vectors x and y, respectively. For more details see [88, 89, 155].

In the case of a metastable conformation $C \subset \Omega$, we are not interested in the alignment of one geometry to another one, but in the alignment of all geometries to each other. Let $q_1, \ldots, q_m \in C$ be the geometries to be aligned to each other. Then, we want to identify

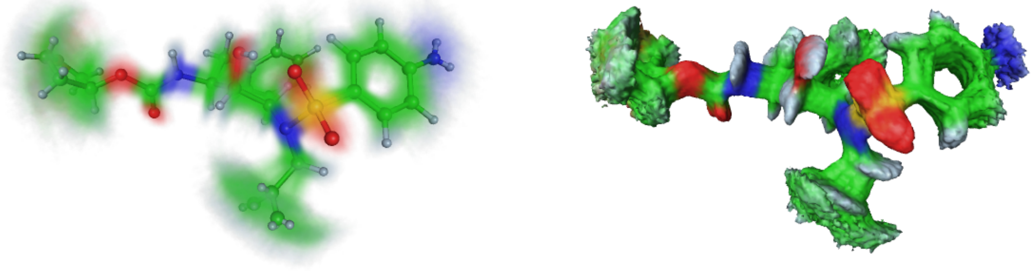


Figure 2.9: Conformational density of one metastable conformation of amprenavir, displayed using volume rendering (*left*) and an isodensity surface (*right*). The density was computed using the bond cylinder primitive only. In the left image the mean geometry is displayed in addition to the density. The coloring is according to the atom types.

rigid transformations $T_1, \dots, T_m$ such that the objective function

$$O = \sum_{1 \le i < j \le m} d_{\text{Cart}}(T_i(q_i), T_j(q_j))$$

is minimized. The rigid transformations $T_1, \dots, T_m$ minimizing O can also be determined in two steps. Again all geometries are translated such that their barycenters coincide. The remaining objective function only depends on rotations of the geometries around the common barycenter. The optimal rotations can be computed by an iterative scheme, where every iteration optimizes the rotation of one of the geometries by solving an eigenproblem analogous to the case of alignment of two geometries [138]. Effectively, this step aligns each geometry to the mean of all other geometries. To get a reasonable starting point for the iteration, we can align all geometries to the first geometry.

Based on the described alignment strategy, we can compute representative geometries as well as geometry densities, whereby different geometries, such as balls and sticks, or molecular surfaces, can be used.

Representatives of Metastable Conformations

The geometry of a metastable conformation C can be understood as a “fuzzy” molecular shape. One way of visualizing this shape is to depict some representative geometry which can be interpreted as a shape around which the geometries of C are distributed. Representatives can also be used to compute an alignment between two metastable conformations.

An approximation of such a representative shape can be generated by first computing an alignment as described in the previous paragraph and then averaging the Cartesian coordinates of each atom over all geometries. This procedure yields mean coordinates for all atoms that generally do not define a realistic molecular shape. For example, parts of the molecule that are very flexible, especially parts performing large rotations, will be distorted in the mean geometry. To get a more realistic representative shape, we can

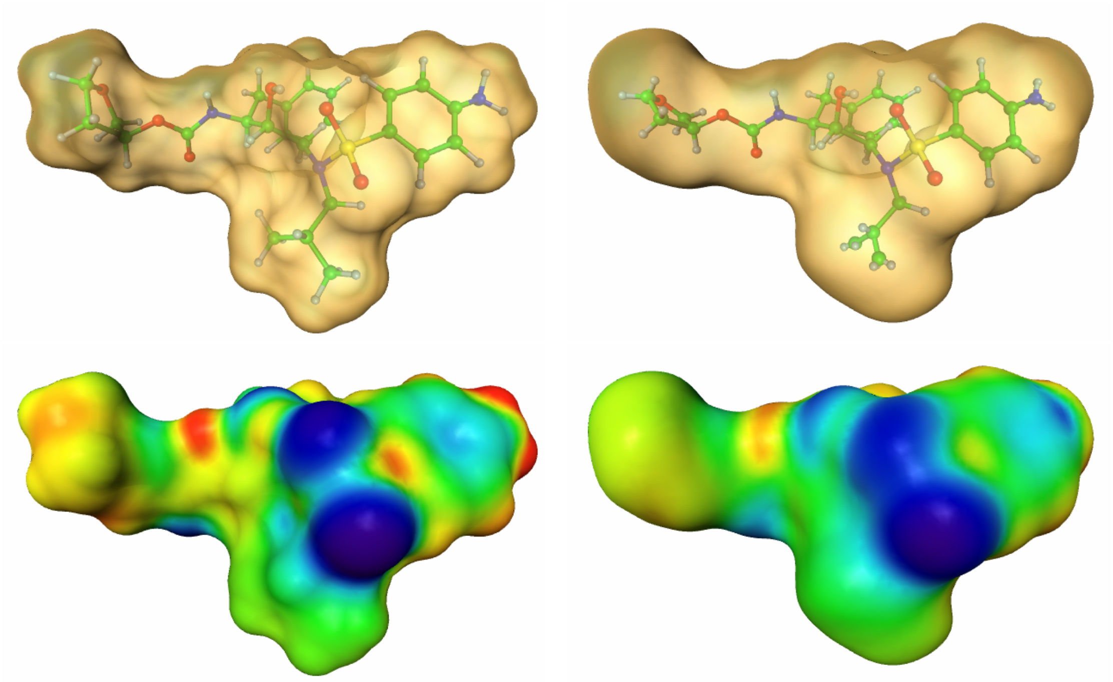


Figure 2.10: Comparison of molecular surface of representative geometry (*left column*) with isodensity surface of conformational density (*right column*). The density was computed from the solvent excluded surfaces of all geometries. The ball-and-stick representations show the representative and the mean geometry, respectively. *Bottom row:* Electrostatic potential mapped onto the surfaces.

search for the geometry $q \in C$ with the smallest distance to the mean geometry. Mean geometry and representative are compared in Figure 2.10.

Conformational Density Based on Molecular Skeleton

Although the representative of a molecular conformation is an important means for understanding the essential properties of a conformation, it is obviously of limited value. The information about the flexibility of the conformation is lost. What is needed is some kind of probability density of the location of the molecule. One possibility is to accumulate the density of skeletal primitives abstracting the molecule, namely cylinders. We call such a density *conformational density*, since we accumulate the density over all geometries of a metastable conformation.

In order to filter out the noise due to arbitrary transformations of the geometries of the metastable conformation, we apply the alignment procedure described above. Once all geometries have been aligned, we compute the bounding box of all transformed geometries, discretize the bounding box and compute a 3-dimensional histogram over all aligned geometries of the metastable conformation w.r.t. the primitives representing each geometry. Rheingans et al. described a technique for the visualization of molecules with positional uncertainty in [147]. However, they only consider spheres as volume filling primitives. We found that using spheres only, it is hard to match certain parts of the density to the molecular structure. In order to allow a visual investigation of the internal structure of the conformation, we have extended the computation of positional uncertainty to a second primitive, the bond cylinders. Furthermore, Rheingans et al. only compute the scalar field for the density. We have extended the method to the computation of color fields in addition to the scalar field. Details of the described approach can be found in [155]. For one metastable conformation of the HIV-1 protease inhibitor amprenavir, the conformational density based on the bond cylinders of the molecule is displayed in Figure 2.9 using both volume rendering and an isodensity surface.

Conformational Density Based on Molecular Surfaces

Even though for the visual inspection of conformational densities it is favorable to use cylinder primitives, for other purposes volume filling representations, such as the solvent excluded surface, might be more useful. For example, if we want to apply the surface alignment approach described in this thesis to metastable conformations, we could use isodensity surfaces of the conformational densities. However, as can be seen in Figure 2.9, right, isodensity surfaces of conformational densities based on cylinders are skeleton-like, as would be expected. Thus, for the purpose of aligning metastable conformations, we extended the method described in [155] to molecular surfaces. To do so, for each geometry we compute its molecular surface, scan-convert [94] it and accumulate the obtained voxel representations on a 3-dimensional grid similar to the approach above. On the isodensity surface of such a conformational density, we can also compute the electrostatic potential by averaging it over all geometries of the metastable conformation. A comparison between the representative of one metastable conformation and its mean geometry is shown in Figure 2.10. While the representative is depicted with its solvent excluded surface, the mean geometry is shown together with an isodensity surface of the conformational density based on the solvent excluded surface.