University
of Antwerp

Chemo-informatics and computational drug design

Prof. Dr. Hans De Winter

University of Antwerp
Campus Drie Eiken, Building A
Universiteitsplein 1, 2610 Wilrijk, Belgium

Gefinancierd door
de Europese Unie
NextGenerationEU

Chapter 6. Molecular mechanics and conformational analysis

1. Force fields

A force field refers to the functional form and parameter sets used to calculate the potential energy of a molecule. It is nothing more than a set of functions that take as input the coordinates of the atoms, and returns an energy value out of these. *All-atom* force fields provide parameters for every type of atom in a system, including hydrogen, while *united-atom* force fields treat the hydrogen and carbon atoms in each methyl group and each methylene bridge as one interaction center, thereby reducing the number of particles in the calculation with a significant calculation speedup as result.

1.1. Functional form of a force field

The basic functional form of potential energy in molecular mechanics includes 1) bonded terms for interactions of atoms that are linked by covalent bonds, and 2) nonbonded terms that describe the long-range electrostatic and van der Waals forces between atoms that are not bonded by covalent bonds. The specific decomposition of the terms depends on the force field, but a general form for the total energy in an additive force field can be written as:

$$E_{total} = E_{bonded} + E_{nonbonded}$$

and in which the bonded and nonbonded terms can be further defined as:

$$E_{bonded} = E_{bonds} + E_{angles} + E_{dihedrals}$$

$$E_{nonbonded} = E_{electrostatic} + E_{vdw}$$

with E_{bonds}, E_{angles} and $E_{dihedrals}$ describing the energy contributions of all covalent bonds, angles and dihedral angles, respectively, and $E_{electrostatic}$ and E_{vdw} the terms describing the nonbonded interactions between atoms due to the atomic partial electrostatic charges and the van der Waals interactions. In the following sections, each of these terms are described in more detail.

1.2. Bond potential

In standard force fields, the contribution by covalent bonds to the total relative energy is given by a Hooke's law in which bonds are treated as springs:

$$E_{bonds} = \sum_{all\ bonds} k(b - b_o)^2$$

with *k* being a bond force constant, *b* the actual bond length for a given bond, and b_0 the reference length. For example, in the case of a C-C bond, *k* could be 100 kcal/mol/Å^2 and b_0 is 1.54 Å. Hence, when the bond is at its reference value of 1.54 Å, then the relative energy contribution equals 0.0 kcal/mol, but when the bond is stretched or compressed by 0.1 Å, then the energy rises by 1.0 kcal/mol.

The force constant *k* and reference bond length b_0 depends on the actual bond type (single, double, triple) and on the constituting atoms (Figure 49).

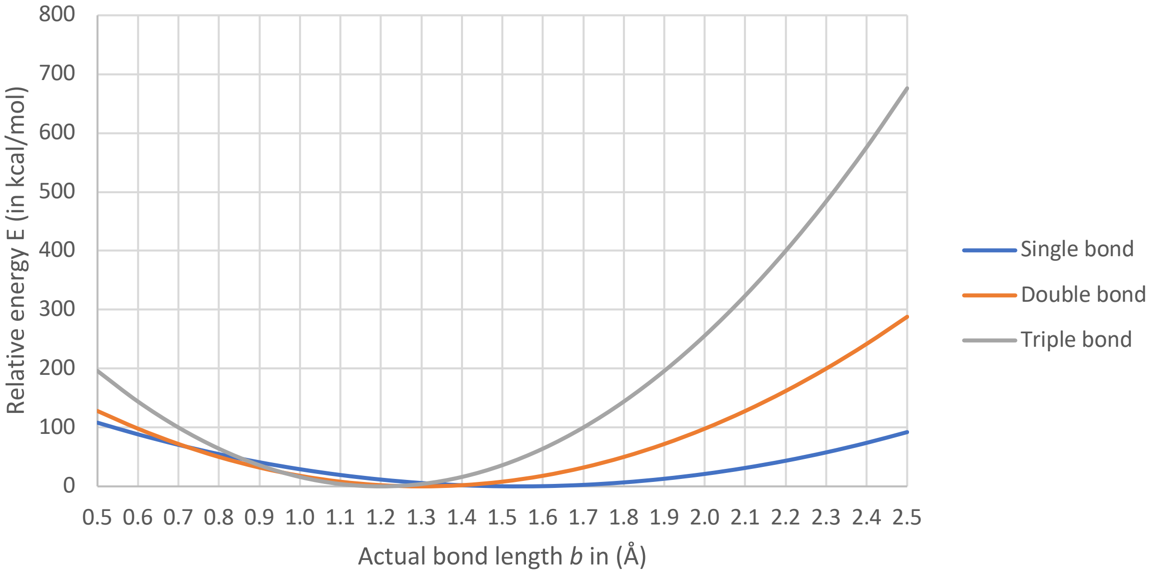


Figure 49. Bond potential illustrated for a single, double and triple C-C bond. Force field parameters are 1.54, 1.3 and 1.2 Å for b_0, and 100, 200 and 400 kcal/mol/Å^2 for k, respectively. The triple bond is more rigid than the single bond, as the amount of compression or stretching leads to a higher increase in relative energy.

1.3. Angle potential

The contribution of angles to the total energy of a molecule is similar to that from bonds:

$$E_{angles} = \sum_{all\ angles} k(\theta - \theta_0)^2$$

which sums over all angles and in which *k* being the angle force constant, θ the actual angle for a given bond angle, and θ_0 the reference angle. Again, the actual values for *k* and θ_0 depend on the constricting atoms that make up the bond angle.

1.4. Dihedral angle potential

A dihedral angle is an angle that is defined by four points, *in casu* four atoms. It is the angle between the planes that are defined by two sets of three atoms each, having two atoms in common (Figure 50):

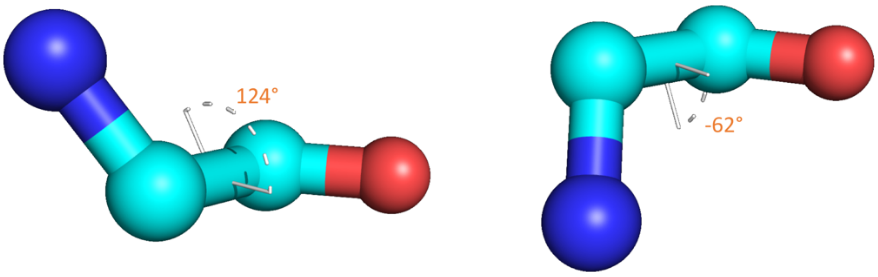


Figure 50. The dihedral angle for the N (blue) – C (cyan) – C (cyan) – O (red) sequence is defined as the angle between the plane through N-C-C and the plane through C-C-O.

Dihedral angle potentials are a bit more complex than bond or angle potentials, since the dihedral angle potential is defined as a periodic function with an optional phase shift:

$$E_{dihedrals} = \sum_{all\ dihedrals} k(1 + \cos(n\phi - \delta))$$

in which *n* is an integer ranging from 1, 2 or 3, ϕ being the actual dihedral angle and δ the phase shift (normally this is a value of 0° or 180°) (Figure 51).

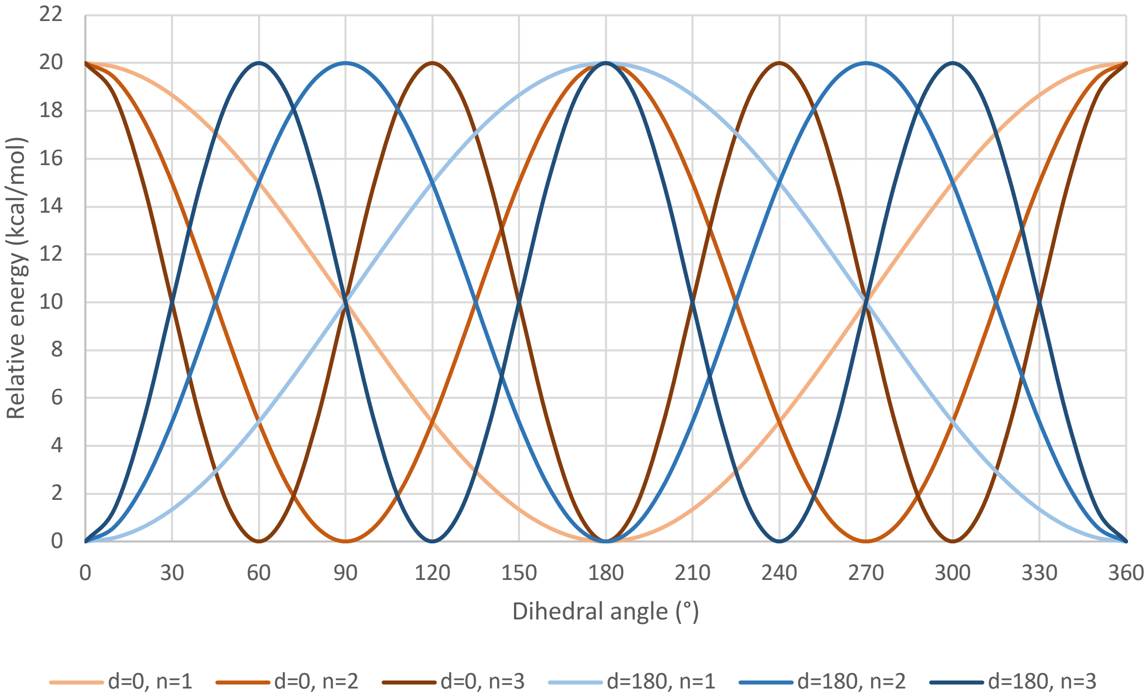


Figure 51. Illustration of a dihedral potential and its dependence on the force field parameters. In this particular example, k was set to 10 kcal/mol, and both δ (shown as d in the legend) and n were varied as indicated. A phase shift δ of 0° is often the most appropriate, since this gives energy minima near 60, 90 and/or 180°.

1.5. Electrostatic potential

The first nonbonded potential is the electrostatic potential, which is derived from the interaction between the partial atomic charges between all nonbonded atom pairs:

$$E_{electrostatic} = \sum_{all\ nonbonded\ atom\ pairs} \frac{q_i q_j}{kDr_{ij}}$$

with q_i and q_j the atomic partial charges of atom *i* and atom *j*, r_{ij} the actual distance between atom *i* and atom *j*, *k* a constant and *D* the dielectric constant. The atomic partial charges can be calculated using a wide variety of methods, including quantum chemical calculations from small molecule systems, and going up to less accurate approaches from protein systems. In most cases, fixed values for the atomic partial charges are used, with electronegative atoms such as N and O being assigned a negative value (for example -0.3 electrons), and carbons and hydrogen atoms a positive value. Depending on the sign of the atomic partial charge, atom pairs may attract or repel each other (Figure 52).

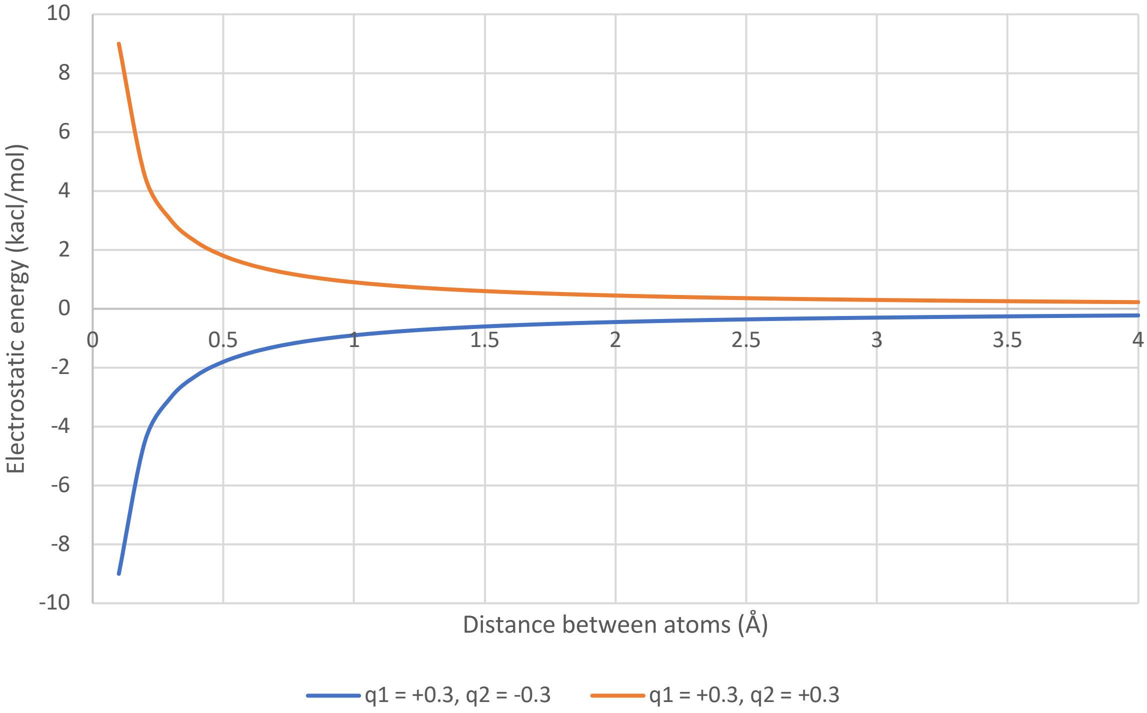


Figure 52. Illustration of the electrostatic energy as a function of the distance between the two atoms. Two cases are shown: the first case (blue) shows the change in electrostatic energy when both atoms carry an opposite charge (+0.3 and -0.3), while the second case shows the change in energy when both atoms carry a partial charge of the same sign (here +0.3 and +0.3). In the former case, the electrostatic energy is attractive, while in the latter situation the potential is repulsive. Note that the potential slowly limits a value of zero at long distances between the two atoms.

1.6. Van der Waals potential

The second nonbonded potential is the van der Waals (VDW) interaction, named after the Dutch scientist Johannes Diderik van der Waals. In contrast to the electrostatic interactions, these attractions are comparatively weak and vanish quickly at longer distances. Van der Waals interactions between two atoms arise from the balance between repulsive and attractive forces. Repulsion is due to the overlap of the electron clouds of both atoms, while the interactions between induced dipoles result in an attractive component. There are many mathematical models to describe the van der Waals interaction, but the 12-6 Lennard-Jones (LJ) potential is the one which is most often used to represent these interactions:

$$E_{vdw} = \sum_{all\ nonbonded\ atom\ pairs} \varepsilon_{ij}\left[\left(\frac{R_{min,ij}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{min,ij}}{r_{ij}}\right)^{6}\right]$$

with r_{ij} the distance between atom *i* and *j*, $R_{min,ij}$ the reference distance at which the interaction is strongest, and ε_{ij} a constant. At interatomic distances shorter than $R_{min,ij}$, the VDW interaction potential is increasing and leads to repulsion between the two atoms. At distances longer than $R_{min,ij}$, the VDW potential also increases and will approach zero at long distances (Figure 53).

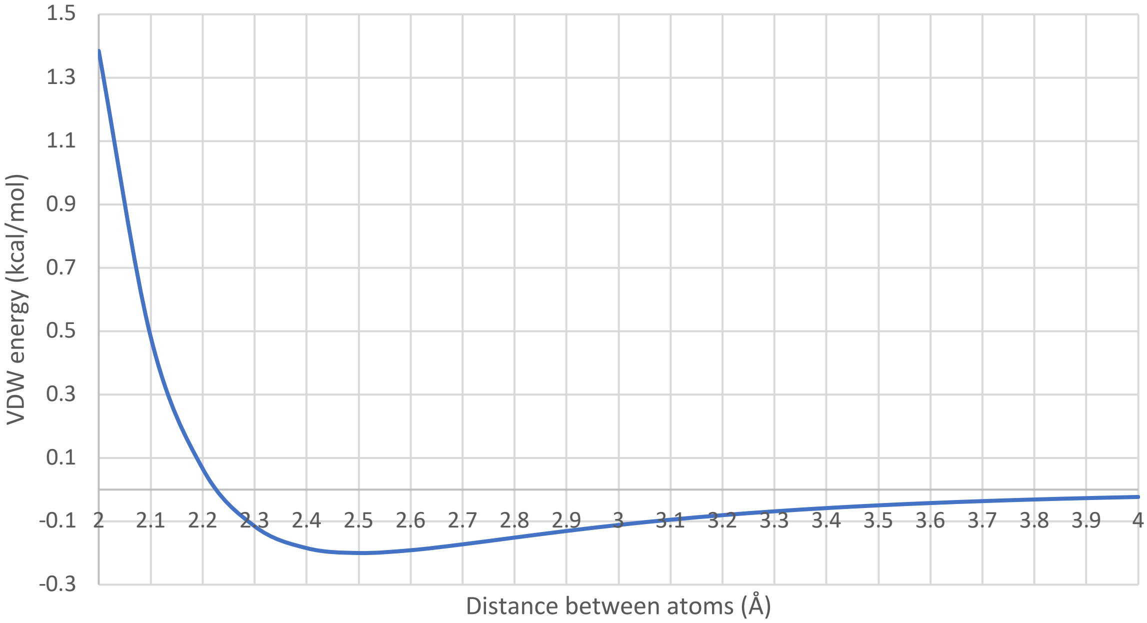


Figure 53. Illustration of the van der Waals potential between two atoms. The distance at which the potential reaches its minimum (in this case 2.5 Å) is called the van der Waals contact distance. This distance is dependent on the atom types; in general atoms with a higher atomic number also have a larger van der Waals radius.

1.7. Popular force fields

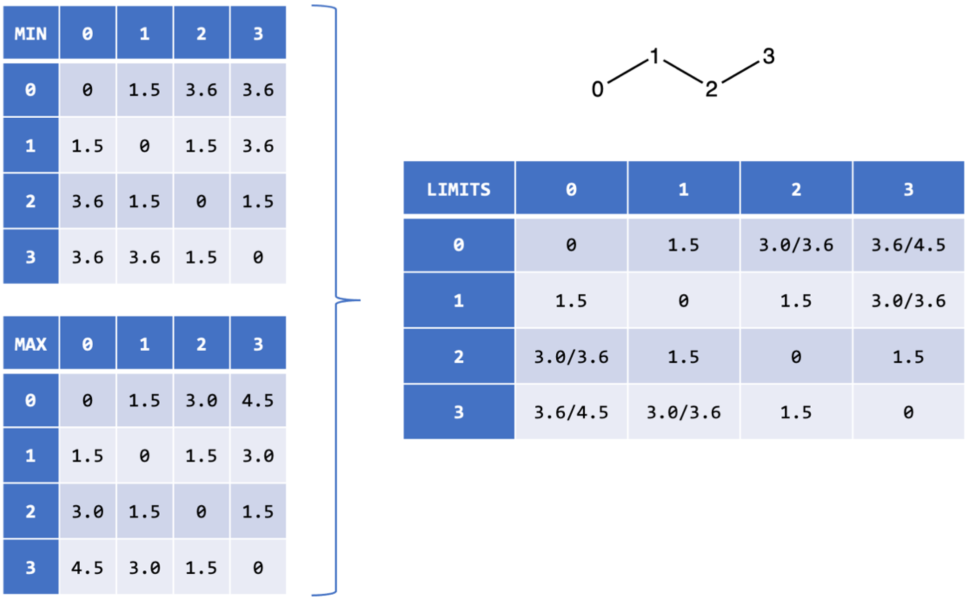
In the preceding sections, an overview was given of the different functions that are commonly used in force fields. The majority of the current force fields all rely on the same kind of functionalities although that small differences exists in both the functional form and/or the parameters used. There exist therefore a number of different force fields and some of them are listed here:

- AMBER – 'Assisted Model Building and Energy Refinement'. Widely used force fields for protein and DNA/RNA.
- CHARMM – 'Chemistry at HARvard Molecular Mechanics'. Originally developed at Harvard, nowadays maintained by Alexander MacKerrell in Baltimore. Widely used for both small molecules and macromolecules.
- GROMOS – 'GROningen MOlecular Simulation'. A force field that comes as part of the GROMOS software, a general-purpose molecular dynamics computer simulation package for the study of biomolecular systems. GROMOS force field A-version has been developed for application to aqueous or apolar solutions of proteins, nucleotides, and sugars. The B-version to simulate gas phase isolated molecules is also available.
- OPLS – 'Optimized Potential for Liquid Simulations'. Developed by William Jorgensen at the Yale University Department of Chemistry. OPLS variants include OPLS-AA, OPLS-UA, OPLS-2001 and OPLS-2005.
- UFF – 'Universal Force Field'. A general force field with parameters for the full periodic table up to and including the actinoids, developed at Colorado State University.
- MMFF – 'Merck Molecular Force Field'. Developed at Merck and suitable for a broad range of molecules.
- MARTINI - A coarse-grained potential developed by Marrink and coworkers at the University of Groningen, initially developed for molecular dynamics simulations of lipids, later extended to various other molecules. The force field applies a mapping of four heavy atoms to one CG interaction site and is parameterized with the aim of reproducing thermodynamic properties.

2. From 1/2D to 3D: Conformation generation with distance geometry

As already described before, molecules are often represented in 1D using the SMILES format, or in 2D with SD-format. However, in order to be able to explore the conformation of molecules, for example within the context of molecular dynamics or docking (see below), these 1D or 2D structures need to be converted into a real 3D representation, with the location in space of each atom represented with a *x*-, *y*-, and *z*-coordinate. There are many approaches that can be used for this purpose, but distance geometry is very often used and quite robust. In this section, we will briefly explain the rationale behind distance geometry, however without going into the mathematical details.

One way to describe the 3D-conformation of a molecule is in terms of the distances between all pairs of atoms. For a molecule that consists of N atoms, there are *N* (*N*-1) / 2 interatomic distances in the molecule, which can be described using a *N* x *N* symmetrical matrix. In this matrix, both the elements (*i*,*j*) and (*j*,*i*) contain the distance between atoms *i* and *j*. The diagonal elements are by definition all zero. The crucial feature about distance geometry is that this matrix cannot contain random distances; rather there are some constraints that can be defined to restrict the distances to a small set of potential solutions. For example, given a simple molecule such as butane, each cell within the distance matrix can be filled with both a lower and a upper distance, as examplified in Figure 54:



MIN	0	1	2	3
0	0	1.5	3.6	3.6
1	1.5	0	1.5	3.6
2	3.6	1.5	0	1.5
3	3.6	3.6	1.5	0

MAX	0	1	2	3
0	0	1.5	3.0	4.5
1	1.5	0	1.5	3.0
2	3.0	1.5	0	1.5
3	4.5	3.0	1.5	0

LIMITS	0	1	2	3
0	0	1.5	3.0/3.6	3.6/4.5
1	1.5	0	1.5	3.0/3.6
2	3.0/3.6	1.5	0	1.5
3	3.6/4.5	3.0/3.6	1.5	0

Figure 54. The principle of distance geometry explained using butane (four carbon atoms) as an example. The bond length between each of the bonded atoms is 1.5 Å, and the VDW radius of each carbon atom is 1.8 Å, implying that the closest distance that is possible between a pair of unbonded carbon atoms is 3.6 Å. The longest distance possible between any pair of unbonded carbon atoms is equal to the number of bonds between them, multiplied by the bond lengths. In practice, there are many additional restraints that can be generated to limit the boundaries in the distance matrix. For example, the minimal and maximal distances between atoms 0 and 2 is not 3.0/3.6 as is shown in the figure, but rather a single 2.45 Å given the fact that these two atoms are 1-3 linked and that the bond angle of 109.5° between 0-1-2 imposes additional limitations on the distance between 0 and 2. Using chemical knowledge many additional restraints can be formulated, like in the case of ring systems or between atoms separated by a torsion angle.

With a completed distance matrix at hand (containing upper and lower bounds), random values are assigned to each interatomic distance between its upper and lower bonds. In the following step, the distance matrix is converted into a trial set of Cartesian coordinates using a series of matrix operations, which are then refined in the final step.

Conformation generation with distance geometry is standard incorporated in RDKit, and can be called very conveniently to generate 3D-conformations from SMILES representations:

```
from rdkit import Chem
from rdkit.Chem import AllChem
mol = Chem.MolFromSmiles("C1CCOCC1NC=O")
mol = Chem.AddHs(mol)
AllChem.EmbedMolecule(mol)
```

It is common to refine the generated conformation(s) using some kind of energy minimization procedure.

3. Energy minimization

Energy minimization (also called energy optimization, geometry minimization, or geometry optimization) is the process of finding a molecular conformation, according to the energy calculated by the used force field, the net inter-atomic force on each atom is acceptably close to zero and the position on the potential energy surface is a (local) minimum. The motivation for performing a geometry optimization is the physical significance of the obtained structure: optimized structures often correspond to a substance as it is found in nature and the geometry of such a structure can be used in a variety of experimental and theoretical investigations such as quantitative structure-activity relationships.

Starting from a random conformation, and because most minimization algorithms can only go downhill, each energy minimization process results in the identification of the nearest local minimum; however, this is not always the global energy minimum (Figure 55).

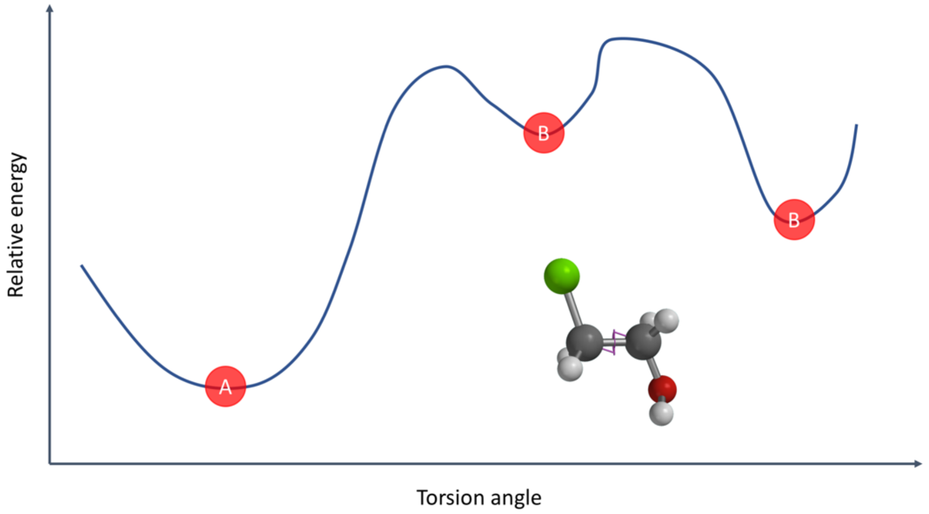


Figure 55. The global minimum (A) and two local minima (B) on the energy profile when rotation along the indicated torsion angle. The relative energy is calculated using a molecular mechanics force field.

A number of methods are available to calculate to local minimum, the steepest descent and conjugate gradients algorithm are two of the most commonly used ones. Of these two methods, steepest descent is several times faster than conjugate gradients, but the latter method converges after a smaller number of iterations and it therefore more productive (Table 10).

Table 10. Comparison between the steepest descent and conjugate gradient methods for energy minimisation of netropsin, a molecule with 13 flexible torsion angles. Metrics for an initial minimisation and a stringent minimisation are given. From reference 6.

Method	Initial minimisation (gradient < 1 kcal/Å^2)		Stringent minimisation (gradient < 0.1 kcal/Å^2)	
	CPU time (s)	Number of iterations	CPU time (s)	Number of iterations
Steepest descent	67	98	1,405	1,893
Conjugate gradient	149	213	257	367

4. Conformational analysis

Conformation generation is the process of transforming the 1D- or 2D-structure of a molecule into a 3D-structure, hence a structure in all atoms of the molecule have *x*-, *y*-, and *z*-coordinates assigned. Examples of 1D-representations include SMILES and InChi, and 2D-representations can include 'flat' SDF or PDB-formats (with 'flat' meaning that the file does not contain *z*-coordinates). Since the majority of molecules are conformationally flexible (rotation around single bonds), multiple conformations exist for a given molecule and conformational analysis is the process of generating these conformations. Conformational analysis can therefore be regarded as the analysis of the conformations that molecules can adopt as a result of single bond rotations, with the intention to locate the global energy minimum and several other minima.

A molecule can adopt an equilibrium between several such minima, the relative abundance of which is determined by the Boltzmann distribution, and which in turn is merely determined by the relative free energy of each pose. The shape of a molecule is not static but is a dynamic equilibrium between a number of conformations, the preferred ones being those we would encounter more times than any other if we were to take a series of snapshots of the population, because they have lower free energies. Conformational analysis can be considered to consist of two parts (Figure 56):

- The equilibrium **distribution** of the different energy minima of a molecule is driven by the **thermodynamics** (the differences in Gibbs free energies between the minima):

$$K_e = e^{-\frac{\Delta G}{RT}}$$

- The **rate** of interconversion between the different minima of a molecule is driven by the **kinetics** (the height of the free energy of activation barrier):

$$\ln\left(\frac{k}{T}\right) = 23.76 - \frac{\Delta G^{\ddagger}}{RT}$$

  where $\Delta G^{\ddagger}$ is the free energy of barrier and *k* the rate constant (in *s*$^{-1}$).

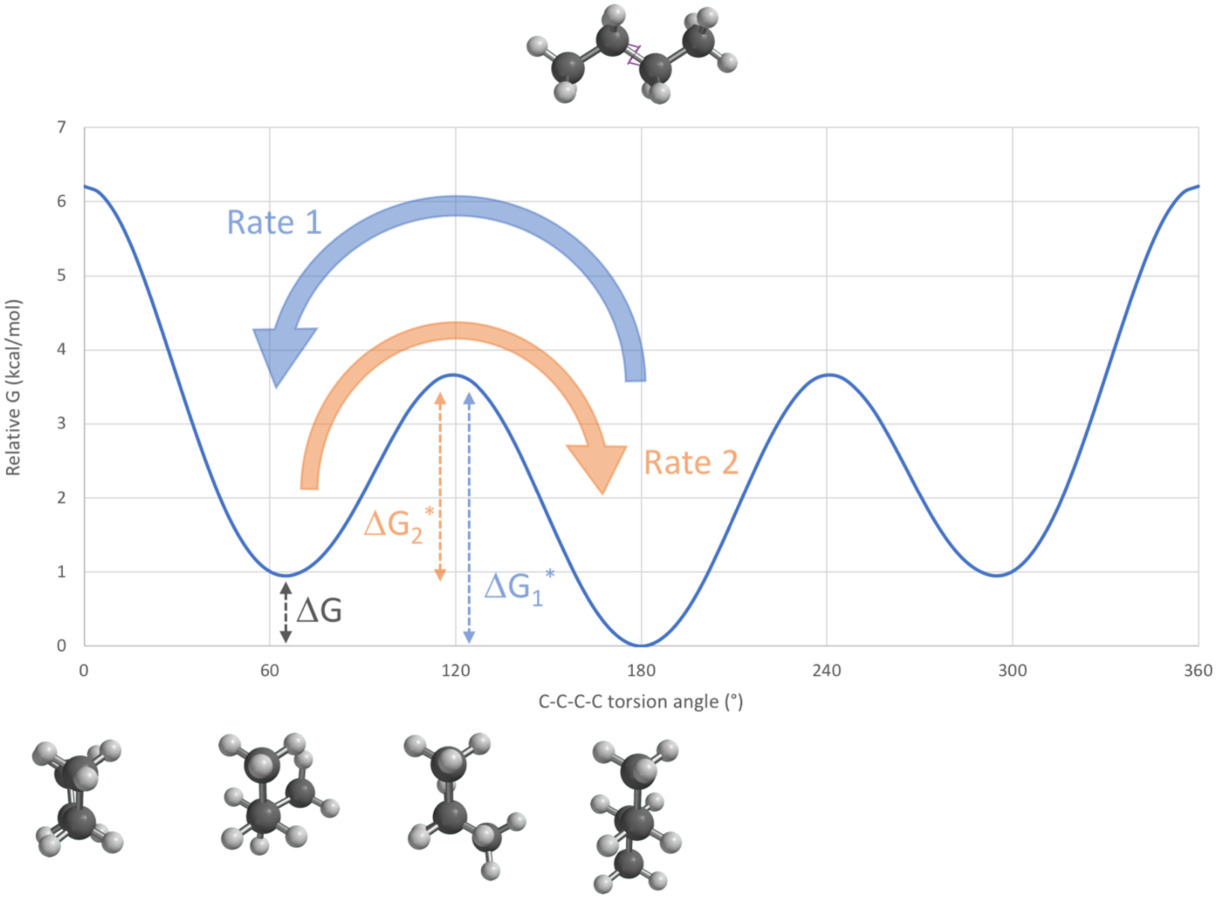


Figure 56. Conformational profile of n-butane, thereby focusing on the C-C-C-C bond. The different conformers for butane are, from left to right, syn-planar (0°), gauche (60°), anti-clinal (120°) and staggered (180°). The free energy difference between the gauche and staggered conformations is about 1 kcal/mol, corresponding to 16% of the conformations in gauche and 84% in staggered. Conversion from staggered to gauche will be slower than the corresponding conversion from gauche to staggered, as can be seen from the differences in activation free energies $\Delta G^{\ddagger}$.

Conformational analysis normally consists of two steps, and which might be iterated over and over again until a satisfactory solution has been identified:

- Energy calculation process.
- Exploration of the conformational space.

4.1. Systematic search

A systematic search is conceptually the simplest of all conformational analysis methods. Using a starting 3D-structure, torsion angles are varied in regular increments and at each step the corresponding energy is calculated. The conceptual simplicity of the systematic search method is in sharp contrast to the combinatorial complexity of its calculation. If A is the torsion angle increment and T is the number of rotatable torsion angles in the molecule, then the total number of possible conformers is $(360/A)^T$. The relative growth in the number of energy calculations is given Table 11.

Table 11. Relative computational complexity of a systematic search as a function of torsion angles and angle increment.

Angle increment (°)	Number of torsions			
	5	10	20	40
30	1	9.1 x 10^{5}	2.1 x 10^{17}	3.2 x 10^{39}
15	3.2 x 10^{1}	9.1 x 10^{8}	2.2 x 10^{23}	3.5 x 10^{51}
8	7.4 x 10^{2}	5.0 x 10^{11}	6.5 x 10^{29}	2.9 x 10^{62}
4	2.4 x 10^{4}	5.0 x 10^{14}	6.8 x 10^{35}	3.2 x 10^{72}
2	7.6 x 10^{5}	5.3 x 10^{17}	7.1 x 10^{40}	3.5 x 10^{86}

As a result of this computational complexity, systematic search methods to identify the global minimum can only be used in case of molecules having a limited number of flexible torsion angles, making the method less useful in practice. For this reason, other methods have been developed, including the Monte Carlo method and genetic algorithms.

4.2. Monte Carlo

Monte Carlo represents a technique to find a good solution to an optimization problem by trying random variations of the current solution. A worse variation is accepted as the new solution with a probability that decreases as the computation proceeds. Monte Carlo is not exhaustive however, meaning that some predefined heuristics are needed to define a suitable endpoint.

In its simple form, starting from a given starting structure random alterations are made and the internal energy of the resulting structure is calculated and an energy minimization step is undertaken. The new conformer produced is saved if the energy or the difference in energies between the new and the best conformer is less than the threshold optimum value. The search is automatically terminated after a user-defined number of unproductive attempts.

The Metropolis Monte Carlo approach amplifies the changes of finding the global minimum. It involves number of sequences where the Monte Carlo algorithm is run at different temperatures. The first phase runs at temperature *T* and an assortment of conformations are generated. The most stable conformation is used as starting origin for next phase where the temperature is set at a lower temperature. The process is repeated until the probability equation becomes selective towards which structure is accepted. Thus, a small part of the conformation space is meticulously investigated. The Metropolis Monte Carlo method introduces a probabilistic criteria to accept new conformers:

$$p = e^{-\Delta E / T k_b}$$

with ΔE being the energy difference between the new conformation and the initial starting one, *T* the temperature and K_b the Boltzmann constant. The acceptance test is performed by choosing a random number *r* (between 0 and 1) which is then compared to *p*. If $r < p$, the change is accepted and the new conformations is taken as the new starting point.

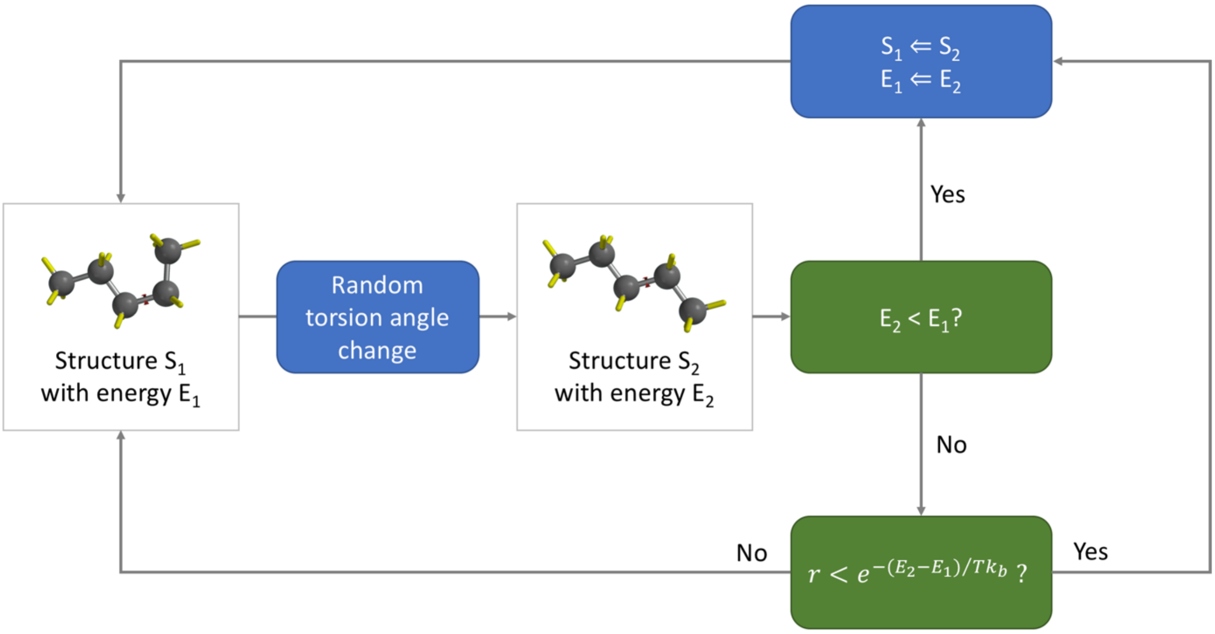

$S_1 \Leftarrow S_2$
$E_1 \Leftarrow E_2$
Yes
Structure S_1 with energy E_1
Random torsion angle change
Structure S_2 with energy E_2
$E_2 < E_1$?
No
No
$r < e^{-(E_2-E_1)/Tk_b}$?
Yes

Figure 57. The Monte Carlo method for conformational analysis. The procedure is stopped after a user-defined number of cycles. The random number is defined as r, T the temperature and k_b the Boltzmann constant. In the Metropolis approach, the temperature T is slowly decreased as the number of runs increase.

4.3. Genetic algorithm

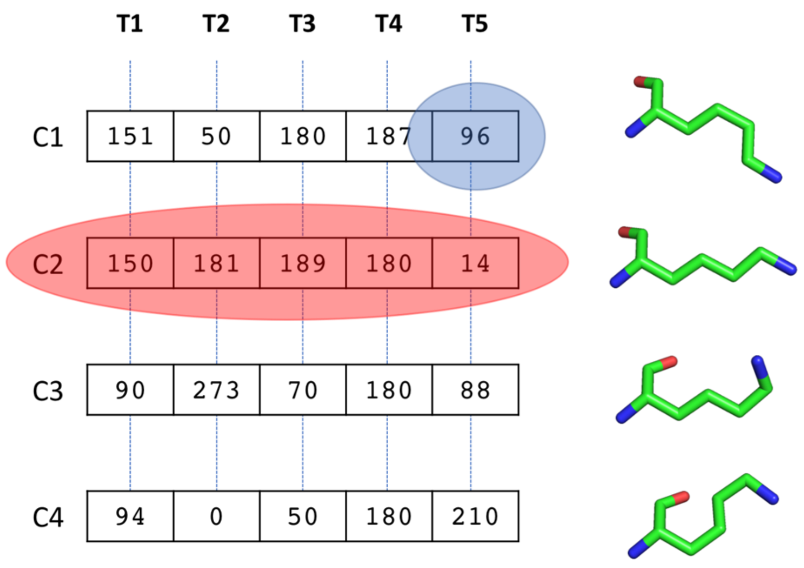
A genetic algorithm (GA) is search heuristic optimization method that is based on various computational models of Darwinian evolution. The genetic algorithm is a large-scale optimization algorithm mimicking a biological evolution in a randomly generated population. The algorithm reflects the process of natural selection where the fittest individuals are selected for reproduction in order to produce offspring of the next generation. A number of conformations forms this population. New chromosomes are generated by modifying some of the torsion angles, and a new population is created in accordance to operators (crossover and mutation). The process is repeated until it converges to a minimum energy structure.

Five phases are considered in a genetic algorithm:

1. Initial population
2. Fitness function
3. Selection
4. Crossover
5. Mutation

Initial population

The process begins with a set of individuals which is called a population. In the context of conformational analysis, each individual is a set of torsion angle values representing a single conformation of the molecule of which one wants to find the global minimum (Figure 58).



	T1	T2	T3	T4	T5
C1	151	50	180	187	96
C2	150	181	189	180	14
C3	90	273	70	180	88
C4	94	0	50	180	210

Figure 58. Genes (blue region), chromosomes (red region) and the population (all chromosomes). A single chromosome defines a single conformer, while each gene represents a single torsion angle in the molecule (in this example, each conformation is represented by five torsion angles).

Fitness function

The fitness function determines how fit an individual is (the ability of an individual to compete with other individuals). It gives a fitness score to each individual. The probability that an individual will be selected for reproduction is based on its fitness score. In the case of a genetic algorithm for conformational analysis, the fitness function is simply the total molecular mechanics energy of each molecule, as calculated by the force field.

Selection

The idea of selection phase is to select the fittest individuals and let them pass their genes to the next generation. Two pairs of individuals (parents) are selected based on their fitness scores (energies). Individuals with high fitness (in casu the lowest energies) have more chance to be selected for reproduction.

Crossover

Crossover is the most significant phase in a genetic algorithm. For each pair of parents to be mated, a crossover point is chosen at random from within the genes. Offspring are created by exchanging the genes of parents among themselves until the crossover point is reached, and the new offspring are added to the population (Figure 59).

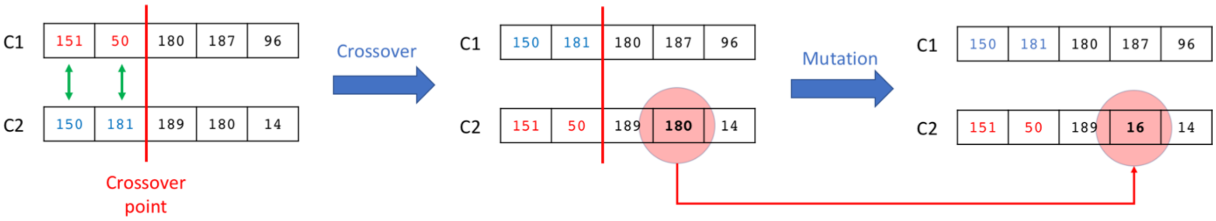


Figure 59. The workflow of the crossover and mutation operators in a genetic algorithm, illustrated on a population of two chromosomes.

Mutation

In certain new offspring formed, some of their genes can be subjected to a mutation with a low random probability. This implies that some of the torsion angles can be changed randomly (between 0 and 359°). Mutation occurs to maintain diversity within the population and to prevent premature convergence (Figure 59).

Pseudocode

```
START
Generate the initial population
Compute energy
REPEAT
      Selection
      Crossover
      Mutation
      Compute energy
UNTIL population has converged
STOP
```