University of Biskra

Department of Materials Science

Faculty of Exact Sciences

Mohamed Khider University of Biskra-Algeria

Molecular modeling

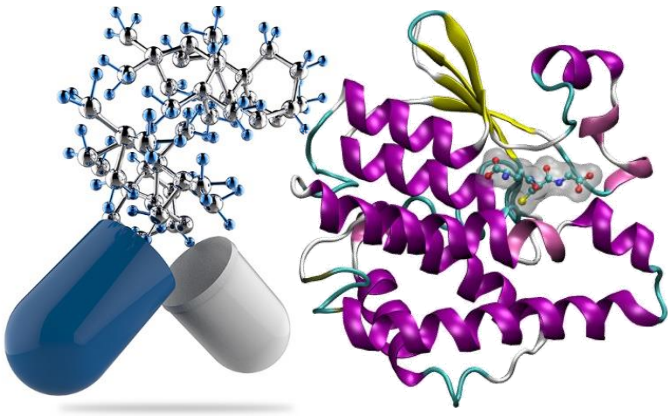
Course and Practical work

Intended for L3 chemistry students

Prepared by

Dr. Hanane Djouama

(Lecturer "A")



2025-2026

Preamble

L3 Pharmaceutical chemistry

Semester: 6

Lecture Unit Title: Methodology.

Subject Title: Molecular modeling

(Credits: 4. Coefficients: 2.)

(Assessment Method: Continuous Assessment: 50%. Final Exam: 50 %.)

Subject Content:

Chapter 1: Simple Hückel Method (4,5 h)

-Principle and characteristic approximations

-Hückel Molecular Orbital Theory of Conjugated

-Hückel Molecular Orbital Theory of Cyclic Systems

-The Hückel Rule

Chapter 2: Molecular Mechanics (4,5 h)

- Force Fields

- Minimization Methods

- Conformational Research

Chapter 3: QSPR Modeling(6 h)

-Simple linear regression

- Linear and multiple regression

- Modeling of physicochemical properties

- Principle and methodology of QSPR modeling

- Molecular descriptors

- Application

Chapter 4: QSAR Modeling(4,5 h)

Modeling of biological activities

- Principle and methodology of QSAR modeling

- Applications

Chapter 5: Molecular Docking(3 h)

- Ligand-receptor interactions

- Principle and methodology of molecular docking – Application

Table of contents

Introduction to Molecular Modeling

1. Introduction to Molecular Modeling

Molecular modeling is anything that foundational the use of a computer to describe, visualize, analyze, and predict molecular properties or evaluate any aspect of the properties of the structure of a molecule (Pensak, 1989).

Recent advancements in artificial intelligence (AI) and machine learning (ML) have significantly improved the ability to analyze complex datasets, predict molecular properties, and simulate complex molecular behaviors at the atomic.

Computational methods in biomolecular modeling include molecular dynamics, Monte Carlo simulations, quantum and molecular mechanics, as well as solvation and free energy estimation tools. This discipline also leverages biochemical databases, chemical informatics, and SAR analyses. Furthermore, it serves as a vital tool for optimizing experimental datasets obtained from NMR and X-ray crystallography.

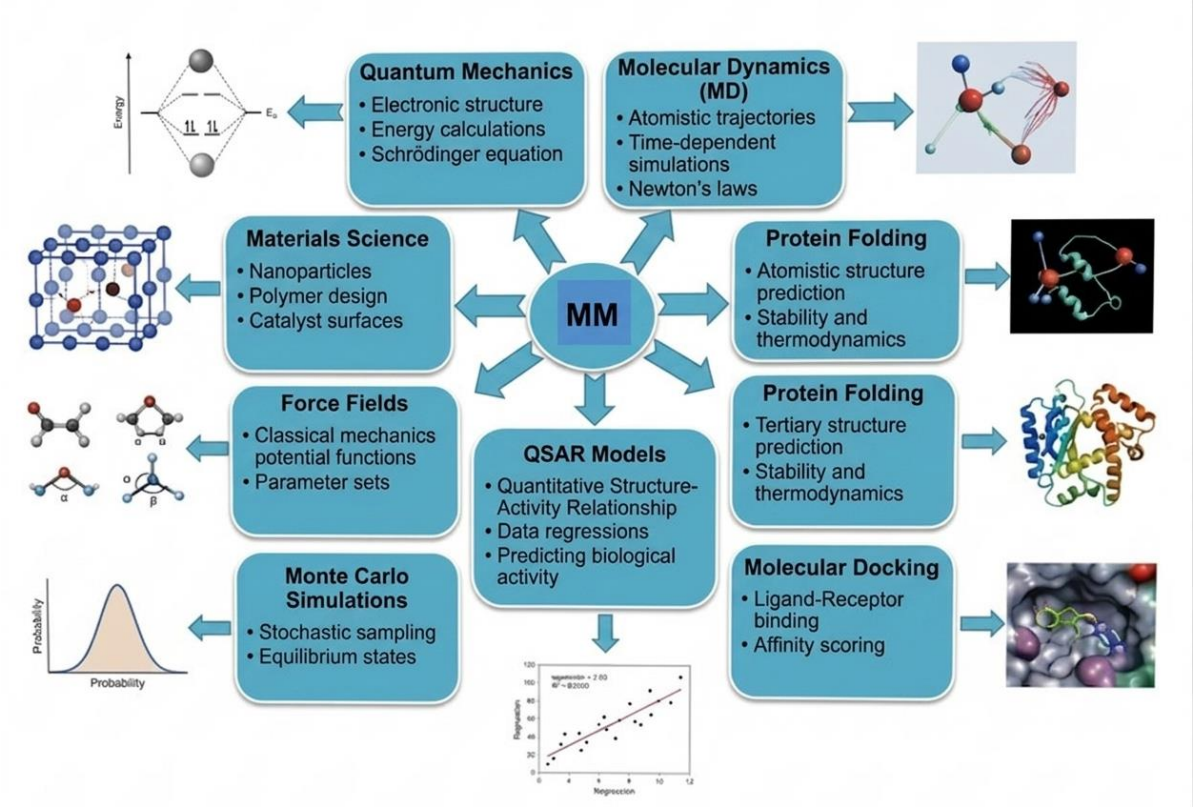


Figure 1. Applications of Molecular Modeling

2. Historical Development of Molecular Modeling

Initiating in the 1920s and 1930s, molecular modeling stems from early breakthroughs in quantum mechanics pioneered by Erwin Schrödinger and Werner Heisenberg.

Their theoretical frameworks unlocked the ability to interpret atomic behavior, primarily through the Schrödinger equation.

As the quantum equivalent of Newtonian classical mechanics, this equation solves for a system's wave function, mapping electron behavior to predict bond structures and molecular energetics.

The discipline evolved further in the 1940s with models capable of calculating spatial geometries, characterizing repulsive steric forces, and identifying thermodynamic energy minima.

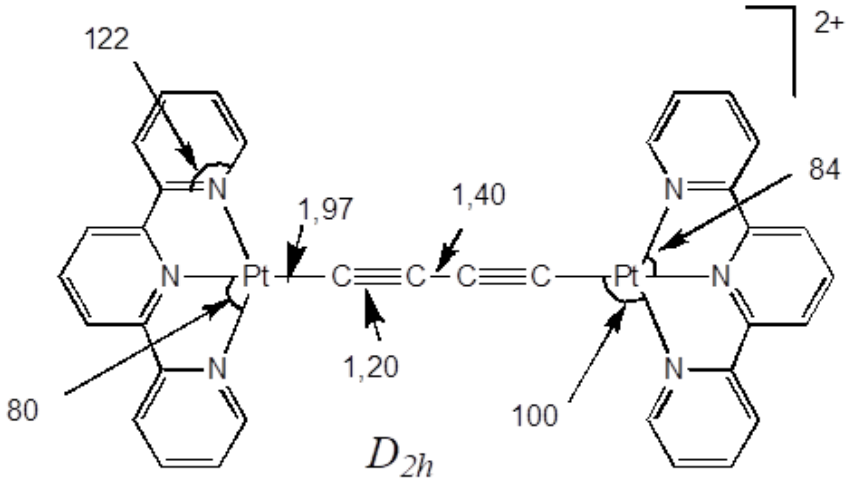


Figure 2. Example of Bond lengths and bond angle values

Modern modeling techniques originated in 1953 with the introduction of Monte Carlo frameworks and the Metropolis algorithm. This was followed by the first-ever molecular dynamics simulation in 1957, conducted by Alder and Wainwright using their "Hard Sphere" model.

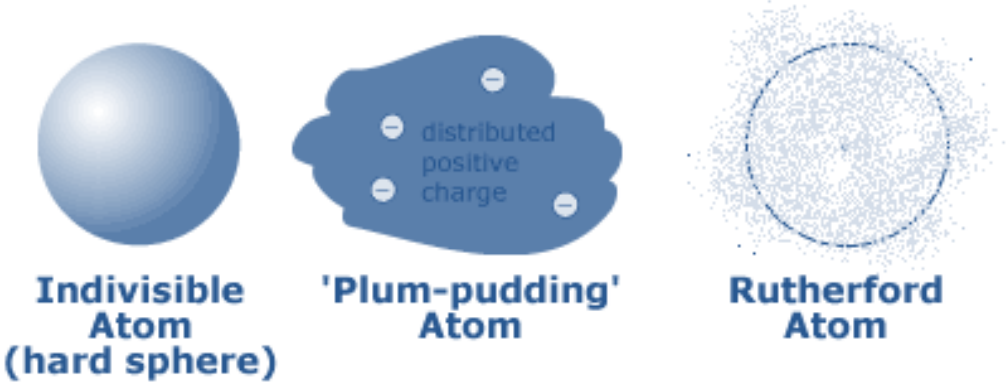


Figure 3. Models of the atom

In the 1960s, computers were first introduced to analyze force fields and run conformational analysis algorithms. A milestone occurred in 1964 when the universities of Oxford and UC Berkeley leveraged computer processors for Hückel (LCAO) calculations to evaluate π-electron energies in conjugated networks. During the 1970s, pioneering programs like GAUSSIAN and ATMOL significantly optimized ab initio molecular orbital calculations. Additionally, the Protein Data Bank (PDB) was founded in 1971, and the initial simulation of protein dynamics was achieved in 1977. The field expanded rapidly in the 1980s as desktop computers and graphical user interfaces (GUIs) made modeling tools accessible to a broader audience. From the 1990s onward, the advent of supercomputing and advanced software allowed researchers to model increasingly complex molecular structures.

Figure 4. High-Performance Computers

Walter Kohn and John Pople were awarded the 1998 Nobel Prize in Chemistry for "the development of density functional theory" and "the development of computational methods in quantum chemistry", respectively.

3. Software and Tools Used in Computational Chemistry

Modern molecular modeling relies heavily on specialized software to execute theoretical tasks like energy calculations, reaction simulations, and molecular dynamics.

Gaussian: A leading industry standard for analyzing electronic structures, refining molecular geometries, and identifying transition states.

ORCA: A highly flexible, freely accessible quantum chemistry suite known for its robust computational performance.

GAMESS: An advanced toolkit designed to compute molecular geometry, electronic wavefunctions, and total system energies.

Q-Chem: A comprehensive program optimized for implementing cutting-edge methods in quantum chemistry.

CP2K: A powerful framework tailored specifically for density functional theory (DFT) computations and complex molecular dynamics.

NWChem: A scalable platform that integrates both quantum mechanics and classical physics to perform ab initio, DFT, and dynamic simulations.

Quantum ESPRESSO: An open-source suite optimized for DFT research on periodic structures, heavily utilized in materials science and solid-state physics.

4. The Importance of Computer Modeling

Employing computational chemistry software grants the ability to investigate numerous molecular attributes, notably:

-Determination of electronic configurations,

-Optimization of spatial geometries,

-Calculation of vibrational frequencies,

-Characterization of transition structures and reaction trajectories,

-Biomolecular simulations, particularly protein docking,Mapping of charge and electron distributions,

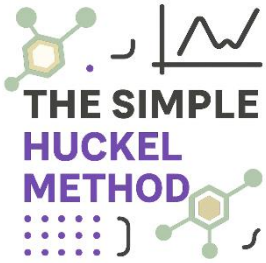
-Generation of potential energy surfaces (PES).

-Kinematics calculations regarding reaction rate constants.

-Evaluation of thermodynamic functions like enthalpy changes and activation barrier heights.

-Estimation of a broad spectrum of macroscopic physical and chemical attributes.

These simulations rely heavily on three established numerical techniques:
ab initio, semi-empirical, and molecular mechanics. Their foundational definitions include:
ab initio: Latin for "from first principles," indicating methods that evaluate molecular systems exclusively via the Schrödinger equation, atomic identities, and universal constants.
Semi-empirical: Hybrid methodologies that rely on empirical data adjustments to streamline computational algorithms.Molecular mechanics: A classical physics approach that treats atoms as spheres governed by empirical or semi-empirical force fields to model structural behavior.



Chapter 1: The Simple Hückel Method

1. Introduction

Erich Hückel proposed the initial form of the molecular orbital (MO) theory of conjugated molecules in 1931.

The method has aided obtain main theoretical results in studying of the properties of organic molecules and the mechanisms of their chemical reactions.

For example, he provided the theoretical basis of Hückel's rule for the aromaticity of (4n + 2) π cyclic planar systems of electrons.

2. Principle and characteristic approximations

We have to use approximation methods to solve Schrodinger's equation for complicated systems. Approximation methods give us an approximate answer – not exactly the true value, but a value close to the exact value. **Hückel** theory uses the variational approximation method to Linear Combination of Atomic Orbitals (LCAO).

In the π-electron approximation, the nπ -π electrons are treated separately by incorporating the effects of the electrons and the nuclei into some sort of effective p electron Hamiltonian H_π.

For the OM ϕ_i, we use the LCAO approximation and we therefore have: $\phi i = \sum_j cij\aleph j$ (where the $\aleph j$ are the OAs). By injecting ¸ca into the single-electron Schrodinger equation, we project onto the χ_k, and we arrive at the secular equations:

$$\sum_j c_{ij} h_{kj} = \varepsilon_i \sum_j c_{ij} S_{kj}$$

where, c_i are the coefficients of the p_z orbitals of individual carbon atoms, whose values are determined by minimizing the variational energy.

The coefficients that minimize the energy are obtained from the non-trivial solution of the n secular equations, represented in the following determinant form,

$$\begin{vmatrix} H_{11} - ES_{11} & \dots & H_{1n} - ES_{1n} \\ \vdots & \ddots & \vdots \\ H_{n1} - ES_{n1} & \dots & H_{nn} - ES_{nn} \end{vmatrix} \cdot \begin{vmatrix} c_1 \\ \vdots \\ c_n \end{vmatrix} = 0$$

Hückel introduced approximations into modified wave equation, and they are: *Coulomb integral (α)*

For carbon *i*,

$$H_{ii} = \int \Phi_i \varkappa \Phi_i d\tau = \alpha,$$

where H_{ii} is a matrix of elements which appear on the diagonal of the determinant. These terms represent the energy of the interactions of electrons in the isolated 2pz orbital with its own nucleus.

Integral of the resonance (β) represents the energy of the interaction of electrons of two nuclei at the same time, made by the overlapping of 2pz AO on the observed atoms

$$H_{ij} = \int \Phi_i \varkappa \Phi_j d\tau = \beta \text{ or } 0$$

Under these conditions, the uniform lengths of the σ bonds combined with the shared nodal plane of the $2p_z$ atomic orbitals ensure highly symmetric interactions across all i j pairs.

Consequently, these non-bonded parameters are set to zero to simplify the mathematical matrix. Meanwhile, the overlap integrals (S_{ij}) serve to gauge the physical overlap between the atomic *p* orbitals.

$$S_{ij}=\int \Phi_i \Phi_j d\tau=1 \ (i=j)$$
$$=0 \ (i\neq j)$$

Introducing all three approximations, secular equations and determinants are becoming simpler and easier for solving, even for the more complex systems:

$$C1(\alpha-\varepsilon)+C2\beta12+C3\beta13+\ldots+Cn\beta1n=0$$
$$C1\beta21+C2(\alpha-\varepsilon)+C3\beta23+\ldots+Cn\beta2n=0$$
$$C1\beta31+C2\beta32+C3(\alpha-\varepsilon)+\ldots+Cn\beta3n=0$$
$$\cdot$$
$$\cdot$$
$$\cdot$$
$$C1\beta n1+\ldots+Cn(\alpha-\varepsilon)=0$$

$$\begin{vmatrix} \alpha-\varepsilon & \beta_{12} & \beta_{13} & \cdots & \beta_{1n} \\ \beta_{21} & \alpha-\varepsilon & \beta_{23} & & \beta_{2n} \\ \beta_{31} & \beta_{32} & \alpha-\varepsilon & & \vdots \\ \vdots & & & \ddots & \\ \beta_{n1} & & \cdots & & \alpha-\varepsilon \end{vmatrix} = 0$$

Hückel's method consists of parameterizing the integrals h_{kj} and S_{kj}. It is valid if there is only one atomic orbital per atom considered. This is the case for polyenes, where the basis considered is then the basis of 2pz atomic orbitals. We define:

$h_{ii} = \alpha_i$ (called the Coulomb integral)

$h_{ij} = 0$ if i and j are not adjacent

$h_{ij} = \beta_{ij}$ otherwise (called the resonance integral)

$S_{ij} = \delta_{ij}$

The Coulomb integral represents the energy of the atomic orbital χi before interaction; it is therefore a characteristic parameter of the atom bearing the atomic orbital. The resonance integral is characteristic of the bond between atoms i and j.

Pauling and Wheland showed (1950) that calculations could be simplified for heteroatomic molecules by expressing the Coulomb and resonance integrals as functions of the carbon parameters and the electronegativity of the atoms: $\alpha_X = \alpha + h_X\beta$ and $\beta_X = k_X\beta$. We often write $\alpha_C = \alpha$ and $\beta_{CC} = \beta$. Some of these parameters are given in Table 1

Table1. Parameters of hetero-elements in Hückel's method

Hückel Method Parameters	Coulomb Integral	Resonance Integral
Carbon	$\alpha_C = \alpha$	$\beta_{CC} = \beta$
Oxygen (1 electron)	$\alpha_O = \alpha + \beta$	$\beta_{CO} = \beta$
Oxygen (2 electrons)	$\alpha_O = \alpha + 2\beta$	$\beta_{CO} = 0.8\beta$
Nitrogen (1 electron)	$\alpha_N = \alpha + 0.5\beta$	$\beta_{CN} = \beta$
Nitrogen (2 electrons)	$\alpha_N = \alpha + 1.5\beta$	$\beta_{CN} = 0.8\beta$
Fluorine	$\alpha_F = \alpha + 3\beta$	$\beta_{CF} = 0.7\beta$
Chlorine	$\alpha_{Cl} = \alpha + 2\beta$	$\beta_{CCl} = 0.4\beta$
Bromine	$\alpha_{Br} = \alpha + 1.5\beta$	$\beta_{CBr} = 0.3\beta$
Methyl (heteroatomic model)	$\alpha_{Me} = \alpha + 2\beta$	$\beta_{CMe} = 0.7\beta$

An oxygen atom with one electron is called an oxygen atom with a single 2pz orbital (like carbon): this is the case for an oxygen atom in a carbonyl group.

Oxygen atoms with two electrons involve a pair of electrons, as in an enol. It is normal to have $\alpha_O < \alpha_C$ because oxygen is more electronegative than carbon; therefore, the energy of its orbital is lower. The parameter βij should take into account the length of the i-j bond, and thus the Hückel method is sometimes used with a parameter β that depends on Rij. Finally, when we have branches (a methyl group for example), it is common to replace the group with a fictitious heteroatom X which contributes an electron to the π system and to which certain parameters are given.

3. Hückel Molecular Orbital Theory of Conjugated

The ethylene molecule serves as a simple model to understand Hückel's molecular orbital theory.

First, we want to establish the simplest method of how Hückel calculated molecular orbitals. Anyone actually interested in the method is invited to study Hückel's treatment of the **Ethylene** molecule.

a. Secular Determinant and Energies:

$$\phi i = \sum_j cij\aleph j \; + H = \sum_i hi$$

we find;

$$\begin{vmatrix} H_{11} - ES_{11} & H_{12} - ES_{12} \\ H_{12} - ES_{12} & H_{22} - ES_{22} \end{vmatrix} = 0$$

with;

-The Orbitals are Normalized: Sii = 1

-The overlap between orbitals is 0: Sij = 0 (i≠j)

-The diagonal Hamiltonian elements are given by an empirical parameter, α

- Off-diagonal Hamiltonian elements are given by an empirical parameter, β, if the carbons are adjacent

$H_{ij} = \beta$: Adjacent Carbons

$H_{ij} = 0$: Non-Adjacent Carbons **Note:** $\alpha<0$ and $\beta<0$

In consequence, we employ the following secular equation to describe a diatomic molecule:

$$(\alpha - E)\, c1 + \beta\, c2 = 0$$

$$\beta\, c1 + (\alpha - E)\, c2 = 0$$

The secular equation is obtained as

$$\begin{vmatrix} \alpha\text{-}E & \beta \\ \beta & \alpha\text{-}E \end{vmatrix} = 0$$

Division by β and denoting (α-E)/β as x yields

$$\begin{vmatrix} x & 1 \\ 1 & x \end{vmatrix} = 0$$

and leades to the solution

$$\begin{vmatrix} x & 1 \\ 1 & x \end{vmatrix} = 0$$

$$\downarrow$$

$$x^2 - 1 = 0$$

$$x_1 = -1 = \frac{\alpha - E_1}{\beta} \qquad x_2 = +1 = \frac{\alpha - E_2}{\beta}$$

So the energies of the molecular orbitals are:

$E_1 = \alpha + \beta$ - Highest Energy

$\alpha = E$

$E_2 = \alpha - \beta$ Lowest Energy

This implies that $E_i = \alpha \pm \beta$ and the ground state electron configuration of the π electrons is:

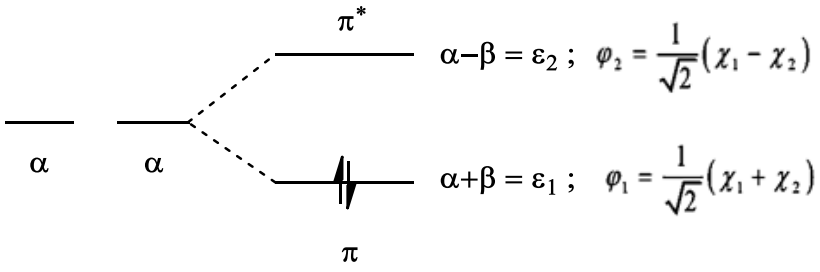


Figure 1. Hückel energy levels in ethylene

b. Molecular Orbitals

Hückel molecular orbital (MO) theory is a simplified method for calculating the electronic structure of conjugated π-electron systems. It's a valuable starting point for understanding computational quantum chemistry and has been extended, analyzed, and applied in various ways. Within HMO theory, the electronic Schrödinger equation of a planar conjugated system

is solved with a linear variational approach where the molecular wave function is expressed as a linear combination of the p_z atomic orbitals (ϕ) of n carbon atoms.

The basic form of the secular determinant for the bonding arising from the overlap of two orbitals is reproduced below.

$$\begin{vmatrix} \alpha - E & \beta \\ \beta & \alpha - E \end{vmatrix} = 0$$

$$\phi = c_1\chi_1 + c_2\chi_2 \longrightarrow \begin{matrix} (\alpha - E)c_1 + \beta c_2 = 0 \\ \beta c_1 + (\alpha - E)c_2 = 0 \end{matrix}$$

$$xc_1 + c_2 = 0$$
$$c_1 + xc_2 = 0$$

$$\frac{c_2}{c_1} = -x \qquad \begin{matrix} x_1 = -1 \\ E_1 = \alpha + \beta \end{matrix} \longrightarrow \frac{c_2}{c_1} = +1 \longrightarrow c_2 = c_1$$

If we now apply the normalisation condition ($\sum c_i^2 = 1$)

$$\begin{matrix} 1 = c_1^2 + c_2^2 \\ c_2 = c_1 \end{matrix} \longrightarrow \varphi_1 = \frac{1}{\sqrt{2}}(\chi_1 + \chi_2)$$

$$\begin{matrix} x_2 = +1 \\ E_2 = \alpha - \beta \end{matrix} \longrightarrow \frac{c_2}{c_1} = -1 \longrightarrow c_2 = -c_1$$

$$\begin{matrix} 1 = c_1^2 + c_2^2 \\ c_2 = -c_1 \end{matrix} \longrightarrow \varphi_1 = \frac{1}{\sqrt{2}}(\chi_1 + \chi_2)$$

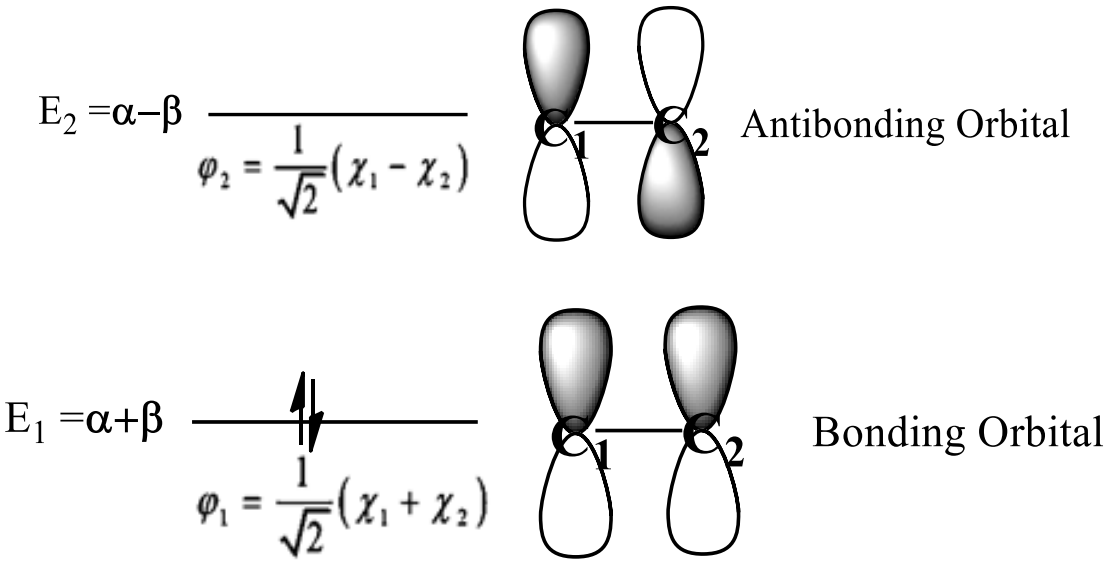


Figure 2. Huckel energy levels and MOs for ethylene

c. π Electron Charge and π Bond Order

The π electron charge on atom μ is defined by:

$$q_\mu \equiv \sum_i^{orbs} n_i c_{i\mu}^2$$

$C_{i\mu}$ is the coefficient of the i'th . MO on atom μ.

–π Bond Order

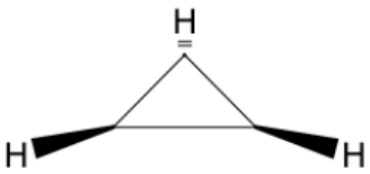
The π bond order between atoms μ and ν is defined by:

$$p_{\mu\nu} \equiv \sum_i^{orbs} n_i c_{i\mu} c_{i\nu}$$

4. Hückel Molecular Orbital Theory of Cyclic Systems

Cyclic conjugated systems are those in which all of the carbons bearing the 2p AOs form a ring. Here we discuss monocyclic conjugated systems in which the conjugated carbon atoms are contained in a single ring, such as cyclopropenyl, cyclobutadiene, benzene, etc.

The simplest cyclic conjugated system is the three-carbon atom cyclopropenyl system



a. Secular Determinant and Energies:

We have π- molecular orbitals, which are delocalized over the three atoms. For cyclopropenyl system, the interaction matrix leads to the following determinant:

$$\begin{vmatrix} \alpha\text{-}E & \beta & \beta \\ \beta & \alpha\text{-}E & \beta \\ \beta & \beta & \alpha\text{-}E \end{vmatrix} = 0$$

$$x = \frac{\alpha\text{-}E}{\beta} \longrightarrow$$

$$\begin{vmatrix} x & 1 & 1 \\ 1 & x & 1 \\ 1 & 1 & x \end{vmatrix} = 0$$

$$\Rightarrow x(x^2-1) - 1(x-1) + 1(1-x) = 0$$

$$\Rightarrow x^3 - 3x + 2 = 0$$

$$\Rightarrow (x-1)(x-1)(x+2) = 0$$

$$\Rightarrow x = 1,1,-2$$

Cyclopropenyl, has one pair of degenerate MO's which are antibonding. The display of MO's is shown below. Note that there is one very highly bonding MO, along with the degenerate pair.

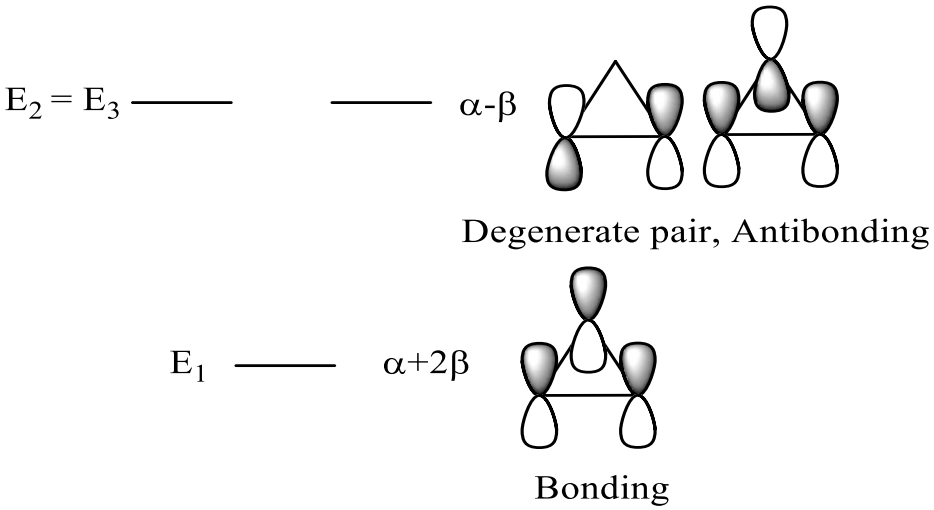


Figure 3. Huckel energy levels and MOs for cyclopropenyl

5. The Hückel Rule

The π -electron amount is defined by the series of numbers generated from:

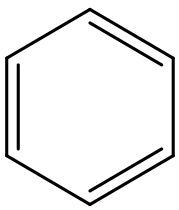
4n+2 where n = zero or any positive integer (i.e., n = 0, 1, 2, etc.).

This rule is Hückel's rule.

-It is used to detect the aromaticity of the ring-shaped planer molecule or ion.

The most common case is six pi electrons (n = 1), which is found in benzene, pyrrole, furan, and pyridine

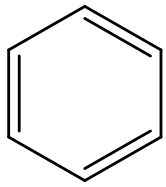
According to this rule, a cyclic, planar molecule is considered aromatic if it has (4n + 2)π electrons, where n is a non-negative integer. For n=1: 4n+2 = 6; benzene is stable and the electrons are delocalized.



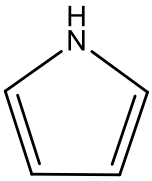
Benzene
Three double bonds;
six π electrons

-Examples of Hückel's Rule

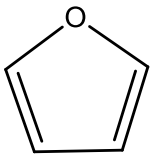
Aromatic molecules can be both homocyclic and heterocyclic. Heterocyclic molecules have atoms other than carbon in their molecular structure



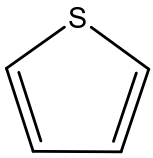
Benzene
π electrons = 6
n= 1



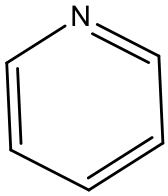
Pyrrole
π electrons = 6
n= 1



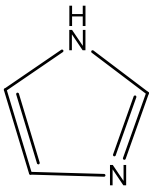
Furan
π electrons = 6
n= 1



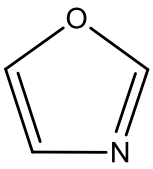
Thiophene
π electrons = 6
n= 1



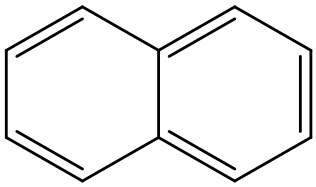
Pyridine
π electrons = 6
n= 1



Imidazole
π electrons = 6
n= 1



Oxazole
π electrons = 6
n= 1



Napthalene
π electrons = 10
n= 2

Figure 4. Hückel's Rule for Aromatic molecules (Number of π electrons = 4n + 2)

6. Aromaticity – Antiaromaticity

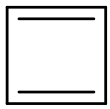
The aromatic compounds are known based on the Hückel's rule.

- Aromatic systems are cyclic, planar, π-conjugated structures with 4n+ 2π electrons, while antiaromatic systems have 4nπ electrons.

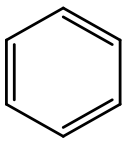
-To be aromatic, a molecule must be planar conjugated, and obey the 4n+2 rule.

For example, is this compound aromatic?

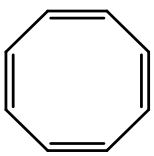
2 π electrons
Aromatic



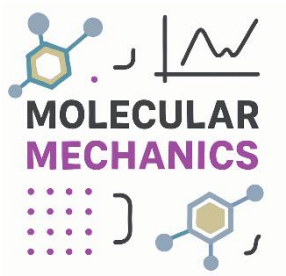
4 π electrons
Not Aromatic



6 π electrons
Aromatic



8 π electrons
Not planar (flat)
Not Aromatic



Chapter 2: Molecular Mechanics

1. Introduction

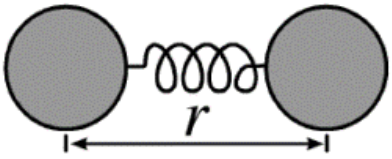
Molecular mechanics (MM) is an essential computational tool that models chemical systems by treating atoms as classical particles governed by empirical force fields. While quantum chemistry directly solves the Schrödinger equation for electrons, MM avoids this complexity by using parametrized functions to approximate the potential energy surface. This simplification allows researchers to simulate incredibly large systems, including materials and macromolecules, without the heavy computational burden of quantum methods. For this reason, MM remains a core component of modern multiscale simulations.

2. Definition

Molecular mechanics (MM) is an empirical method for calculating properties of molecules that includes aspects such as heat of formation, strain energy, dipole moment, and vibrational frequencies, utilizing molecular force fields to visualize the movement of atoms in molecules.

3. Basics of Molecular mechanics

The mechanical molecular model considers atoms as spheres and bonds as springs. The mathematics of spring deformation used to describe the ability of bonds to stretch, bend, and twist. Only the position of nuclei (atom) taken into account to calculate the energy. We adopt Born-Oppenheimer approximation.



4. Basics of Born-Oppenheimer (BO) approximation

In molecular physics, the Born-Oppenheimer (BO) approximation simplifies the molecular Schrödinger equation by treating nuclear and electronic wave functions separately. This formulation relies on the fact that nuclei are drastically heavier than electrons.

Molecules have N nuclei (μ, v,….) and n electrons (I, j,…..) the complete Hamiltonian operator are,

$$\hat{H} = -\sum_{\mu=1}^{N}\frac{\hbar^2}{2m_\mu}\nabla_\mu^2 - \frac{\hbar^2}{2m_e}\sum_{i=1}^{n}\nabla_i^2 + \sum_{\mu<\nu}^{N}\frac{Z_\mu Z_\nu e^2}{4\pi\epsilon_0 r_{\mu\nu}} - \sum_{i,\mu}\frac{Z_\mu e^2}{4\pi\epsilon_0 r_{i\mu}} + \sum_{i<j}^{n}\frac{e^2}{4\pi\epsilon_0 r_{ij}}$$

Above this equation there is a five-term is described as a different indication, first term act as nuclear kinetic energy, the second electronic kinetic energy; third, nuclear potential energy; the fourth, nuclear – electron attraction potential energy, and the fifth, electric repulsion potential energy.

Schrodinger equation for the molecule is as follows,

$$\hat{H}\psi = \mathrm{E}\psi$$

where E = Total energy,

ψ = Total wave function

Under the assumption of decoupled motions, the total wave function of the system is mathematically formulated as a product state of the electronic wave function ψ_e and nuclear function ψ_n :

$$\Psi_{total} \approx \psi_e(r;R)\cdot\psi_n(R)$$

$$\hat{H}_{elec}\psi_e(r;R) = E_e(R)\psi_e(r;R)$$

$$\left[-\frac{\hbar^2}{2m_e}\sum_i^n \nabla_i^2 + \sum_{\mu,\nu}\frac{z_\mu z_\nu e^2}{(4\pi\epsilon_0)r_{\mu\nu}} - \sum_{i,\mu}\frac{z_\mu e^2}{(4\pi\epsilon_0)r_{i\mu}} + \sum_{i,j}\frac{e^2}{(4\pi\epsilon_0)r_{ij}}\right]\psi_e = E\psi_e$$

The Born – Oppenheimer approximation of the electronic and nuclear motions has been separated;

$$E_e = E_e(R)$$

For a diatomic molecule plot, a graph E_e (R) vs R gives the potential energy curve shown in the graph;

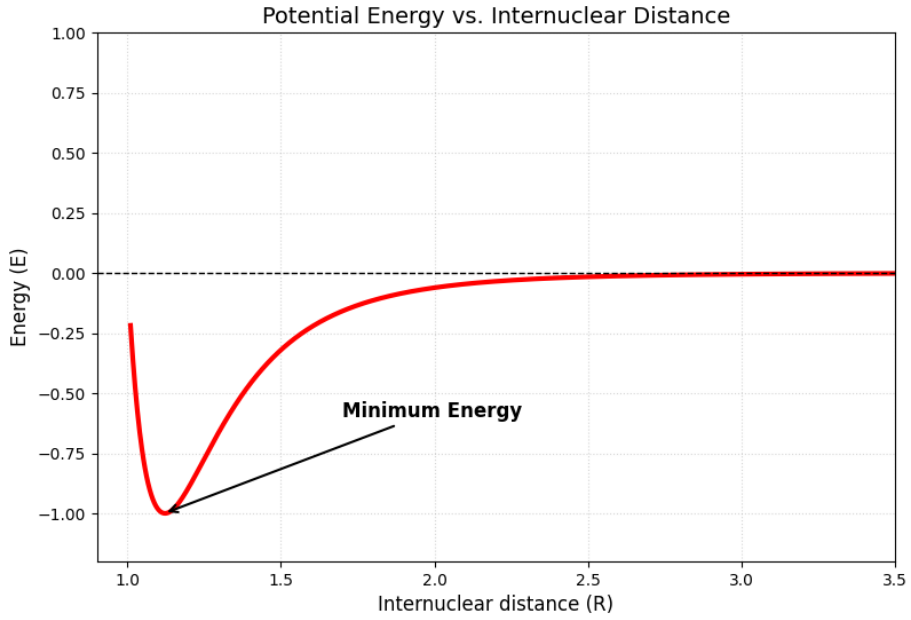


Figure1. Energy vs. Inter-nuclear distance

5. Force fields

A force field is a mathematical expression describing the dependence of the energy of a system on the coordinates of its particles. It consists of an analytical form of the interatomic potential energy, U (r1, r2, . . . , rN), and a set of parameters entering into this form.

Molecules are distinct as a set of atoms that is held together by simple elastic (harmonic) forces and the FF replaces the true potential with a simplified model valid in the region being simulated.

A force field is a set of empirical energy functions and parameters used to calculate the potential energy U as a function of the molecular coordinates.

Potential energy function is composed of non-bonded and bonded interactions:

$$U(\vec{r}) = \sum U_{bonded}\ (\vec{r}) + \sum U_{non-bonded}\ (\vec{r})$$

Unlike coarse-grained methods, all-atom molecular mechanics (MM) explicitly accounts for every single atom within a system. Under this framework, the comprehensive potential energy is formulated as the cumulative sum of both bonded and non-bonded contributions. The bonded components encompass bond stretching, angle bending, and torsional (dihedral) rotations, whereas the non-bonded interactions are governed by electrostatic and van der Waals forces.

$$U(R) = \sum_{bonds} k_r (r - r_{eq})^2 \quad \text{bond}$$

$$+ \sum_{angles} k_\theta (\theta - \theta_{eq})^2 \quad \text{angle}$$

$$+ \sum_{dihedrals} k_\phi (1 + \cos[n\phi - \gamma]) \quad \text{dihedral}$$

$$+ \sum_{impropers} k_\omega (\omega - \omega_{eq})^2 \quad \text{improper}$$

$$+ \sum_{i<j}^{atoms} \varepsilon_{ij} \left[\left(\frac{r_m}{r_{ij}} \right)^{12} - 2 \left(\frac{r_m}{r_{ij}} \right)^{6} \right] \quad \text{van der Waals}$$

$$+ \sum_{i<j}^{atoms} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \quad \text{electrostatic}$$

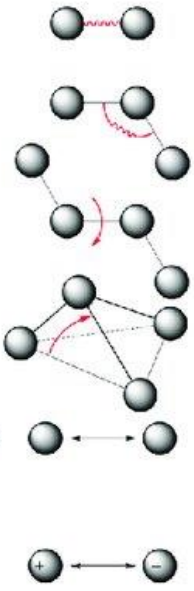


Figure 2. Potential energy function is composed of non-bonded and bonded interactions

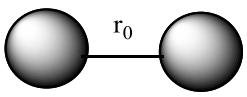
a. Bonded interactions (Intramolecular terms)

-The bond potential

Bond stretching is very often represented with a simple harmonic function that controls the length of covalent bonds.

Oscillation about an equilibrium bond length r_0 with bond constant k_b:

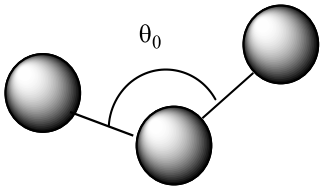
$$V_{Bond} = k_b (r_{ij} - r_0)^2$$



-The angle potential

Oscillation about an equilibrium angle θ_0 with force constant k_θ:

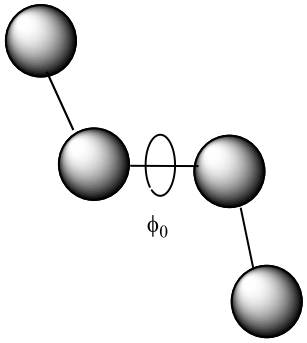
$$V_{Angle} = k_\theta(\theta_{ijk} - \theta_0)^2$$



-The torsion (dihedral) angle potential

Distinct for every four successively bonded atoms.

$$V_{torsion} = k_\phi(1 + cos(n\phi - \delta)) + \cdots$$

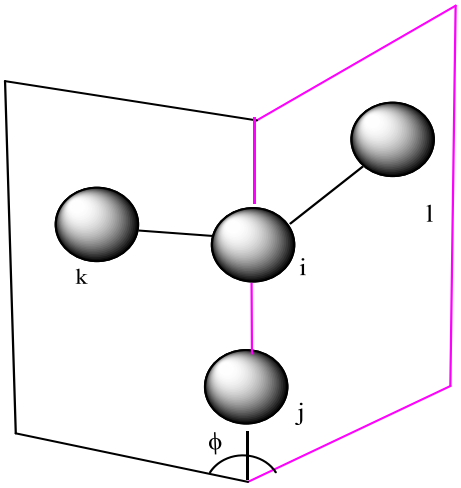


n represents the number of potential maxima or minima generated in a 360° rotation.

- The improper torsion potential

Defined for a group of four bonded atoms where the central atom i is connected to the three peripheral atoms j,k, and l. Used to enforce planarity.

Given by a harmonic function: $V_{Improper} = k_\phi(\phi - \phi_0)^2$



Where the dihedral angle ϕ is the angle between planes ijk and ijl.

The dihedral angle ϕ is the angle between planes ijk and ijl

b. Non-Bonded interactions (Intermolecular terms)

Describe non-elecrostatic and electrostatic interactions between all pairs of atoms.

-The Lennard-Jones potential

Approximates the potential energy of non-elecrostatic interaction between a pair of non-bonded atoms or molecules:

$$V_{LJ}(r) = \frac{C12}{r^{12}} - \frac{C6}{r^{6}}$$

The r^{12} term approximates the strong Pauli repulsion originating from overlap of electron orbitals.

The r^{6} term describes weaker attractive forces acting between local dynamically induced dipoles in the valence orbitals.

The too steep repulsive part often leads to an overestimation of the pressure in the system.

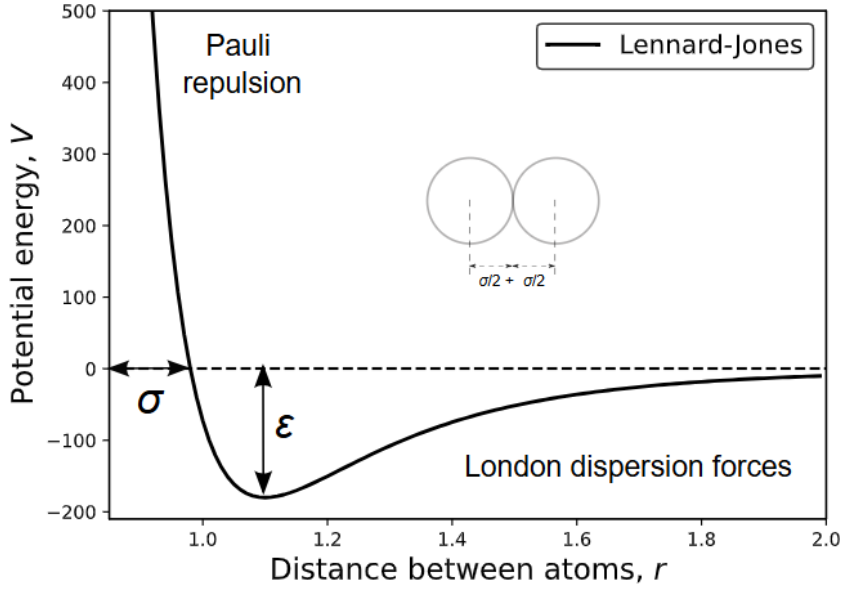


Figure3. The Potential Energy Surface

-The Buckingham potential

Replaces the repulsive r^{12} term in Lennard-Jones potential with exponential function of distance :

$$V_B(r) = Aexp(-Br) - \frac{C}{r^6}$$

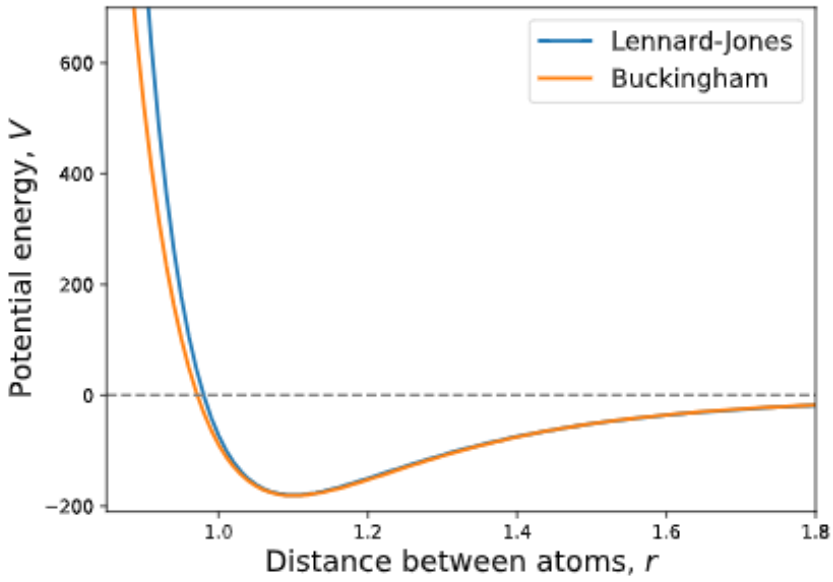


Figure 4. The Potential Energy Surface (**Buckingham potential)**

-The electrostatic potential

Point charges are assigned to the positions of atomic nuclei to approximate the electrostatic potential around a molecule. The Coulomb's law:

$$V_{Elec} = \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r_{ij}}$$

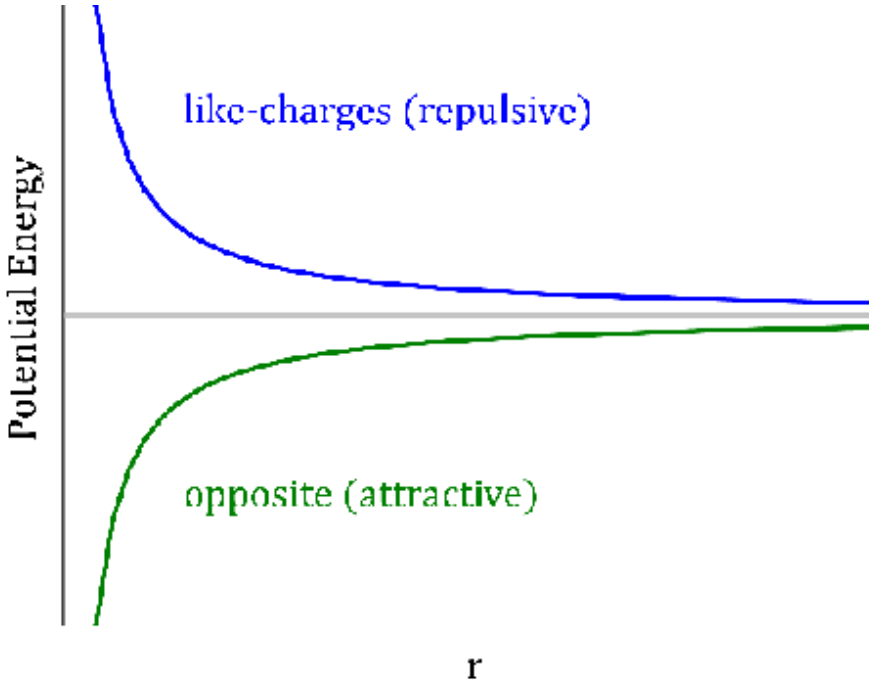


Figure 5. The electrostatic potential

5.1.Types of force fields

- Type one

-Bond stretching and angle bending is defined by simple harmonic motion (quadratic approximation).

-Correlations between bond stretching and angle bending are omitted.

Examples: AMBER, CHARMM, GROMOS, OPLS

AMBER (Assisted Model Building and Energy Refinement)

AMBER mentions both to a set of molecular mechanical force fields for the simulation of biomolecules and a package of molecular simulation programs. **AMBER** is a generally used suite of molecular dynamics software and a family of empirical force fields designed for simulating biomolecules like proteins, nucleic acids, and lipids.

CHARMM (Chemistry at HARvard Macromolecular Mechanics) is a both a set of force fields and a software package for molecular dynamics simulations and analysis.

GROMACS(GROningen MOlecular Simulation) is a force field for molecular dynamics simulation, and a related computer software package.

OPLS (Optimized Potential for Liquid Simulations) is a set of force fields developed by Prof. William L. Jorgensen for condensed phase simulations

- Type two

-Add anharmonic cubic and/or quartic terms to the potential energy for bonds and angles.

-Contain cross-terms describing the coupling between adjacent bonds, angles and dihedrals.

Examples: MMFF94, UFF

MMFF94 (**Merck molecular force field)** describe quite accurately electrostatic and van der Waals interactions, but that the magnitude of hydrogen bonding interactions are severely underestimated.

UFF (**Universal Force Field)** is a molecular mechanics force field designed for the full periodic table. The atom types are characterized by atomic number, hybridization and formal connectivity.

- Type three

-Explicitly add special effects of organic chemistry such as polarization, stereoelectronic effects, electronegativity effect, Jahn–Teller effect, etc.

Examples: During the 1990s the first general polarizable force fields appeared. The PIPF (polarizable intermolecular potential function), DRF90 and AMOEBA force fields are good examples of such developments.

5.2.Importance the force fields

-Molecular Simulations: Force fields enable simulation of molecular systems, providing insights into molecular dynamics, conformational changes, and interactions between ligands and their biological targets.

-Binding Affinity Prediction: They can simulate the ligand-bound and unbound states, which are crucial for predicting ligand binding affinities.

-Energy Minimization: Force fields are used in energy minimization procedures to find the most stable conformation of molecules.

-Molecular Dynamics: They allow for the simulation of molecular dynamics to study the behavior of molecules over time, which is essential for understanding protein folding, ligand binding, and other dynamic processes.

-Lead Optimization: Force fields help in optimizing lead compounds by predicting how modifications to the molecular structure affect binding and stability.

6. Energy-Minimizing Procedures

Energy minimization are characterized based on the derivative order employed to categorize a local minimum on the potential energy surface.

-Zero-order approaches rely exclusively on the evaluation of the energy function itself to find low-energy zones via grid searching.

- First-derivative strategies, such as the conjugate gradient and steepest descent methods, utilize the functional gradient.

-Second-derivative bases, typified by the Newton-Raphson algorithm, integrate the Hessian matrix to pinpoint energy minima.

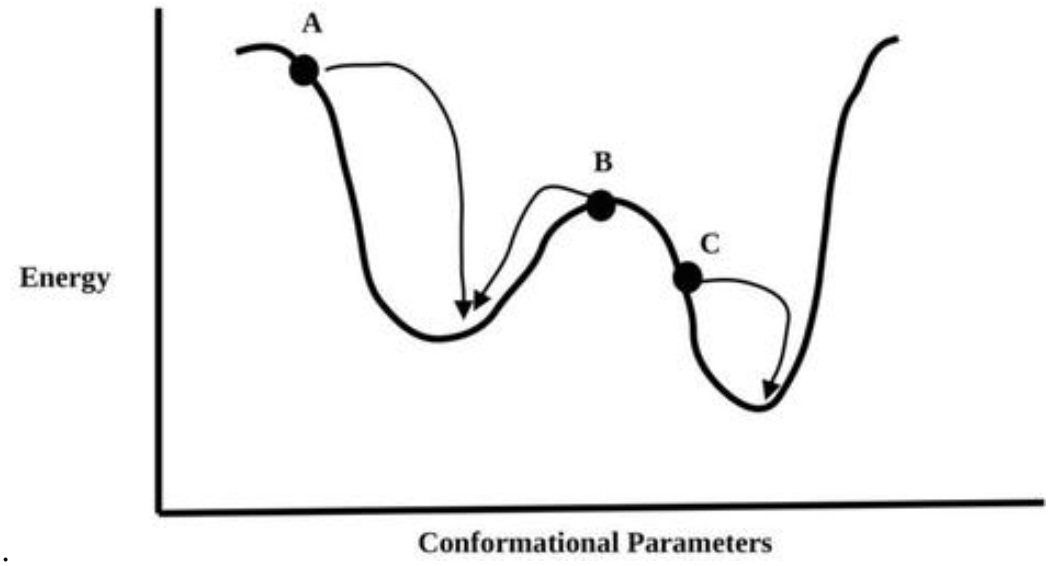


Figure 6. A one-dimensional energy surface showing minimization methods

Energy Minimization is very widely used in molecular modeling and is an integral part of techniques such as conformational search procedures. Energy Minimization is also used to prepare a system for other types of calculation.

For example, energy mini- Minimization may be used prior to a molecular dynamics or Monte Carlo simulation in order to relieve any unfavorable interactions in the initial configuration of the system.

7. Molecular mechanics methods Benefits:

-Calculating energy barriers between various conformers,

-Assessing PES steepness near local minimum points,

-Modeling the kinetics of protein folding,

-Resolving protonation and ionization equilibria,

-Mapping spatial coordination at enzyme active sites,

-Facilitating the targeted design of receptor binding sites.

8. Disadvantages of MM methods:

-each force field provides good results for a limited class of molecules.

-since empirical methods do not consider the electrons, they can not describe bond formation and breaking

-many molecular properties depending on subtle electronic effects are not reproducible by MM methods.

9. Results from MM methods

- Conformation and optimized structural coordinates.

-Characterized potential energy landscapes.

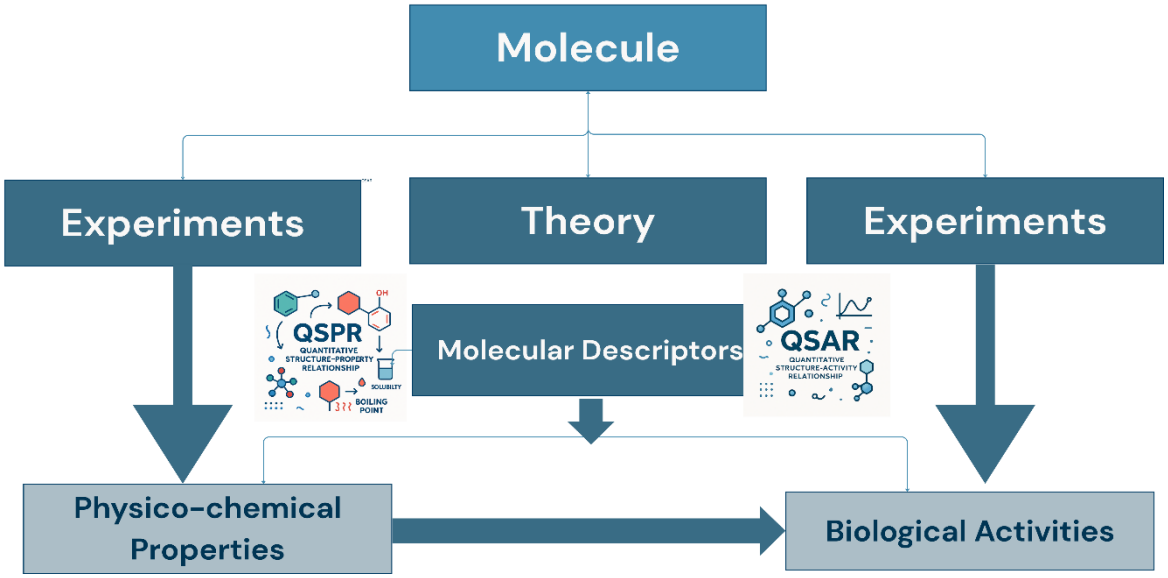
Macromolecular docking configurations.

-Theoretical vibrational spectrum profiles.

-Fundamental thermochemical constants and properties.

QSPR/QSAR Modeling

The structure of molecules contains the characteristics responsible for their physical, chemical, and biological properties, allowing us to understand their activity. Quantitative structure-activity/property relationships (QSAR/QSPR) are increasingly used, due to the growth of computing resources.



Structure-Activity Relationship (SAR) Method

Chapter 3: QSPR Modeling

1. Definition

Quantitative Structure-Property Relationship (QSPR) can be defined as a model that founds relationships between the molecular structure of compounds and their physicochemical properties, permitting estimates of compound behavior in biological systems and helping in the design of novel molecules with particular characteristics.

2. Types of Regression (Linear and multiple regression)

Linear regression analysis serves to evaluate the relationship linking a dependent variable to one or several independent variables. In instances focusing strictly on a single independent factor, the framework is formally classified as a simple linear regression model. Conversely, when the analytical setup incorporates multiple independent variables to predict the outcome, the framework is designated as a multiple linear regression model.

-If there is one independent variable, it is called simple linear regression.

-If there are multiple independent variables, it is called multiple linear regression.

a. Simple linear regression

The linear model

Consider a simple linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Where:

Y: Dependent variable

X: Independent variable

β_0: Intercept

β_1: Slope

ε: Error term

These parameters are generally called as **regression coefficients**.

The primary goal of regression analysis is to estimate the parameters β_0 and β_1, which denote the intercept and slope coefficient of the regression line. Within this framework, the error term ε is assumed to be an independent and identically distributed random variable possessing a mean of zero and a constant variance σ^2 . A subsequent assumption of normality is also applied to ε. visually; the association between the variables can be mapped using a scatter plot. In this representation, data points that cluster tightly around a straight line signify a robust linear relationship, whereas a higher degree of dispersion indicates a weaker correlation. To identify the optimal line of best fit, linear regression relies on the ordinary least squares (OLS) method.

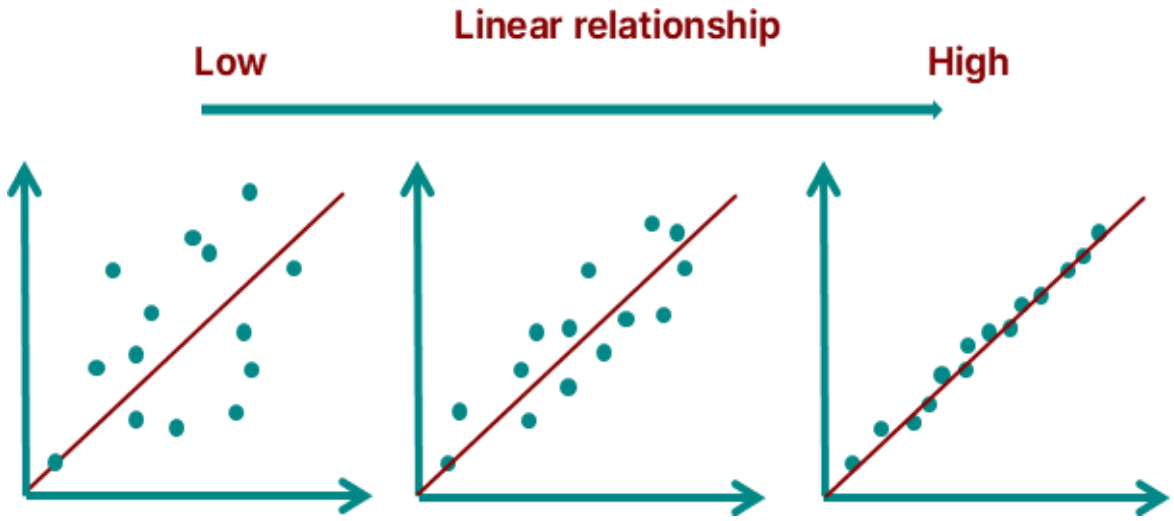


Figure1. The relationship between the variables

-Interpretation of Results

-Intercept: The value of Y when X is zero.

-Slope: The change in Y for a one-unit change in X.

-R^2: The proportion of the variation in Y explained by X.

-Statistical Significance: Assess the statistical significance of the regression coefficients.

-Example Simple Linear Regression

To quantify the relationships between molecular features and physicochemical attributes, researchers frequently pair linear regression models with molecular descriptors. This predictive framework facilitates the estimation of compound behaviors, thereby driving the rational design and comprehensive evaluation of novel drug candidates.

In our example, the Wiener Index(W) of alkanes is used to estimate their Boiling Point (BP).

$$BP = \beta_0 + \beta_1 W + \varepsilon$$

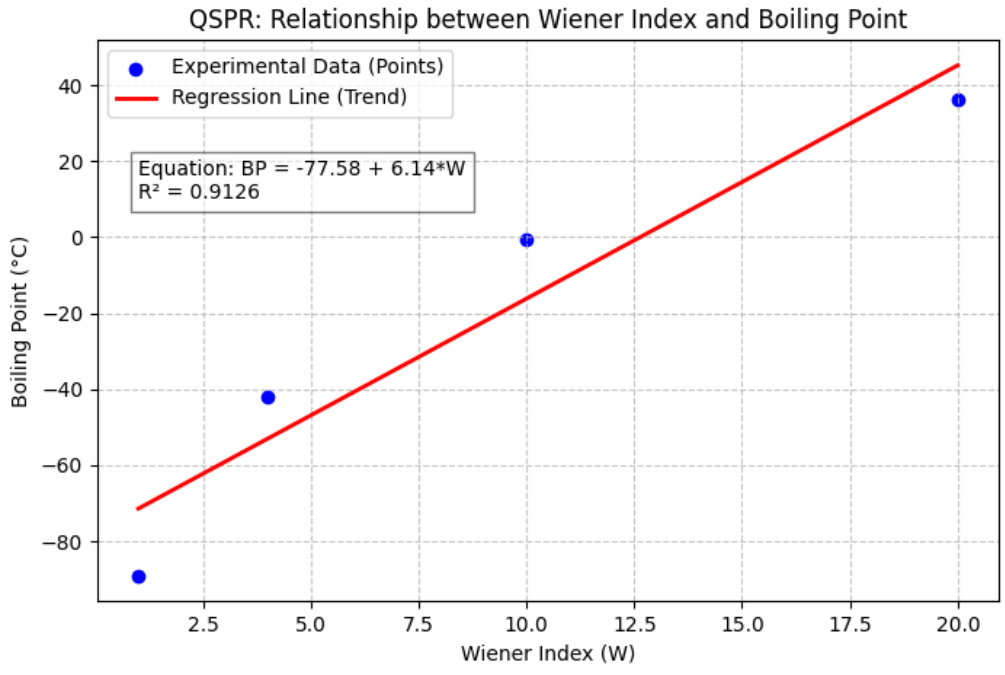


Figure 2. Example Simple Linear Regression -Using Python

b. Multiple Linear Regression

Multiple linear regression (MLR) serves as an extension of the simple linear model by incorporating a broader set of explanatory variables. This statistical approach is designed to evaluate and model the relationship between a single dependent criterion and a combination of two or more independent predictors. Consequently, the mathematical framework of the multiple linear regression model is expressed as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Where:

Y: Dependent variable

$X_1, X_2, \ldots, X_p$: Independent variables

β_0: Intercept

$\beta_1, \beta_2, \ldots, \beta_p$: Coefficients for each independent variable

ε: Error term

Multiple linear regression is generally used in the modeling of property statistics. We used to produce a linear model to describe the relationship between a physicochemical property and the molecular fingerprint c bits:

$$\text{Property} = \sum_{j=1}^{m} c_j f_j$$

Property is one of physicochemical properties.

cj is the contribution coefficient, which is determined by regression analysis;

fj is the binary bit of the jth fingerprint, with its presence or absence represented by the numeric values 1 or 0 respectively.

3. Coefficient of Determination R^2

To evaluate how well the regression model can predict or explain the dependent variable, two main measures are used: the coefficient of determination R^2 and the standard estimation error.

The coefficient of determination R^2, also known as explained variance indicates how much of the variance is explained by the independent variables.

To calculate R^2, the variance of the estimated values is related to the variance of the observed values:

$$R^2 = \frac{s^2_{\hat{y}}}{s^2_y}$$

Variance of the predicted values

Variance of the observed values

Adjusted R^2

The coefficient of determination R2 is influenced by the number of independent variables used. The more independent variables are included in the model, the greater the explained variance R2. To take this into account, the adjusted R2 is used.

$$R^2_{adj} = 1 - (1 - R^2) \cdot \frac{n-1}{n-p-1}$$

4. Assumptions of Linear Regression

Valid interpretation of linear regression outcomes requires that the underlying dataset satisfies several core statistical prerequisites:

Linearity: A straight-line, proportional relationship must exist between the independent and dependent variables.

-Homoscedasticity: The variance of the error terms or residuals must remain uniform across all levels of the predictor variables

-Normality: The error distribution must closely follow a standard normal Gaussian curve.

-Absence of Multicollinearity: Explanatory variables must not exhibit strong mathematical correlations with one another.

-Absence of Autocorrelation: Residual terms must be entirely independent, showing no serial correlation or self-dependence.

4. Modeling of physicochemical properties (QSPR)

Modeling of physicochemical properties consist of using computational techniques, to predict a compound's physical and chemical behavior from its molecular structure.

4.1.Principle of QSPR modeling

QSPRs are computer-assisted mathematical models, which relate the physico-chemical property of compounds to their chemical structure.

QSPR studies are performed on the basis of the correlation between the experimental values and descriptors which are derived from the molecular structure of respective compounds.

Usually, a QSPR model has the form of a mathematical equation:

$$\text{Property} = f(x_1, x_2, \ldots, x_n)$$

This equation relating particular topological, electronic, physicochemical, etc. molecular descriptors $x_1, x_2, \ldots, x_n$ to a property by means of a certain function(f) of n variables.

4.2.Methodology of QSPR modeling

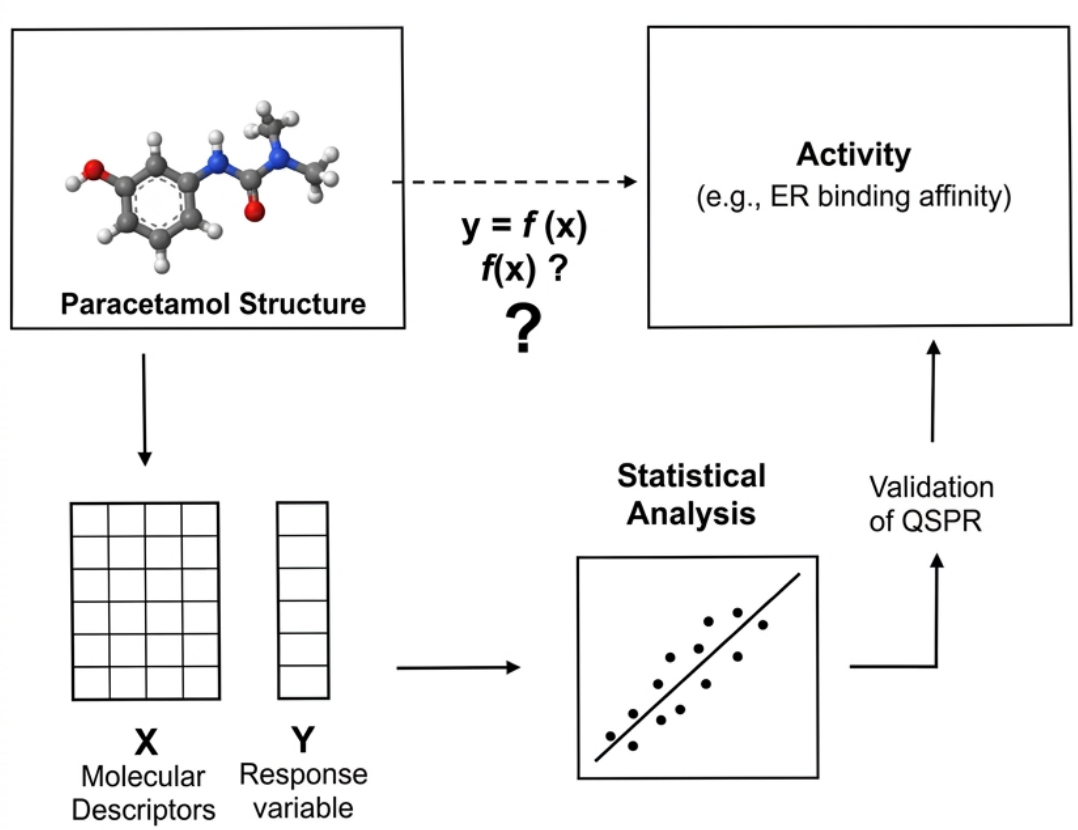
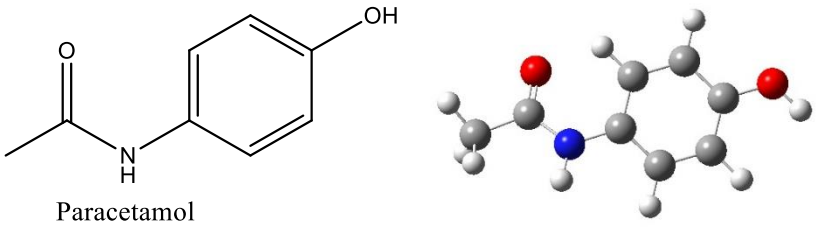


Figure 3. QSPR methodology to a common drug: Paracetamol.

Example: QSPR methodology to a common drug: Paracetamol.

Paracetamol, also chemically designated as N-acetyl-para-aminophenol or commonly referred to as acetaminophen, is a functionalized aromatic molecule. Classified as a non-narcotic analgesic and antipyretic agent, this therapeutic compound is widely prescribed to manage mild-to-moderate pain and reduce fever. Structurally, it features an acetamide group connected to a highly conjugated aromatic ring system. The molecular structure of paracetamol is illustrated below:



Paracetamol QSPR Analysis

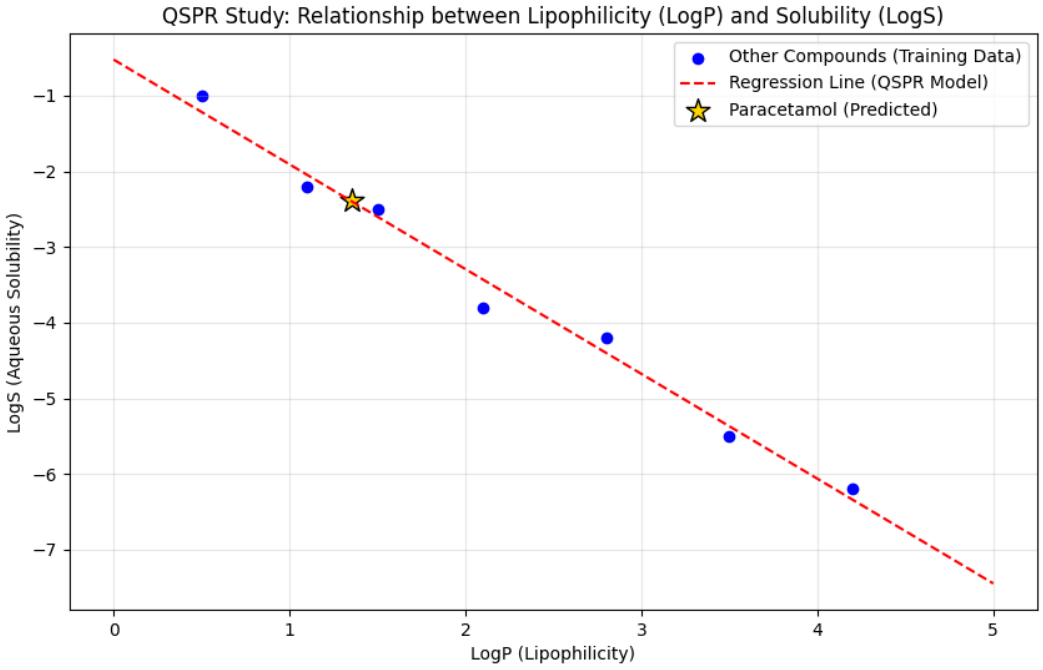


Figure 4. Paracetamol QSPR Analysis-Using Python

Chapter 4: QSAR Modeling

1. Introduction

The notion of QSAR has characteristically been used for drug discovery and development and has increased wide application for correlating molecular information with not only biological activities but also with other physicochemical properties, it found its way into the practice of agro chemistry, pharmaceutical chemistry, and eventually most facets of chemistry.

.

2. Definition

Quantitative structure-activity relationship (QSAR) modeling relates to the construction of analytical models of biological activities as a function of structural and molecular information of a compound library.

QSAR modeling is a computational method that establishes a mathematical relationship between the chemical structure of a compound and its biological activity.

The basic principle assumes that molecules with similar structures exhibit similar pharmacological effects, allowing biological responses to be predicted from structural information alone.

This makes QSAR a powerful tool for identifying new drug candidates even before synthesis.

3. Methodology for constructing the QSAR model

The key elements involved in constructing QSAR models are as follows (Figure 1):

-The initial step involves the classification of compounds with experimentally determined biological activity values.
-Ideally, these compounds belong to a congeneric series but should also exhibit sufficient chemical diversity to ensure significant variation in biological activity.

- Second, study of molecular descriptors that reflect various physicochemical and structural features of the molecules.
This stage entails identifying relevant descriptors that reflect crucial molecular properties and conducting a thorough investigation of their impact on biological activity.

- Establishing a correlation between 2D or 3D molecular descriptors and their respective biological activities to gain insights into the factors influencing variations in activity across the biological dataset.
This process helps us understand the key molecular features responsible for the observed differences in biological responses.

- Assessing the mathematical validity, statistical robustness, and predictive accuracy of the constructed QSAR model to confirm its reliability and effectiveness in analyzing biological activity.

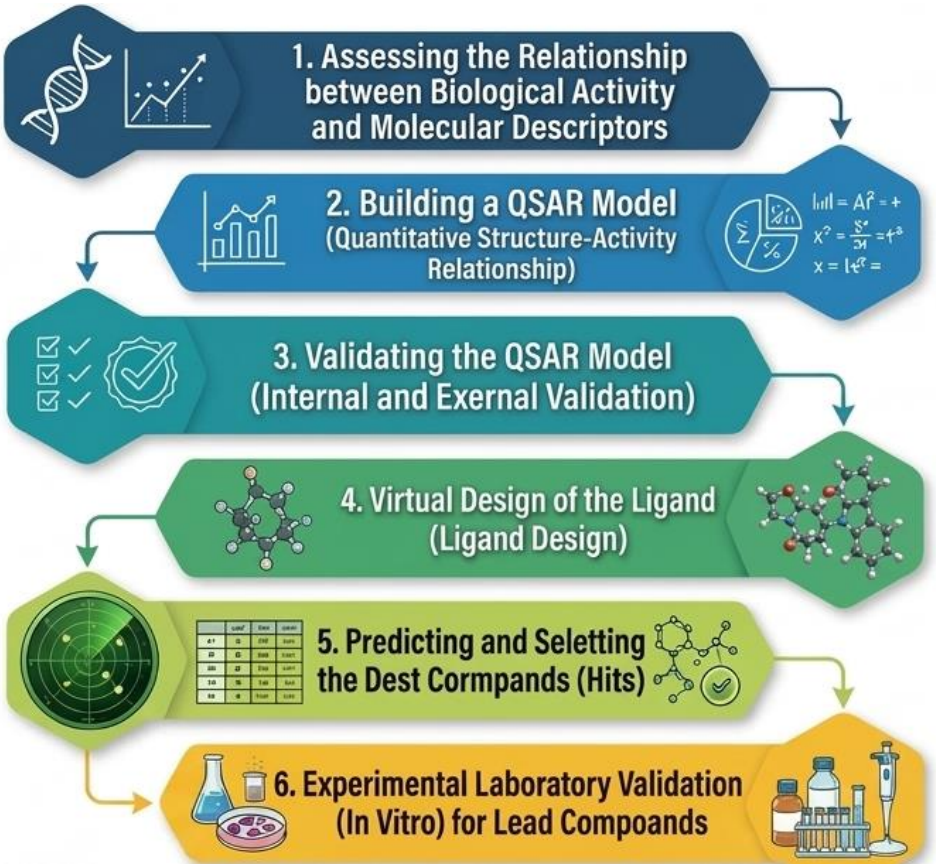


Figure 1. Methodology for constructing the QSAR model

4. Basic Principle of QSAR

QSAR is an equation that relates the biological activity of compounds to their physicochemical, steric, and structural properties, collectively known as molecular descriptors. It can be used as an alternative to structure-based methods such as molecular docking when the 3D structure of the target protein is unknown.

A classical QSAR model generally follows the mathematical form:

$$\textbf{Biological Activity} = f(\text{n})\ \textbf{(Descriptors)}$$

$$\textbf{Activity} = \textbf{f(Physicochemical Descriptors + Structural Descriptors)}$$

Where:

-*Activity* = biological response

- *Descriptors* = numerical representation of molecular features

-*f* = regression or machine-learning technique

QSAR equations permit researchers to predict the activity of unapproved molecules and guide structural modifications for improved potency or lower toxicity.

5. QSAR methods

Many different approaches to QSAR have been established since Hansch's seminal works.

QSAR methods can be analyzed from:

(1) The types of **structural parameters** that are used to describe molecular identities starting from different representation of molecules, from simple chemical formulas 1D to 3D conformations.

(2) **The mathematical procedure** that is employed to obtain the quantitative relationship between these structural parameters and biological activity. A schematic representation of the QSAR methods is showed in Figure 1

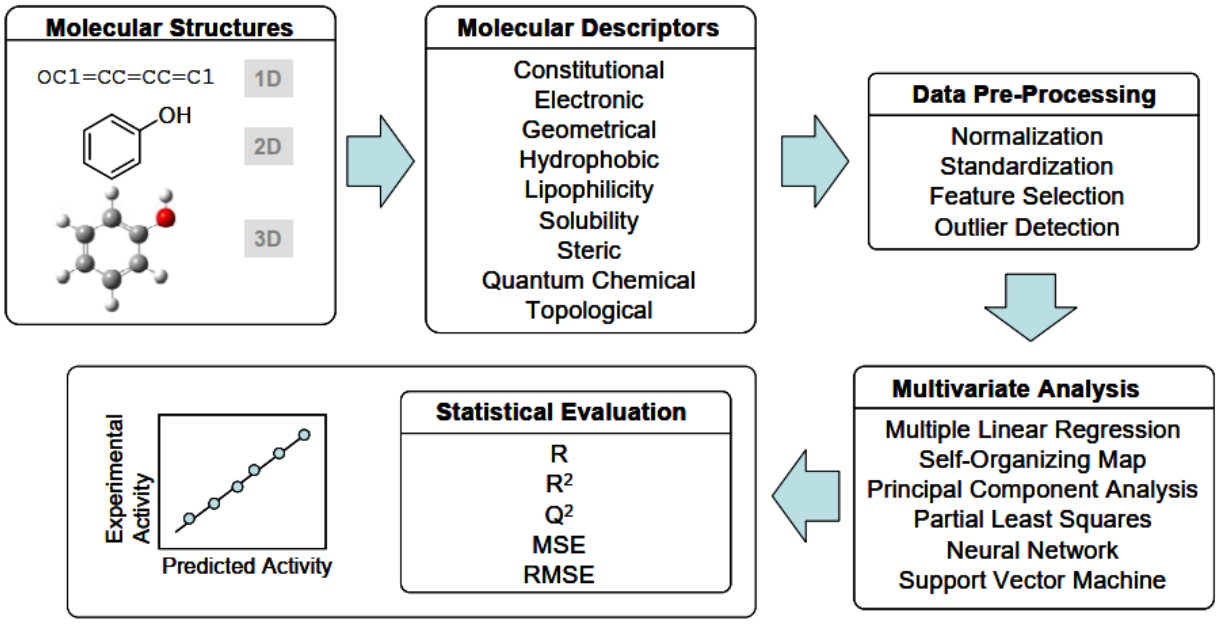


Figure 2. Schematic overview of the QSAR methods.

6. Classifications of Quantitative Structure-Activity Relationships (QSAR)

1. **1D-QSAR**: Correlates biological efficacy strictly with global, one-dimensional physicochemical metrics, including molecular weight, pKa, or logP values.

2. **2D-QSAR**: Frequently termed classical QSAR, this methodology evaluates molecular structural descriptors like fragmentation inputs, steric parameters, electronic factors, and hydrophobicity via Hansch approaches. Predictive modeling typically relies on Free–Wilson analysis or Multiple Linear Regression (MLR).

3. **3D-QSAR**: This approach examines spatial molecular attributes by deploying probe-based sampling throughout a defined molecular lattice, subsequently mapping these three-dimensional characteristics against target biological responses.

4. **4D-QSAR**: Expands on 3D frameworks by integrating spatial orientation and conformational flexibility as a fourth dimension, which is achieved by calculating ensemble averages across training datasets.

5. **5D-QSAR**: Introduces a fifth variable dimension that permits the simultaneous simulation and representation of up to six distinct induced-fit receptor-ligand conformations.

6. **6D-QSAR**: Represents the highest dimensional, enabling researchers to concurrently examine and evaluate multiple diverse solvation models during the modeling process.

7. Molecular descriptors

Biomolecular descriptors are numerical representations that capture structural, physicochemical, or electronic characteristics of a molecule. They are created using welldefined algorithms applied to molecular representations or acquired via established experimental approaches. These descriptors operate as essential, independent factors in QSAR

models, allowing for the prediction of molecular characteristics and biological activities. The primary types of theoretical molecular descriptors are:

– (0)D-descriptors, also referred to as constitutional descriptors or count descriptors, are the most fundamental type of molecular descriptors.

- (1)D-descriptors, also known as structural fragment descriptors or fingerprints, represent a molecule as a collection of predefined structural fragments or molecular subunits.
- (2)D-descriptors, also known as graph invariants, characterize molecular structures based on their two-dimensional topology.
- (3)-D-descriptors describe molecular properties based on the three-dimensional spatial arrangement of atoms. These descriptors account for molecular shape, size, steric effects, and electronic properties.
- 4D-descriptors extend beyond static three-dimensional representations by incorporating conformational flexibility and environmental factors.

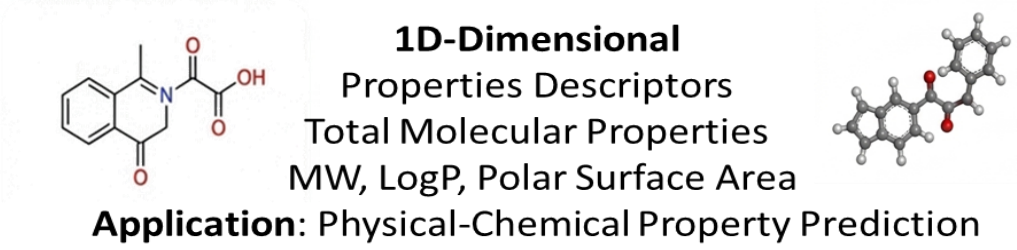
Figure 4 illustrates four primary categories of molecular descriptors utilized in molecular modeling and computer-aided drug design.

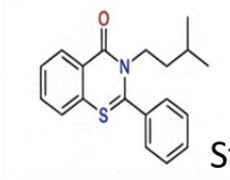
The first category, Physicochemical Descriptors includes fundamental properties such as the partition coefficient log P, Molecular Weight (MW), and Polar Surface Area (PSA).

The second category consists of Structural Descriptors, which focus on the 2D and 3D connectivity of the molecule, represented here by topological indices and skeletal arrangements.

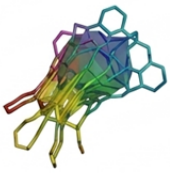
The third category covers Quantum Chemical Descriptors, which describe electronic properties derived from quantum mechanics, specifically showcasing the Highest Occupied Molecular Orbital (HOMO), Lowest Unoccupied Molecular Orbital (LUMO), and the dipole moment.

Finally, the fourth category details Geometrical Descriptors, which define the spatial and three-dimensional characteristics of the molecule, including its molecular volume and surface area.

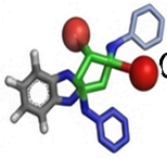




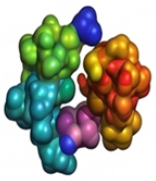
2D-Dimensional:
Structural Descriptors:
Structural Connectivity and Patterr
Substructure and Fragment Fingerprints



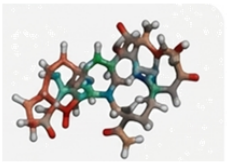
Application: Substructure Search & Graph-Based Methods



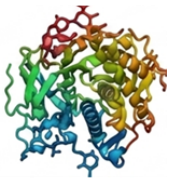
3D-Dimensional
Conformation & Bioactivity Descriptors
3D Shape and Electrostatic Properties
Pharmacophore Modeling



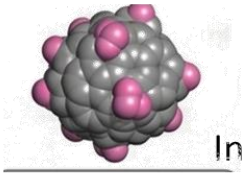
Application: Pharmacophore Search, De Novo Design (CoMFA/CoMSIA)



4D-Dimensional
Conformation Ensemble Descriptors
Molecular Dynamics Trajectories
Multiple Conformations Analysis



Application: Ensemble Pharmacophore and Receptor Modeling



5D-Dimensional
Induced Fit & Protein Flexibility

Descriptors: Representing Conformational Selection & Flexibility

Application: Induced Fit Docking and Response-Ensemble Analysis

Figure 3. Classifications of QSAR Models

8. Statistical resources used for QSAR model

The success of any QSAR model largely relies on choosing the right molecular descriptors and establishing a strong statistical relationship between them and the specific biological activity of interest. Since the inception of QSAR studies, it has been recognized that representing molecular descriptors is a fundamental and crucial aspect of the methodology. The development of a QSAR model can refine a vast number of molecular descriptors generated by modern, advanced software. The three main statistical methods that are used in the linear framework of the QSAR investigation are:

- Multivariable linear regression analysis (MLR).
- Principal component analysis (PCA).
- Partial least square (PLS).

Principal Component Analysis (PCA) was used to overcome the difficulties of MLR investigations by extracting information from several, sometimes redundant variables and translating them into a smaller collection of uncorrelated variables. Thus, PCA is an efficient and effective way to reduce the number of independent variables utilized in QSAR model construction. This method works especially well for models that have a lot of chemical variables compared to forecasts. However, the findings of PCA are sometimes difficult to understand, making it difficult to pinpoint precise physicochemical and structural properties required for biological function.

Multivariable linear regression (MLR) is a basic and easy approach for discovering chemical descriptors with substantial correlations to changes in biological activity.

When making a QSAR model, the MLR method can include either forward or backward stepwise regression.

Statistical tests are used to change a stable model by adding or removing chemical descriptors one at a time until the best model is found. But when there are a lot of molecular descriptors, MLR might not work as well, so researchers have to be careful to leave out variables that have a strong internal correlation. Under certain situations, researchers may use statistical methods and software to automate the MLR process, but constant control remains necessary.

Additionally, PCA and MLR are put together to create the partial least squares (PLS) method, which adds biological activity as a dependent variable to make the correlation and connection stronger.

When multiple dependent variables are involved, the PLS method proves to be highly useful and advantageous. It is an iterative regression technique that derives solutions by linearly transforming a large set of original descriptors into a smaller set of new orthogonal terms known as latent variables. Consequently, this method is recognized as a standard statistical approach. In general, the standard QSAR model equation is written as:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 \ \ldots . + pX_p\alpha$$

where Y is the biological activity or response and x_1, x_2,..., x_p are molecular descriptors or variables, α is the regression coefficient and x represents the input variables (descriptors). Model validation can be conducted using internal and external validation techniques. Validation techniques are critical for evaluating a model's prediction abilities on unknown data and identifying the proper complexity of an equation depending on the current dataset.

Either internal techniques which depend on the same data used for model development or external methods, which utilize a different dataset for evaluation can be verified using the QSAR model.

9. Different phases of QSAR model validation

a. Internal Validation

Internal validation assesses the model's performance against the same dataset used during development. These approaches determine the reliability and robustness of QSAR models by evaluating how well the model matches the training data. One frequently employed internal

validation technique is Pearson's correlation coefficient (r), which quantifies the degree of association between two variables. The r value varies from -1 to +1, with -1 indicating a negative (inverse) correlation and +1 representing a positive (direct) connection. To rate how well a QSAR model can predict, Pearson's correlation coefficient (r) is used to look at how closely actual (x) and expected (y) numbers match up.

This helps to identify the level of fluctuation between the two variables.

The correlation coefficient (r) is computed using the following equation:

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}}$$

r_{xy} is the correlation coefficient among variables x and y, n is the sample size, x is the individual value of variable x, y is the individual value of variable y, xy is the product of variables x and y, and x2 and y2 are the squared values of factors x and y, respectively.

Root Mean Squared Error (RMSE) is a widely used metric for evaluating the relative error of a QSAR model. It quantifies the difference between experimental and predicted values, providing insight into the model's predictive accuracy. RMSE is calculated using the following formula:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(x-y)^2}{n}}$$

RMSE represents the root mean squared error, where x denotes the experimental value of the activity or property of interest, y signifies the predicted value of the activity or property, and n refers to the dataset's sample size.

b. Cross validation

A commonly used method for internal validation of a QSAR model is cross-validation (CV, Q^2, q^2). This process involves performing regression multiple times on different subsets of data. Typically, each molecule is systematically left out once, and the R value is calculated based on the predicted values of the omitted molecule. The coefficient of multiple determination (R^2) is used to evaluate the goodness-of-fit of a QSAR model, measuring how well the model explains the variance in the dataset .

R^2 values vary from 0 to 1, and higher values indicate better model fit. A QSAR model is regarded significant if its R^2 value is 0.6 or higher. Values closer to 1 suggest a greater correlation between anticipated and experimental data.

The R^2 coefficient is determined using the following formula [58]:

$$R^2 = \frac{MSS}{TSS} = 1 - \frac{RSS}{TSS}$$

Where:

- MSS (Model Sum of Squares) represents the variation explained by the QSAR model.
- RSS (Residual Sum of Squares) is the sum of the squared differences betweenexperimental and predicted responses for the training set.
- TSS (Total Sum of Squares) is the total variation, calculated as the sum of the squareddifferences between experimental responses and their mean values.

With:

$$MSS = \Sigma_i (\hat{y}_i - \bar{y})^2$$

$$TSS = \Sigma i (yi - \bar{y})2$$

$$RSS = \Sigma i (yi - \hat{y})2$$

$\bar{y}$: mean value of the dependent variable, yi: observed dependent variable and $\hat{y}i$: calculated dependent variable.

The cross-validated squared correlation coefficient (LOO-R^2CV or Q^2) is a measure of the predictive performance of a QSAR model using leave-one-out cross-validation (LOO-CV). It is calculated using the following formula:

$$R^2_{CV}\,(Q^2) = 1 - \frac{\Sigma(Ypred - Yobs)^2}{\Sigma(Ypred - Ymean)^2}$$

Where y^{pred}, yobs, and $\bar{y}$ represent the predicted, observed, and mean values of the target property (biological activity), respectively. The cross-validated squared correlation coefficient (R2CV or Q2) assesses the model's predictive reliability. A QSAR model is generally considered acceptable if R^2CV (Q^2) exceeds 0.5, indicating a reasonable predictive capability.

c. Y-Randomization test

The Y-randomization technique is a commonly employed method to verify the robustness of a 3D-QSAR model. It helps assess whether the model's predictive performance is genuinely significant or if it has been influenced by chance correlations. The Yrandomization technique ensures that the developed QSAR models are not generated by chance. In this method, the activity values (Y) are randomly shuffled while keeping the descriptor values unchanged, and new models are constructed. For a model to be considered reliable, the coefficient of determination (R^2) from the training phase and the cross-validated R^2 (R^2_{cv}) of the randomized models should be significantly lower than those of the original, non-randomized models.

d. External Validation

The original dataset is divided into two subsets, a training set and a test set, during this validation process. Various methods can be used for this partitioning; for more details, see the report from the QSAR validation workshop. One such method involves selecting a fixed proportion of a homogeneous dataset from the original data to form the test set, while the remaining data is designated as the training set. If the training set comprises two-thirds of the original data and the test set one-third, the data can be ranked based on the magnitude of the biological response. Alternative approaches can also be employed for selecting the training set, such as considering relevant physicochemical properties. The training set serves as the foundation for developing a predictive model, which is then utilized to estimate the activity of the test set members with greater accuracy. *Golbraikh* and *Tropsha* introduced the external validation statistical criteria, which are defined using the following formulas:

$$R^2_{pred} = 1 - \frac{\Sigma(Y_{pred} - Y_{obs})^2}{\Sigma\left(Y_{pred} - \overline{Y}_{mean\,(train)}\right)^2}$$

Where ($test$) and ($test$) representing the observed and predicted activity values for the test set and Y_{mean}(train) is the average value for the dependent variable for the training set. It was mentioned that when $R^2_{pred} > 0.6$, it is considered a good indicator of strong external predictability.

And:

$$r_0^2 = 1 - \frac{\sum(Y_{test(pred)} - kY_{test(pred)})^2}{\sum(Y_{test(pred)} - \overline{Y}_{test(pred)})^2} \quad < 0.1 \ ,$$

$$r'^2_0 = 1 - \frac{\sum(Y_{test} - kY_{test})^2}{\sum(Y_{test} - \overline{Y}_{test})^2} \quad < 0.1$$

$$K = \frac{\sum(Y_{test} - kY_{test(pred)})}{\sum(Y_{test\ (pred)})^2} \quad , \quad 0.85 \leq K \leq 1.15$$

$$K' = \frac{\sum(Y_{test} - kY_{test(pred)})}{\sum(Y_{test})^2} \quad , \quad 0.85 \leq K' \leq 1.15$$

Represent the coefficients of determination through the origin for predicted versus observed activity, and observed versus predicted activity of the test set, respectively. The parameters k and k' denote the slopes of the regression lines through the origin. Furthermore, according to Roy's statistical criteria.

$$\frac{(r^2 - r_0^2)}{r^2} < 0.1 \quad \text{or} \quad \frac{(r^2 - r'^2_0)}{r^2} < 0.1$$

$$|r^2 - r_0'^2| < 0.3$$

To validate the predictive ability of the model, it is important to calculate the difference between the values of $r2$, $r02$ and $r'0\ 2$ according the following equations:

$$r_m^2 = r^2(1 - \sqrt{(r^2 - r_0^2)}$$

$$r'^2_m = r^2(1 - \sqrt{(r^2 - r'^2_0)}$$

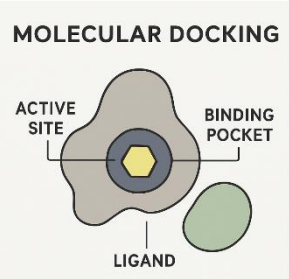
$$\overline{r}^2_m = r_m^2 - r'^2_m$$

Where $r02$ is the squared correlation coefficient without intercept $r'm\ 2$ the same as $rm2$, with

the x and y axes exchanged.

Also, to further validate the predictive capacity of the developed model, the calculation of $\Delta r2m$ and $r'2m$ is obligatory using the following formulas:

$$\Delta r_m^2 = \left|r_m^2 + r'^2_m\right| = \frac{r_m^2 + r'^2_m}{2}$$



Chapter 5: Molecular Docking

1. Introduction

Since its beginning in the mid-1970s, molecular docking has established an indispensable basis in structure-based drug design (SBDD), submission critical insights into the precise binding mechanisms between therapeutic agents and their biological targets. Through the deployment of these computational simulations, researchers can systematically refine chemical structures to enhance both pharmacological potency and target selectivity. Consequently, this methodology not only accelerates the initial identification of promising lead hits but also serves as a cornerstone for the rational design and optimization of clinically valuable compounds.

2. Definition

Molecular docking is a computational technique widely used in structural molecular biology and computer-aided drug design. Its primary purpose is to predict how a small molecule (ligand) binds to a protein whose three-dimensional structure is known. Effective docking methods explore multiple binding conformations and employ scoring functions to rank potential binding modes. This approach is particularly useful for virtual screening of large compound libraries, guiding the identification of molecules that may inhibit a target protein. Proper preparation of the input structures is critical, and careful analysis of the results from stochastic search algorithms is necessary to interpret docking outcomes accurately.

3. Types of molecular docking

3.1. Rigid docking

Rigid/Lock and Key Docking is based on Emil Fischer's lock and key theory (1894), which suggests that enzyme specificity toward its substrate relies on complementary geometric shapes, allowing them to fit together precisely, much like a lock and key. In this rigid docking

approach, the internal geometry of both the ligand and receptor remains unchanged, meaning no conformational flexibility is considered during the docking process.
If structural rigidity is an important part of how ligands and receptors interact (**Figure 1, (a)**), this method can help predict high-affinity binding.

3.2. Flexible docking

Based on the induced-fit model, flexible docking (**Figure 1, (b)**) lets you guess a lot of different binding positions for protein-ligand, protein-protein, and peptide-protein interactions. When you use rigid docking, the structures don't change. But when you use flexible docking, the structures do change because of the ligand.

This is especially true for the side chains of residues in the target protein's binding site. This approach utilizes various receptor conformations, resulting in highly multidimensional data sets that better capture dynamic chemical interactions.

The method also lets you look at the potential energy surface over a number of different coordinate functions, which takes into account the fact that flexible protein structures have a lot of degrees of freedom. This makes it easier to predict binding affinity and find new drugs.

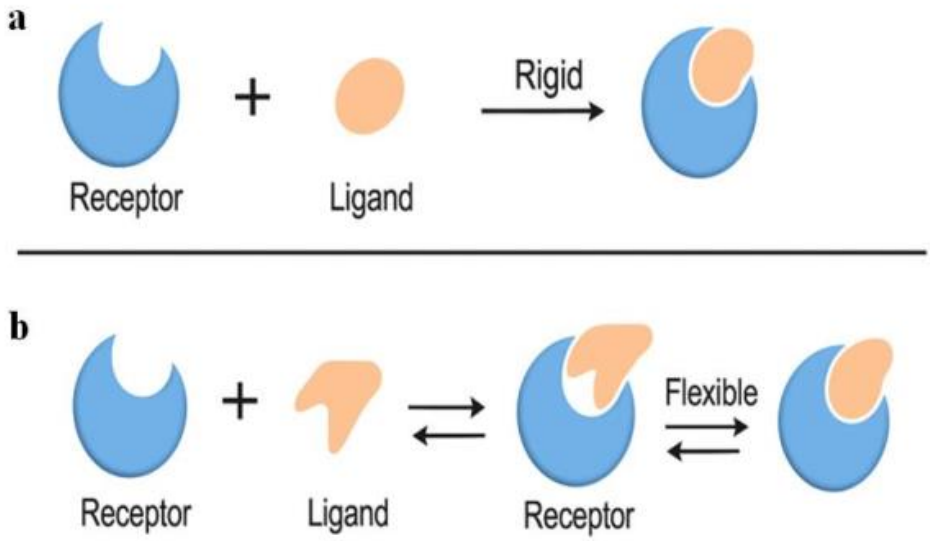


Figure 1. molecular docking approach. (a) A lock-and-key model (Rigid docking) and (b) Induced fit model (Flexible docking).

3.3. Semi-flexible docking

Semi-flexible docking is an intermediate approach where protein conformation varies to a limited extent, making it suitable for studying protein-ligand interactions.

In this method, the ligand stays flexible, letting it rotate freely around certain torsional angles. On the other hand, the protein structure stays mostly rigid or lets some amino acid side chains move around. Unlike rigid docking, where both molecules remain unchanged, semi-flexible docking balances structural adaptability with computational efficiency.

This method works especially well with small molecule ligands because it records changes in ligand conformation while keeping the protein framework mostly stable.

4. Basics of molecular docking

Docking is often used to predict the alignment of small-molecule medicinal drugs with protein targets, therefore assessing binding affinity and activity. It is critical in rational drug design, and substantial attempts have been made to develop docking algorithms for greater accuracy. Docking is a computer process that identifies the optimal orientation of one molecule relative to another, resulting in the development of a stable complex.

This approach is critical for drug structural characterisation since it helps to optimize both the shape of the ligand and protein, as well as their relative orientation, in order to reduce the system's free energy, resulting in better binding predictions for drug development.

Molecular docking is made up of two interrelated steps:

-a searching algorithm that investigates various ligand conformations inside the protein's active site.

-a scoring function that scores these conformations according to their binding affinity.

Ideally, searching algorithms should faithfully duplicate the experimentally observed binding mode, while scoring functions should give this mode the highest rank out of all potential conformations. The search method in molecular docking finds the location and flexibility of the ligand, and the scoring function checks the strength and stability of the binding.

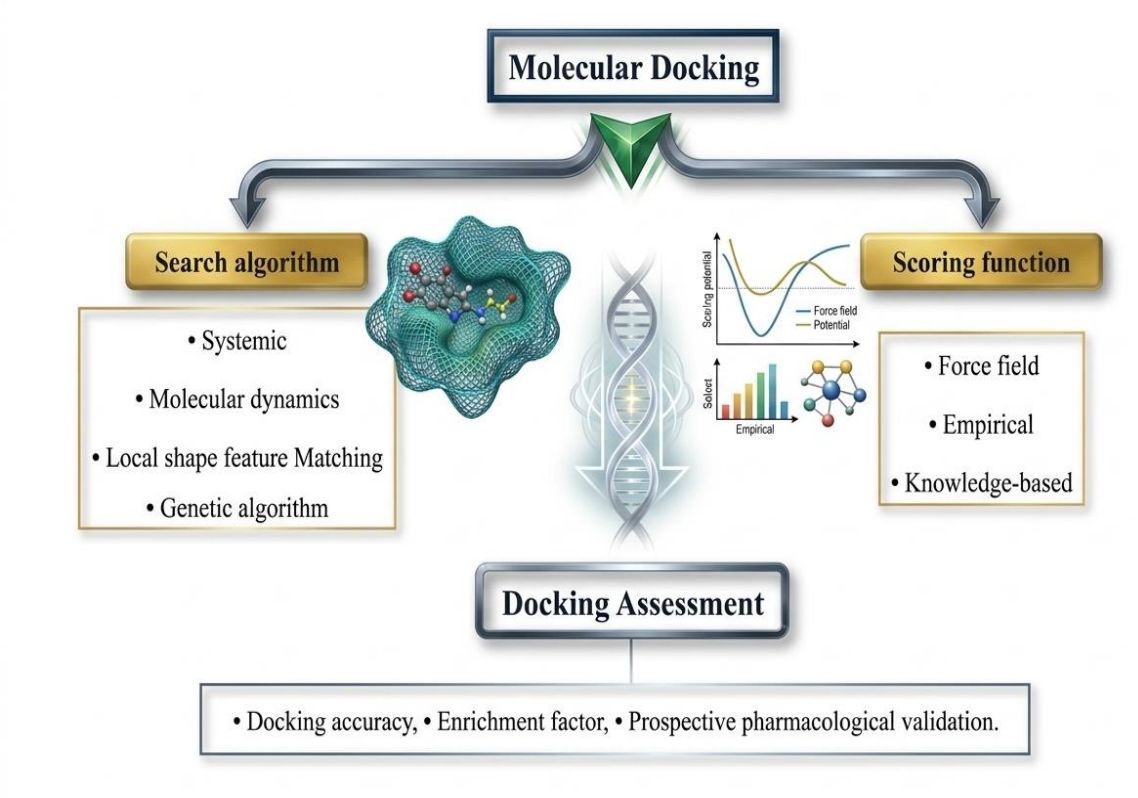
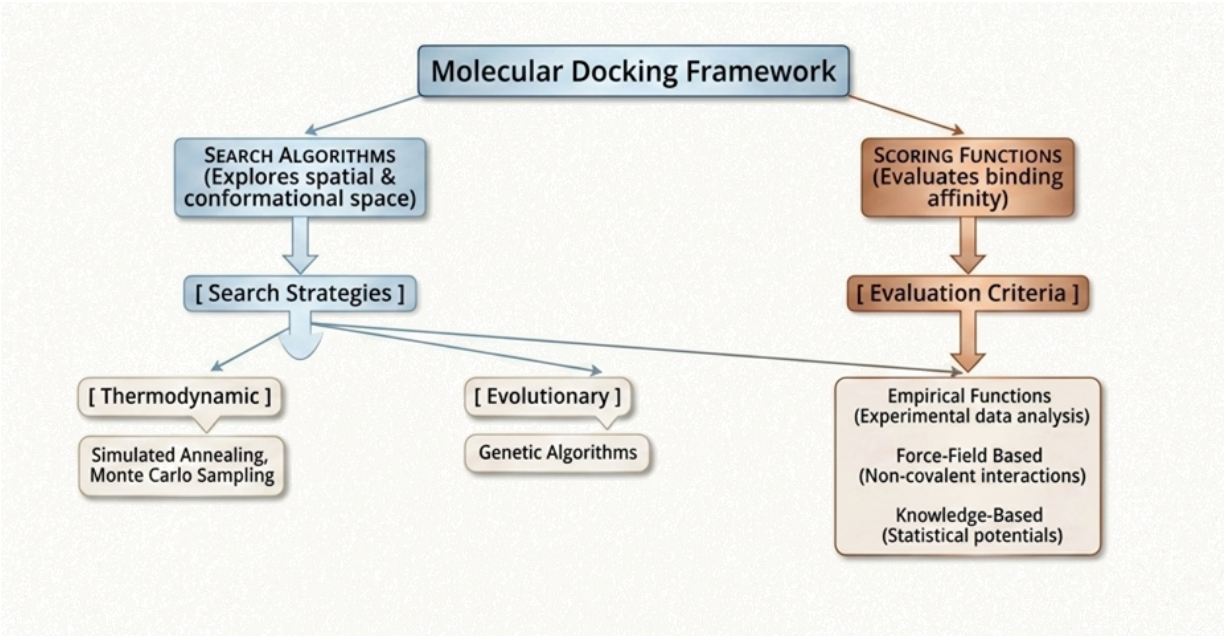


Figure 2. Basics of molecular docking

5. Principles of Molecular Docking Techniques

The basic principle of molecular docking consists of two key components: the docking process and the evaluation of binding strength through various scoring functions (see Figure 3).



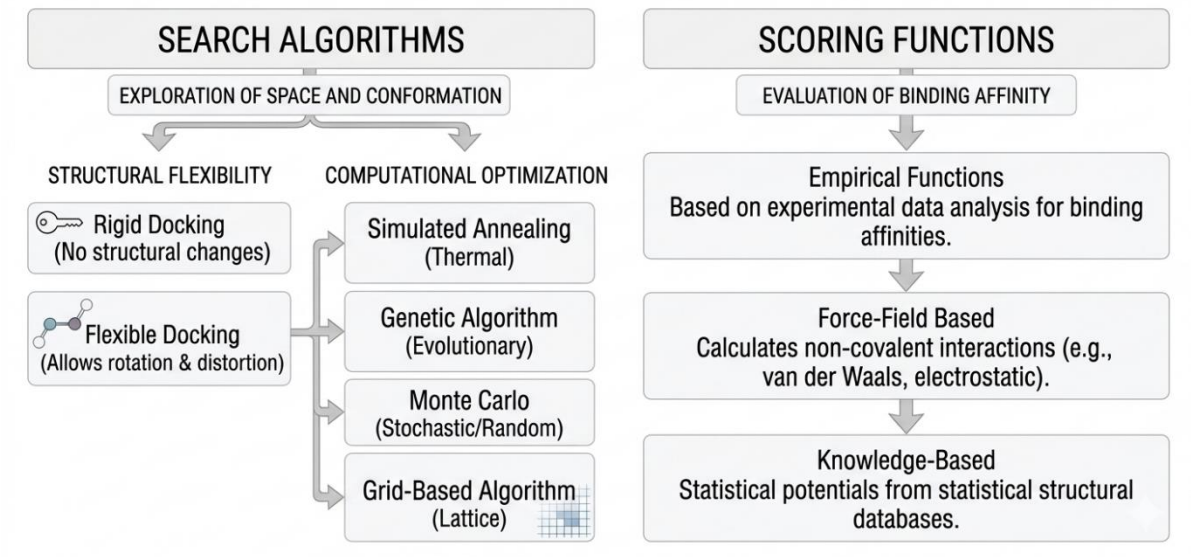


Figure 3. Principles of Molecular Docking Techniques

6. Components of Molecular Docking

a. Ligand and Receptor

-Ligand:

Ligand is a small molecule that interacts with a target molecule, frequently a drug molecule or a candidate molecule.

-Receptor:

A receptor is a target molecule, frequently a protein, enzyme, or nucleic acid, that binds to a ligand and can form a complex with the ligand through specific chemical bonds and non-covalent interactions.

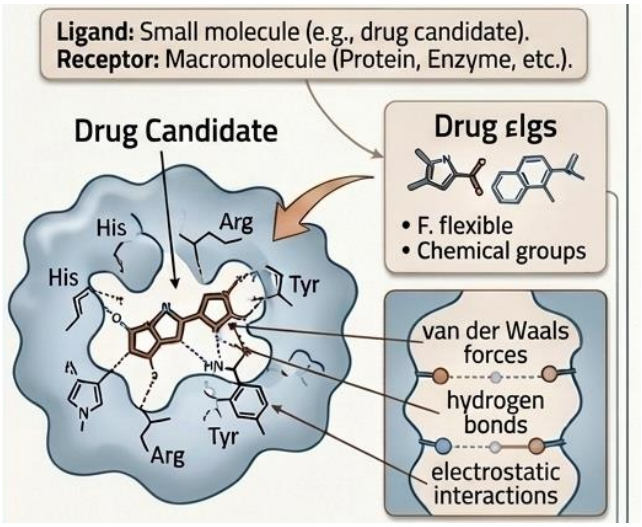


Figure 4. Ligand and Receptor

b. Binding Sites

Binding Sites are precise regions of the receptor molecule that predicament to the ligand. This region is typically composed of a group of amino acid residues that can bind to ligands through van der Waals forces, hydrogen bonding, electrostatic interactions, etc.

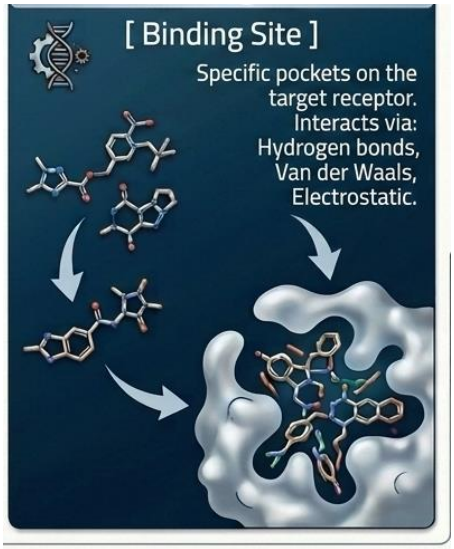


Figure 5. Binding Sites

c. Docking Scores

Docking Scores are numerical pointers used to compute the binding affinity of ligands to receptors. This score is based on computer-simulated docking results that assess the stability and affinity of the ligand binding to the receptor.

7. Molecular Dynamics (MD) simulation docking

Molecular dynamics (MD) simulation docking methods detention the dynamic interaction between ligand and receptor by simulating the movement of molecules over a certain period.

For example, researchers to predict how an antiviral drug binds to a coronavirus protein. To see if this connection is stable, they run a 100-nanosecond Molecular Dynamics (MD) simulation. This simulation places the drug-protein complex in a virtual box filled with water and salts to mimic the human body.

8. Advantages and Challenges of Molecular Docking

-Advantages of Molecular Docking

Molecular docking serves as a cornerstone in rational drug design and structural biology due to several distinct benefits:

-High-Throughput Optimization: It accelerates the screening of expansive chemical libraries against target proteins.

-Resource Efficiency: The technique significantly curtails the financial expenditures and time constraints of traditional wet-lab assays.

-Structural Insights: It yields atomistic-level data on complex configurations.

-Lead Advancement: The generated binding poses offer predictive guidelines for optimizing lead compounds.

-Versatile Probing: The method evaluates diverse macromolecular events, including both protein-protein and ligand-receptor interactions

-Challenges in Molecular Docking

Despite its utility, computational docking is constrained by several intrinsic methodological bottlenecks:

-Scoring Inaccuracies: Scoring functions frequently struggle to accurately compute binding free energies.

-Conformational Sampling: Simulating full macromolecular flexibility remains a persistent obstacle.

-Solvation Effects: Accounting for the thermodynamic contributions of water molecules often yields imprecise results.

-Computational Cost: Executing advanced flexible docking protocols or molecular dynamics simulations requires high computational power.

-Throughput Trade-offs: The long computation times limit the viability of these methods for massive virtual libraries.

-Screening Errors: Virtual screening campaigns remain prone to false positives and false negatives, resulting in discrepancies between in silico predictions and in vitro data.

9. Key Steps in Molecular Docking

Docking comprises discovering the most favorable binding modes of a ligand to the target of interest. The key steps are represented in Figure 6.

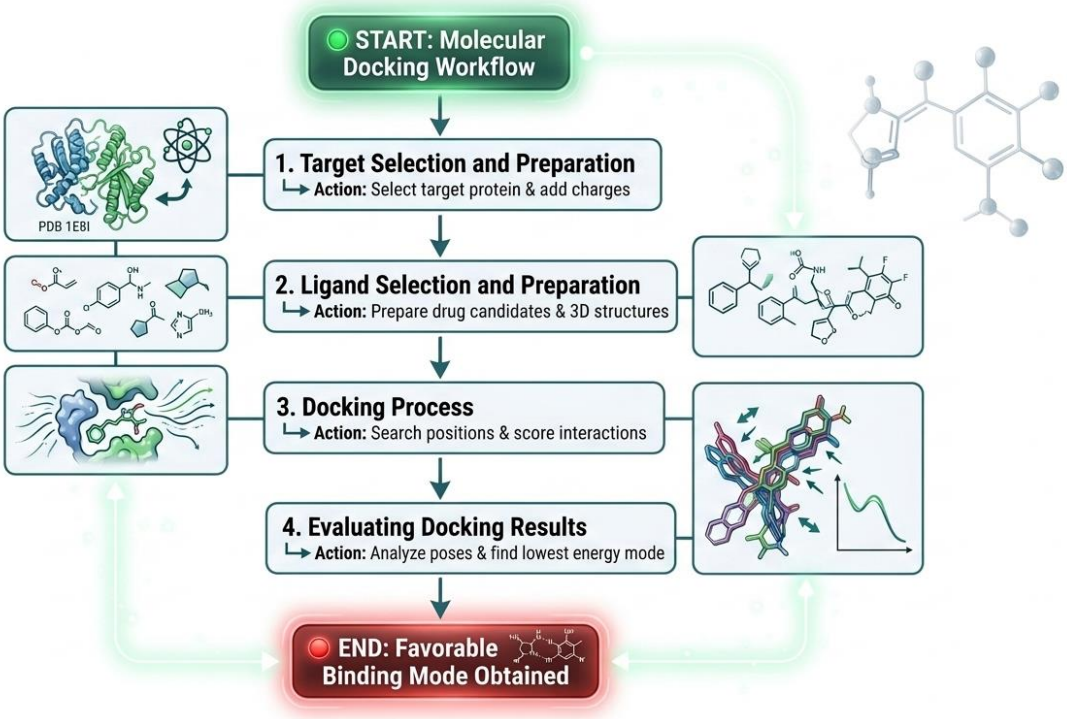


Figure 6. Steps of Molecular Docking

10. File Formats in Molecular Docking

Molecular docking applications require specific file formats for both receptors and ligands to ensure software compatibility. These identical formats provide a consistent method for representing molecular structures. Consequently, they enable unified integration and harmonization across various docking platforms. Common file formats include:

-MOL2 (Tripos Mol2)

Tripos Mol2 is an ASCII-based file that encompasses all necessary data to reconstruct a SYBYL molecule. Unlike rigid, fixed-column formats, it uses a free-format structure, which greatly facilitates its conversion into other file types. A MOL2 file comprehensively describes the molecule's structural details, including three-dimensional atomic coordinates, atom and bond types, and partial charges.

-SDF (structured data file)

Developed by Biovia (formerly Molecular Design Limited, or MDL), the Structure Data File (SDF) is a standard chemical text format that stores 2D and 3D structural data for ligands. A key distinction from the MOL2 format is the ability of SDF files to hold multiple chemical structures within a single document, using a four-dollar-sign delimiter ($$$$) as a separator. Furthermore, it details atomic connectivity and hybridization states, making it a staple format for managing ligand libraries in virtual screening workflows.

-PDB (Protein Data Bank) file format

The Protein Data Bank (PDB) format is the standard file type for storing three-dimensional atomic coordinates. Widely supported by molecular modeling software, these files contain extensive metadata alongside spatial data, including author credits, publication references, and experimental methodologies.

In terms of structure, a PDB file is a plain text document organized into individual lines called "records." A single file typically incorporates multiple record types to categorize different forms of structural information.

-PDBQT (Protein Data Bank, Partial Charge, and Atom Type)

This is an extension of the PDB file format. However, it’s a bit more descriptive, containing information about the ligands’ partial charges (q) and atom types (t).

These partial charges are essential for AutoDock to accurately evaluate electrostatic interactions throughout the docking simulation. Additionally, the PDBQT structure explicitly defines flexible, rotatable bonds, which allows the software to efficiently navigate and sample the conformational space of the ligand.

-XYZ (Cartesian coordinates)

Compared to other file formats, the XYZ file format is simple as it represents the total ligand atoms in the first row and a commentary in the second row.

The three-dimensional coordinates of the ligands are then listed from the third row. These ligands are represented as Cartesian (X, Y, and Z) coordinates. These files can be generated in some docking software quite easily.

11.Applications of Molecular Docking

Molecular docking has found its use in various sectors. Some of them are mentioned below

-Lead optimization

Due to its ability to predict an optimized orientation of biomolecules, it can be used to predict different binding modes of the ligand in the binding site of the target molecule. Such information can be further used to develop potent, selective, and efficient analogs.

-Hit Identifications

Since molecular makes use of the scoring function and search methods, this can be used to screen huge online databases to retrieve potent biomolecules.

-Drug-DNA Interactions Studies

A lot of the drugs available in the market utilize nucleic acids and auxiliary processes as their main cellular target. This is seen in the cases of anticancer therapeutic agents. By being able to study the correlation between a drug's molecular structure and its cytotoxicity, rational design and synthesis of new drugs can be done.

-ADMET prediction

Molecular docking studies can be used to predict the Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) properties of small molecules. These results can then be used to weed out compounds with unfavorable properties early in drug discovery and development.

Practical work

The Simplified Hückel Method: Practical work 1

Using the program "Hulis"

1. Purpose:

-Study of some conjugate systems

-Using the program "Hulis"

2. Procedures

HuLiS program

You can use the hulis. application online: http://www.hulis.free.fr

HuLiS is implemented both as a java applet and a javascript.

We will take the example of the javascript GUI, as shown in Figure 1.

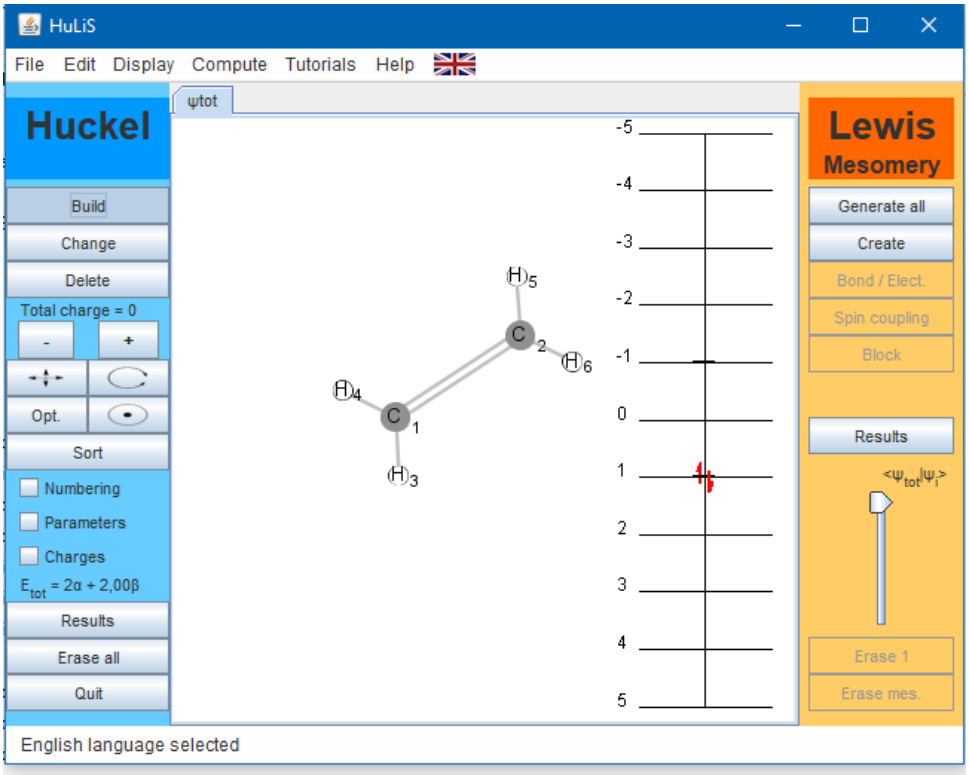


Figure 1. HuLiS GUI

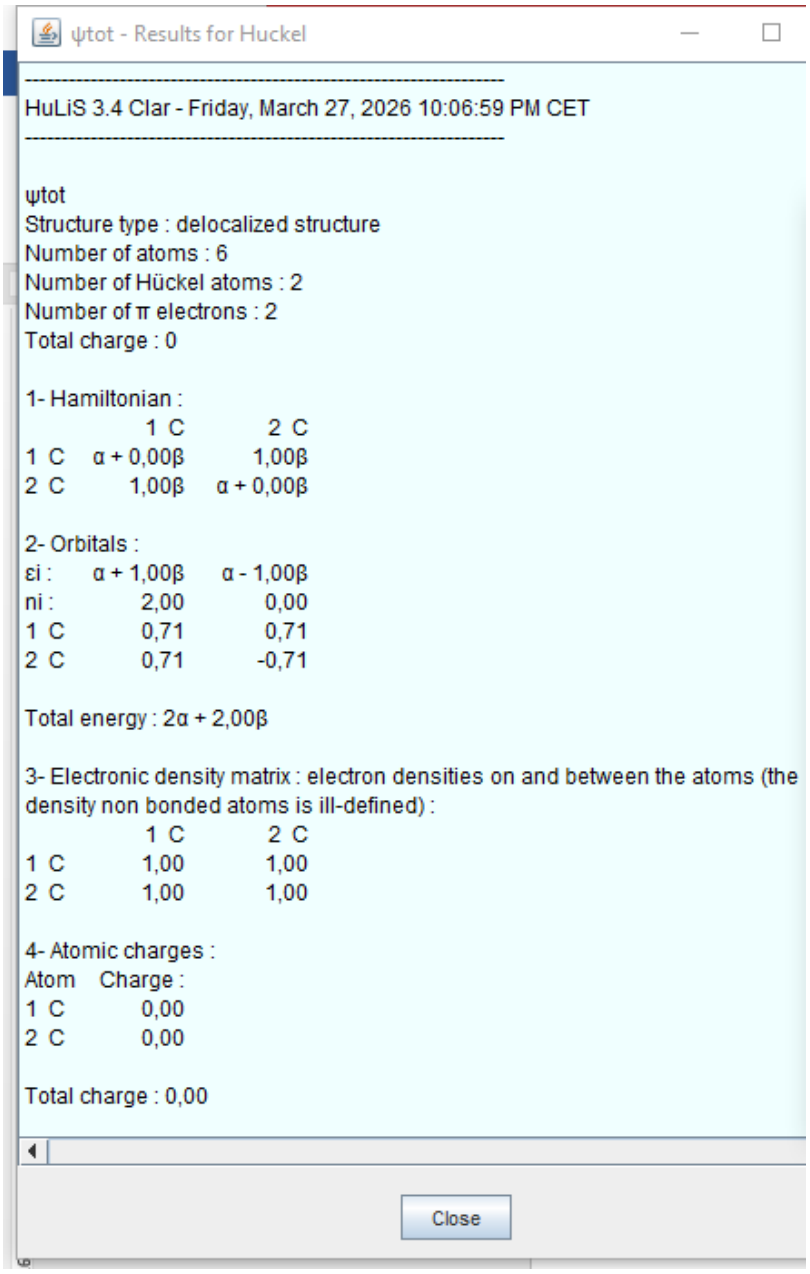
-Double-click on the Hulis icon to start the Program

-Running the Program:

-The main screen shows a molecule builder. Highlight Build on the left side of the screen if it is not already highlighted when the program starts.

3. Results

Click on the Results button, and a Results window appears with significant information about the Hückel calculation.



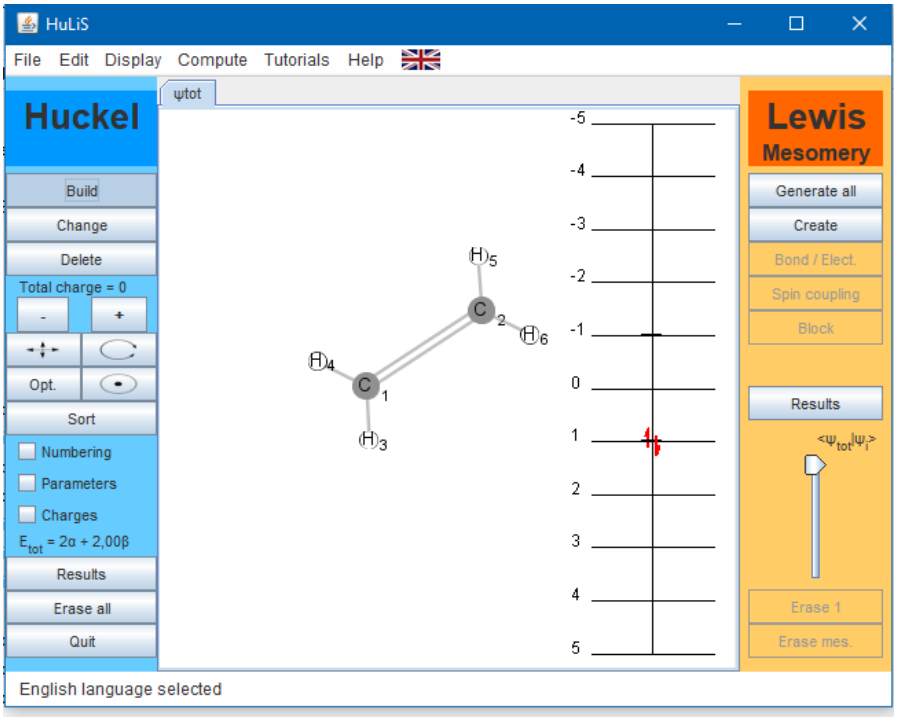


Figure 2. The results windows

To get a visual representation of the π-molecular orbitals generated in the Hückel program, simply click on one of the orbitals in the right-hand pane of the program window. The image below is of the HOMO.

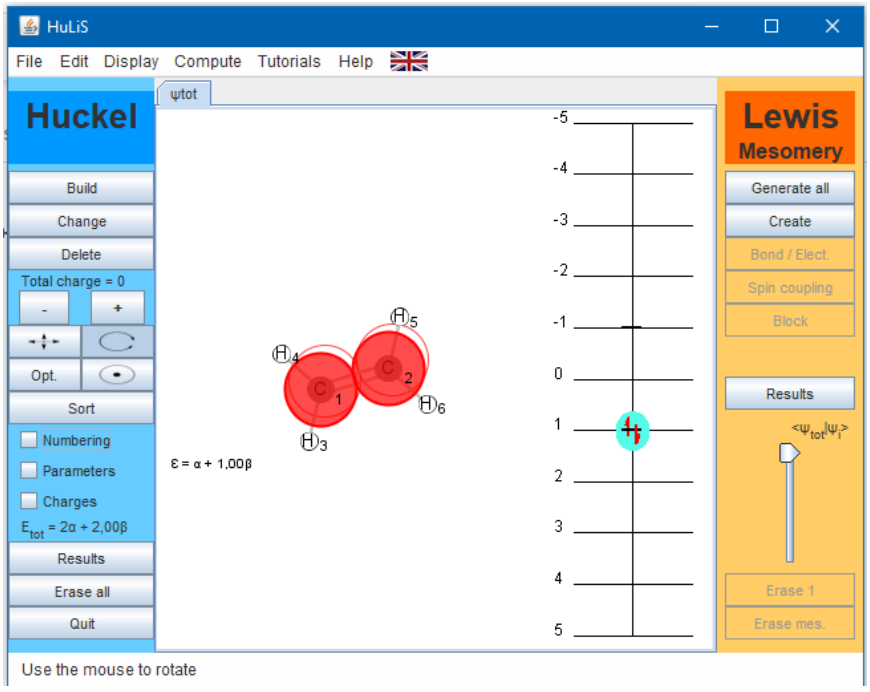
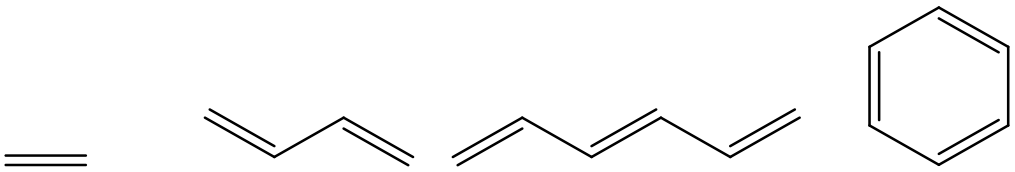


Figure 3. The Highest Occupied Molecular Orbital (HOMO)

Specific Objectives and application

-You are using the **Hulis application**. It is installed on the computers in the classroom.

-Enter the following molecules:



Ethene Buta-1,3-diene Hexa-1,3,5-triene benzene

4.Report

1-Construct the molecular orbital diagram.

2- Construct the diagram of the HOMO and LUMO molecular orbitals, and determine the analytical form of these OMs.

3- Calculate the electron energy.

4- Calculate the energy gap using the Hückel method. Conclude

Conformational Analysis and Exploration of the Potential Energy Surface: Practical work 2

Using the program "Gaussian and Gaussview"

1. Purpose:

- To become acquainted with basic concepts of Molecular Mechanics computations.

- To learn how to use the program Gaussian09W a program for building molecules and computing properties.

We will use the graphical interface program, GaussView, for conveniently preparing input to G09W and for analyzing results.

http://www.gaussian.com/

GaussView5 links:

https://www.d.umn.edu/~psiders/courses/chem5650/gaussviewtutorial/tutorial.html

https://www.mtholyoke.edu/courses/jwijngaa/PS/Gaussview2.pdf

Gaussian09 link: www.molcalx.com.cn/wp-content/uploads/2015/01/Gaussian09W_tutorial.pdf

2. Brief Description of Principles

-Conformational Analysis

Conformational analysis is the study of the different energy levels associated with the different conformations of a molecule. Conformations are the different 3-dimensional arrangements that the molecule can acquire by freely rotating around σ-bonds. Conformational analysis aims

to identify minimum energy structures, while simulation operations give an assembly of states that includes structures not at energy minima.

-Conformer

Conformers refer to the different spatial arrangements of atoms in a molecule that can interconvert through rotation about single bonds, with each conformer potentially exhibiting distinct energy levels and stability.

Conformational analysis requires a large number of energy evaluations and is practical only using molecular mechanics force fields.

-Conformational analysis in computational chemistry

Conformational analysis in computational chemistry uses molecular mechanics (MM) or quantum mechanics (QM) to identify and evaluate stable 3D arrangements of atoms (conformers) generated by single-bond rotations.

-Force fields

Force fields are mathematical models used to describe the potential energy of a system of atoms or molecules.

A force field typically includes terms for bond stretching, angle bending, dihedral torsions, and non-bonded interactions (van der Waals and electrostatic interactions).

$$E_{total} = E_{bonded} + E_{non\text{-}bonded}$$

Where,

$$E_{bonded} = E_{bonds} + E_{angles} + E_{dihedrals} + E_{impropers} \text{ and}$$

$$E_{non\text{-}bonded} = E_{van\ der\ Waals} + E_{electrostatic}$$

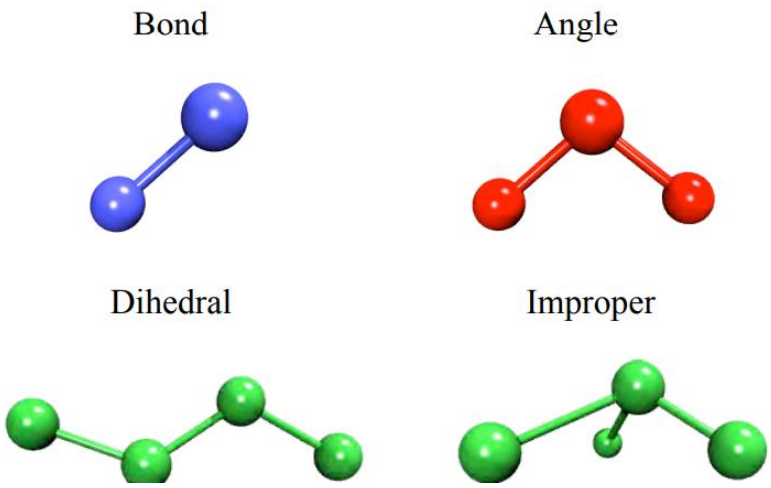


Figure 1. Energy Terms Described in Force Field

-Energy Minimization:

Using a force field to calculate the energy E of a structure is called "Molecular Mechanics" (MM). These calculations allow us to find the lowest energy state by minimizing E.

The total energy function has many minima and maxima. There is no general mathematical method to find the global minimum (i.e., the lowest) of this function. Therefore, numerical analysis methods are used to find local minima.

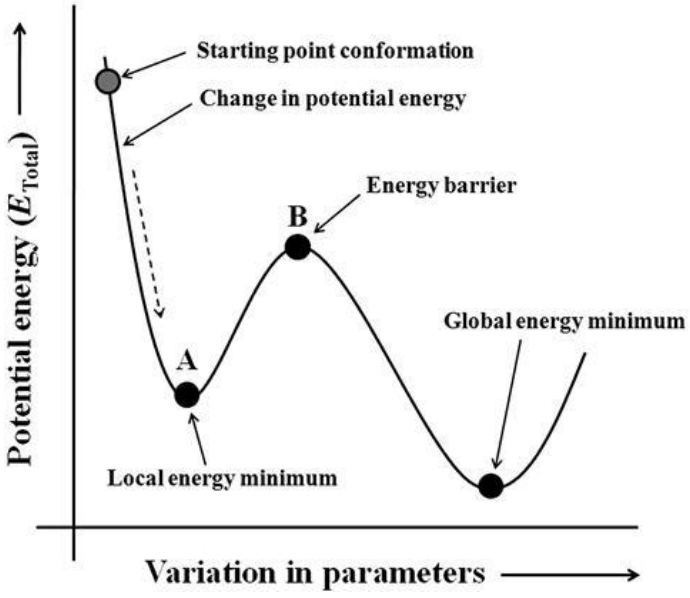


Figure 2. Different phases of a molecule during minimization of its energy.

The energy minimization methodology needs to involve identification of the point closest to the starting structure.

-Potential energy surfaces

The potential energy surface for a molecular describes the relationship between the molecular geometry, the relative positions of the component atoms, and the molecular energy.

Accurate PES are crucial to our understanding of molecular structure and reactivity.

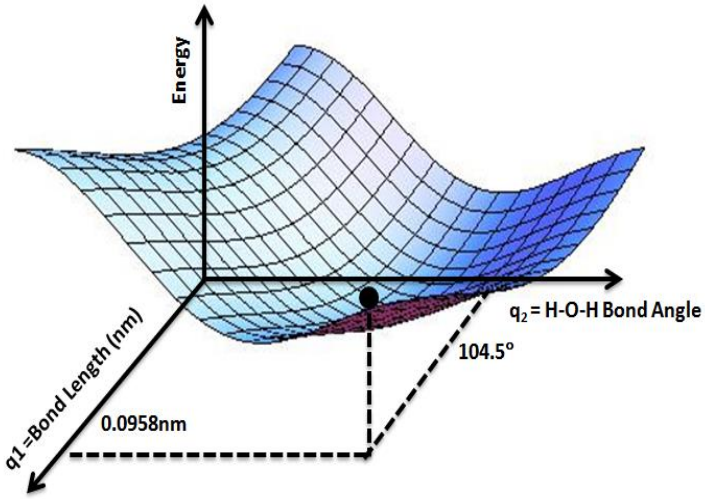


Figure 3. Potential energy surfaces

3. Procedures

1- Building with GaussView

-Click on the GaussView icon to run GaussView, the graphical interface.

The input files can be generated using any text editor

Output can be read in the editor.

NOTE: the .gjf or .com file

-Input File

Input File. (.gjf or .com extension) gjf means gaussian job file

Specify only the atom atomic numbers or symbols, their coordinates, the charge, and spin multiplicity.

The essential elements of this file come in the following order:

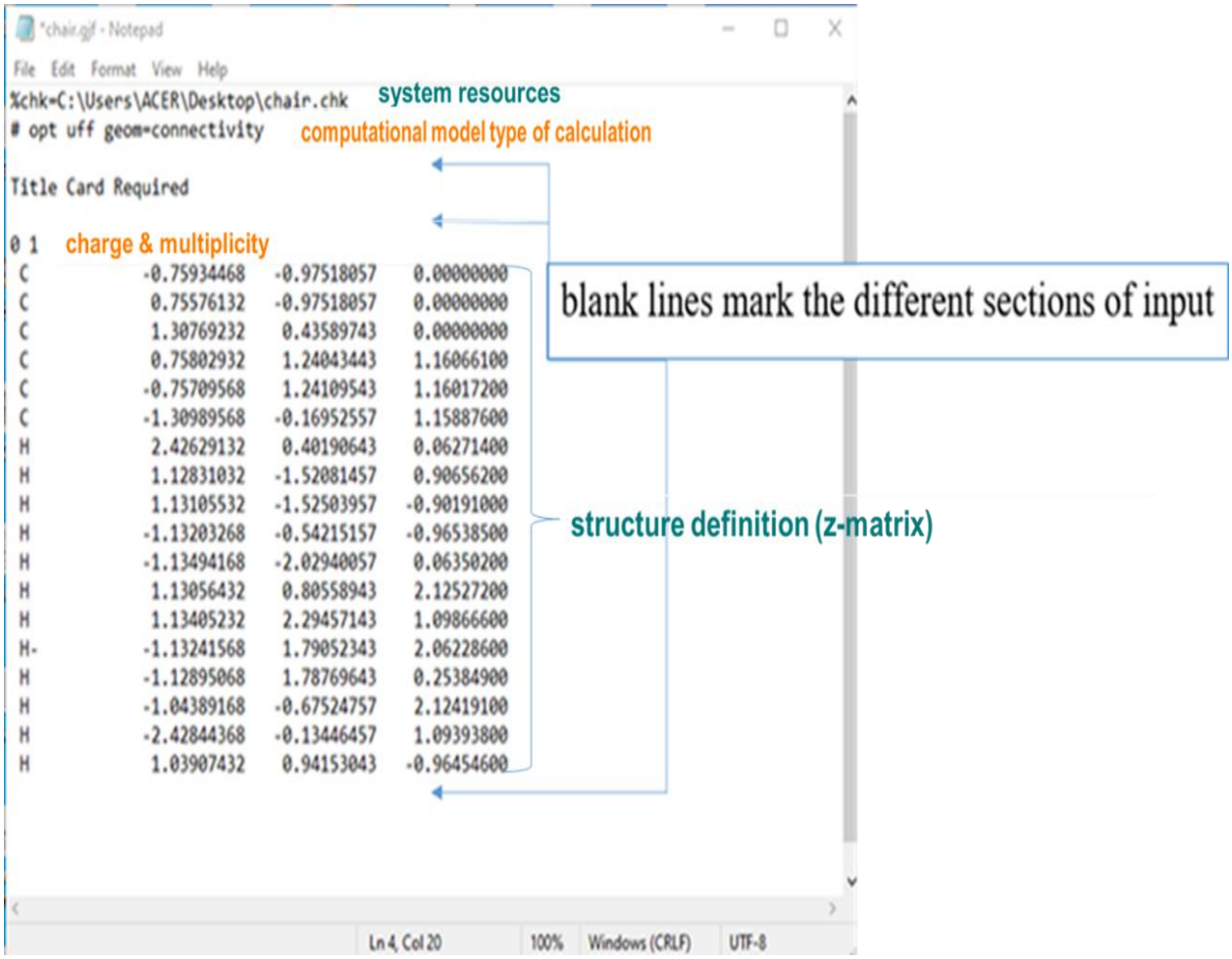


Figure 4. Example input file for cyclohexane-chair: chair.gjf

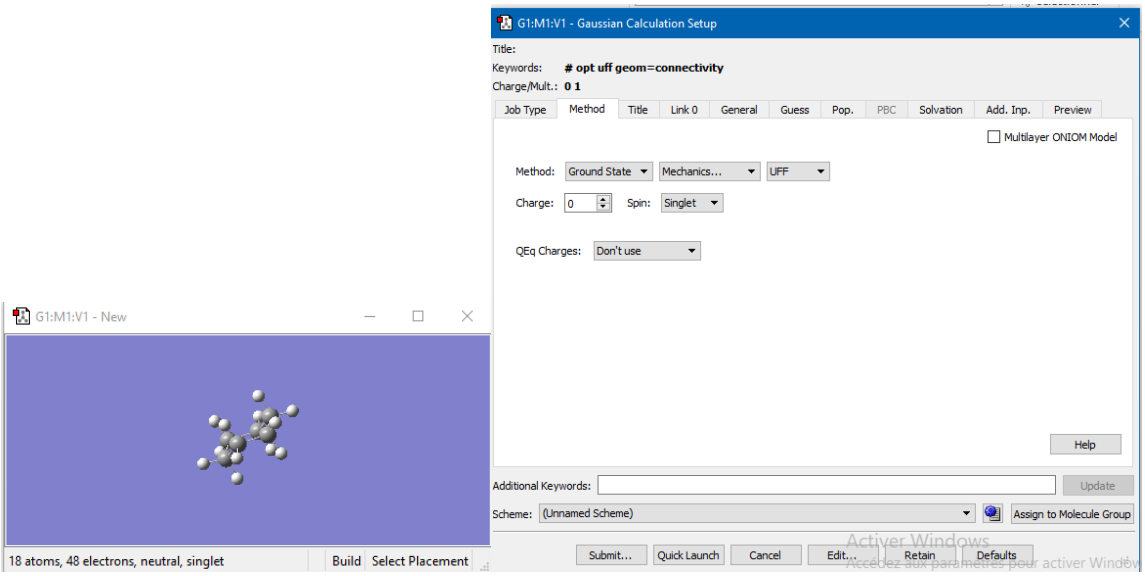


Figure 5. Main window of the Input file (GUI)

2. Submitting Calculations

-The submission dialogue box appear are running GaussView on.

You will name the input file, and then submit the calculation.

When Gaussian is finished running, you will receive a message in Gaussview

-Output file

Output file (.log) or (.out)

The output file (.out) contains a lot of information that was generated from the computation that was performed.

-Checkpoint files (.chk)

The checkpoint file contains the basis set and the SCF coefficients computed during the run.

3. Specific Objectives and application

3.1. Conformational analysis

Cyclohexane, a six-membered ring structure, can adopt several conformations to minimize strain and achieve stability. The most important conformations of cyclohexane are:

1. Chair Conformation:

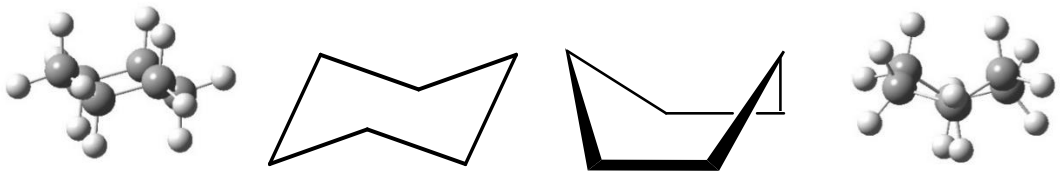
The most stable conformation of cyclohexane, where the carbon atoms are staggered, minimizing torsional strain. The C-C-C bonds are very close to 109.5o, so it is almost free of angle strain. It is also a fully staggered conformation and so is free of torsional strain.

2. Boat Conformation:

The boat conformation is less stable than the chair due. This too is almost free of angle strain, but in contrast has torsional strain associated with eclipsed bonds at the four of the C atoms that form the side of the boat.

1. Construct the cyclohexane molecule (1-chair and 2-boat).

2. Use molecular mechanics methods (energy minimization using the force field) to perform the conformational analysis of the cyclohexane molecule.



1 2

-Report

Examine the structure using GaussView and list all the energy terms describing these molecules.

-Compare relative energies with experiment:

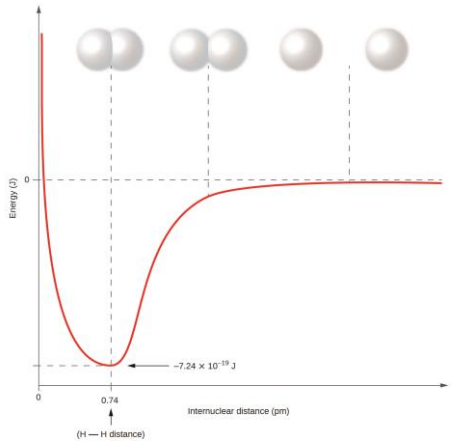
	Calculated	**expt. Estimate (Kcal/mol)**	**θ Angle (°)**	**ϕ Torsion (°)**
E(chair)				
E(boat)				

-Which Conformer is more stable? How much more stable

3.2. Exploration of the Potential Energy Surface

-Variation of Potential Energy as a Function of Internuclear Distance in Diatomic Molecules

The study of the system's energy variation as a function of geometric parameters leads to what is commonly referred to as the Potential Energy Surface (PES). In the case of diatomic molecules, the only geometric parameter is the internuclear distance, meaning the PES is a two-dimensional curve.



- Perform the following single-point calculations with different values of R (distance between the two hydrogen atoms). Use R = 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, and 1.1. - Record the energy value obtained at the end of each calculation.

-Record the resulting energy value at the end of each calculation.

-Report

1. Complete**:**

R	**Energie**	R	**Energie**
0.5		0.8	
0.6		0.9	
0.7		1.0	

1. Present and discuss your results?
2. Draw a graph showing the variation of energy as a function of distance: E(R)

Structure Activity Relationship: Practical work 3 HyperChem Lab and SwissADME, web

Procedure

1. Purpose:

- To become acquainted with basic concepts of (SAR) computations.

- To learn how to use the program HyperChem Lab a program for building molecules and computing properties.

-We will use the SwissADME, web tool in the design for compute physicochemical descriptors

.https://www.swissadme.ch/index.php

2. Brief Description of Principles

-Molecular Descriptors Used in QSAR

Physicochemical descriptors help identify optimal ranges for drug-like properties.

LogP (Partition Coefficient)	Indicates lipophilicity, which affects membrane permeability and drug absorption.
Molecular Weight (MW)	Higher MW compounds often have reduced absorptivity and may violate drug-likeness rules.
Polar Surface Area (PSA)	Measures the surface area formed by polar atoms; strongly associated with oral bioavailability and blood–brain barrier penetration.

Structural Descriptors

These descriptors are essential for modeling activity dependent on **2D structure**

Topological Indices	Include Wiener index, Zagreb index, and Kier–Hall indices; these reflect branching, molecular shape, and connectivity patterns.
Connectivity Indices	Describe how atoms are linked within the structure and help differentiate isomers with similar formulas but different shapes.

Quantum Chemical Descriptors

Quantum descriptors originate from quantum mechanical calculations and describe electronic distribution in molecules.

Quantum descriptors are particularly useful for understanding drug–receptor interactions.

HOMO (Highest Occupied Molecular Orbital)	Indicates electron-donating ability.
LUMO (Lowest Unoccupied Molecular Orbital)	Indicates electron-accepting ability.

-Lipinski's rule of five

Rule of five is a rule of thumb to evaluate drug likeness or determine if a chemical compound with a certain pharmacological or biological activity has properties that would make it a likely orally active drug in humans.

The biologically active molecule must implement five conditions to be potentially used as a drug

for oral administration. Poor absorption or permeation are most likely if:

- Molar mass >500,

-Number of H-bond acceptors >10,

-Number of H-bond donors > 5,

- LogP> 5 (or MlogP>4.15)

3. Procedures

SwissADME web tool

SwissADME was used to estimate the drug similarity parameters and based on their pharmacokinetic parameters, the derivative was proposed to be safe and not against Lipinski's rule of five.

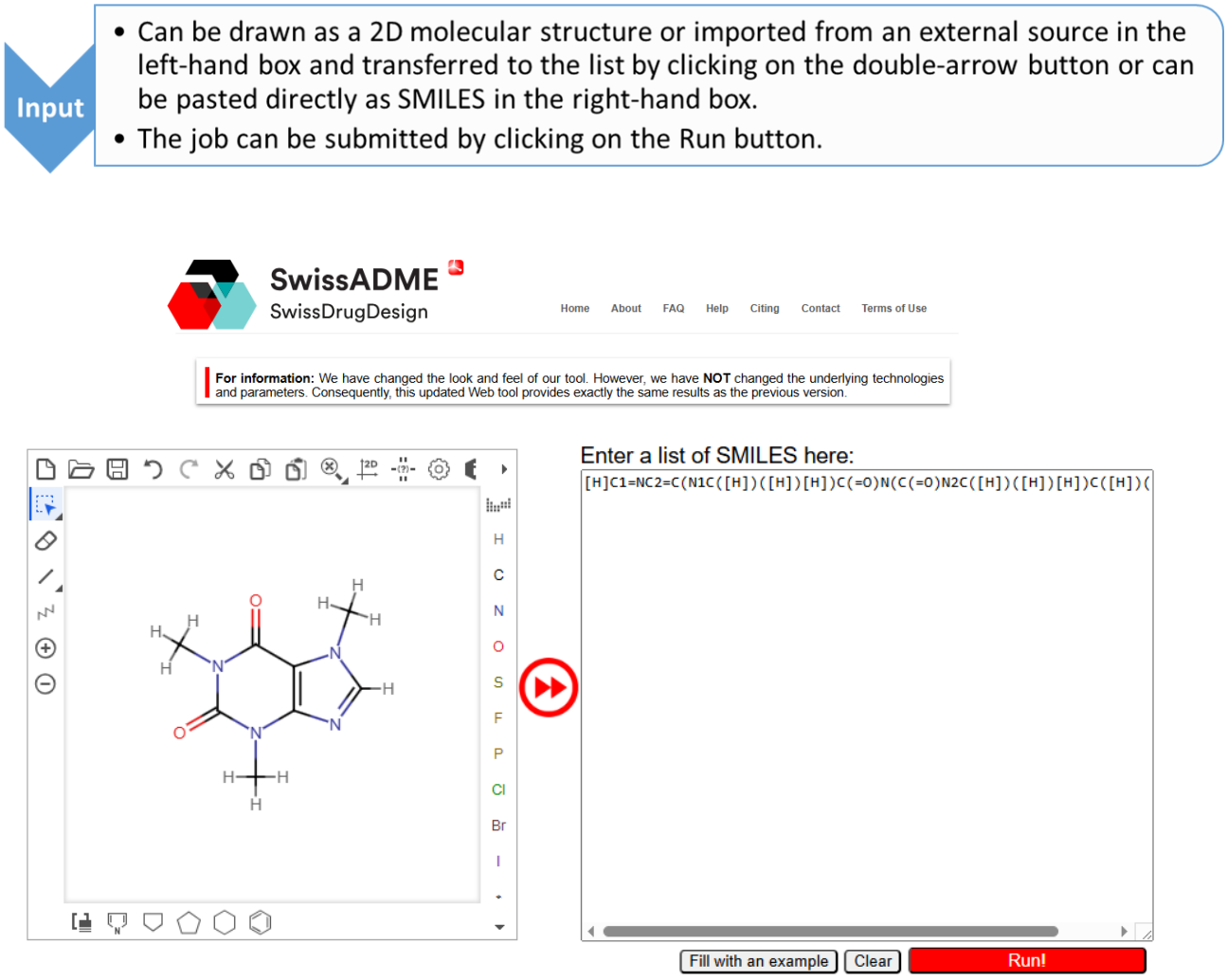
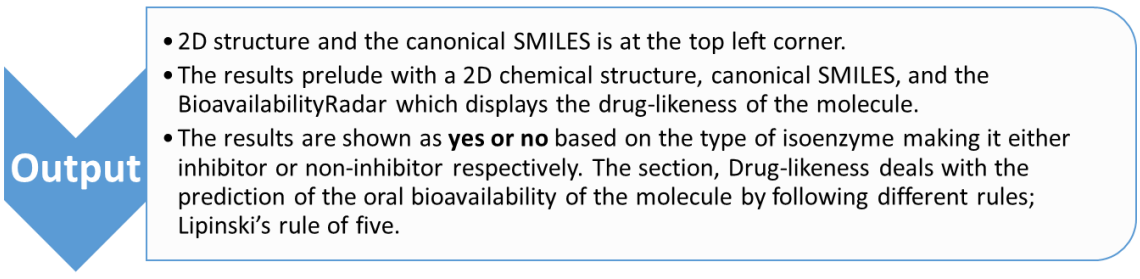


Figure 1. Main input panel of SwissADME.



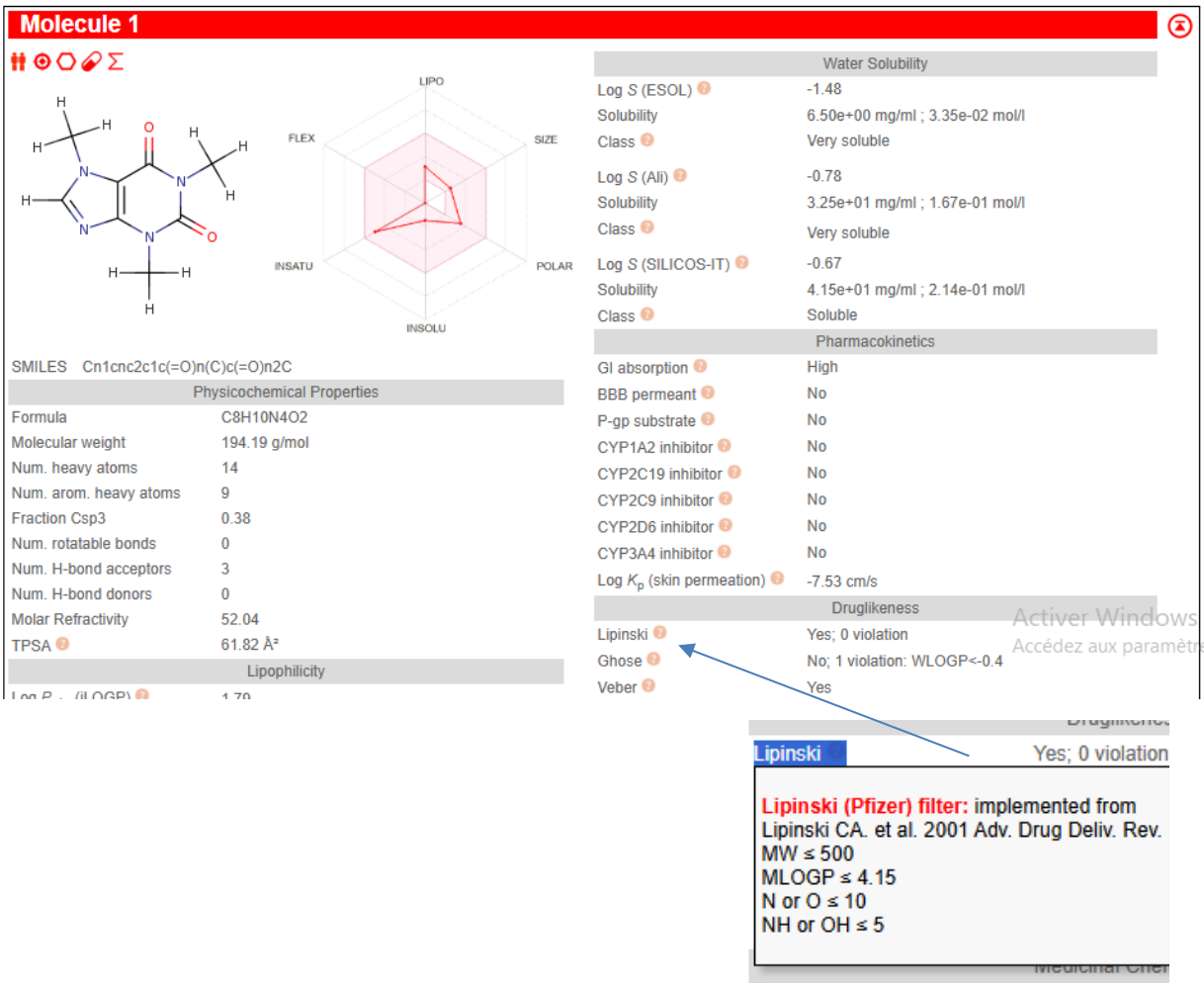
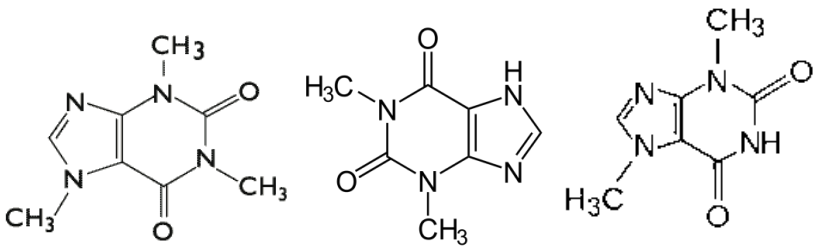


Figure 2. Graphical output

Step 1: Drawing the Molecules

-Use HyperChem to build the following molecules:



Caffeine **Theophylline** **Theobromine**

-Go to the Build menu and select Add H & Model Build to automatically add hydrogens and adjust the geometry.

-Go to the File menu and select Start Log …. Give the log file a significant nameand note its directory location

Step 2: Geometry Optimization

Go to the Setup menu and choose:

Molecular Mechanics (e.g., MM+)

Go to the Compute menu and select Geometry Optimization.

Choose an algorithm (like Polak-Ribiere) and click OK. Wait for the calculation to finish (it will say "Converged" at the bottom).

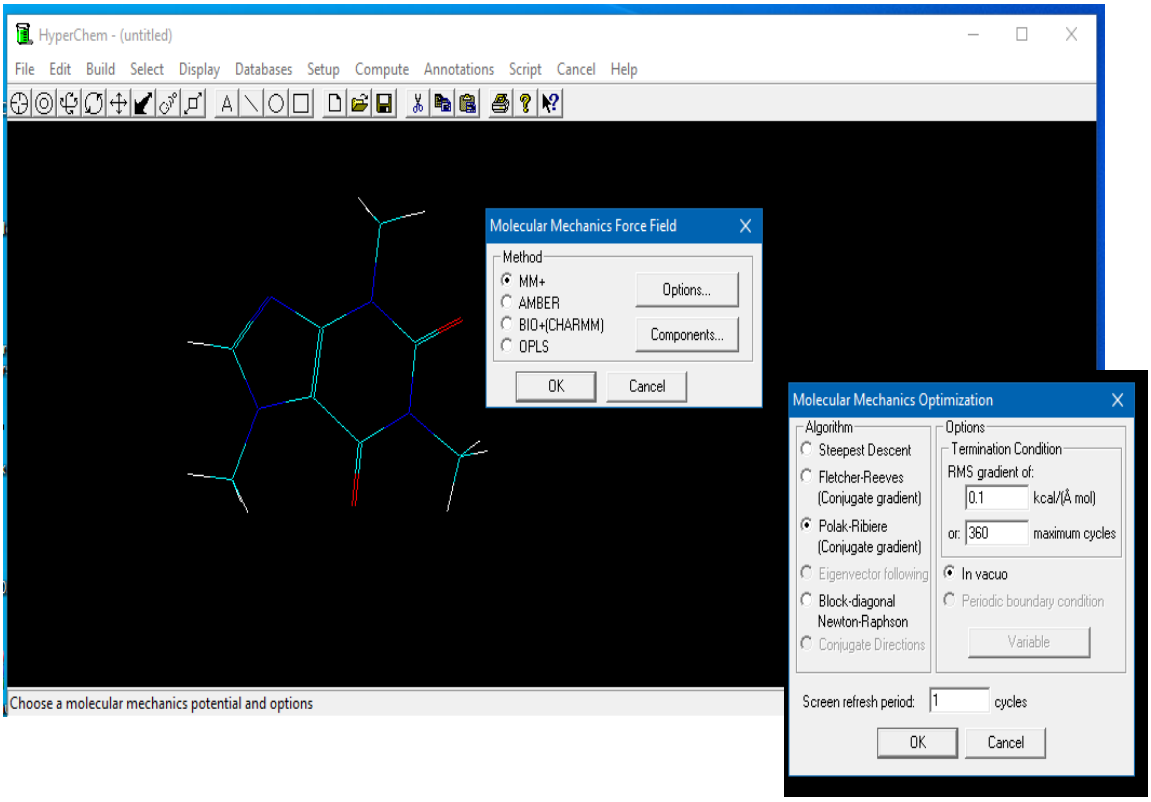


Figure 3. Geometry Optimization

Step 3: Calculating Properties for the Table

Once the geometry is optimized, you can extract the data:

Energy (E): Look at the Status Bar at the bottom of the screen; the "Total Energy" will be displayed there.

Molecular Weight: Go to Edit -> Chemistry -> Molecular Properties.

-QSAR Properties:

Go to the Compute menu.

Select QSAR Properties.

In the dialog box, click on the buttons for Polarizability, Molar Refractivity, Molar Volume, and Hydration Energy.

Click Compute to get the values for each.

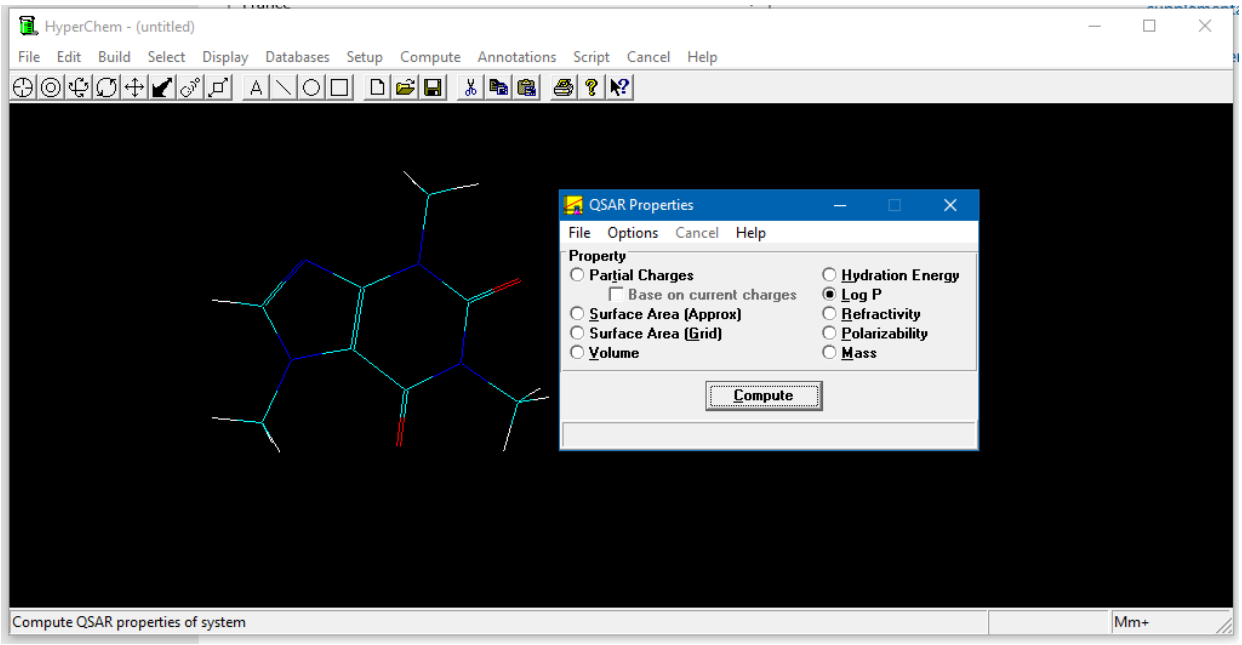


Figure 4. QSAR Properties

-Report

-Complete

	E_{tot}	polarizability	molecular weight (amu)	molecular Volume	Hydration energy
Caffeine					
Theophylline					
Theobromine					

-check Lipinski's rule of five.

Development of a QSAR model: Practical work 4

Using Python tools

1. Purpose:

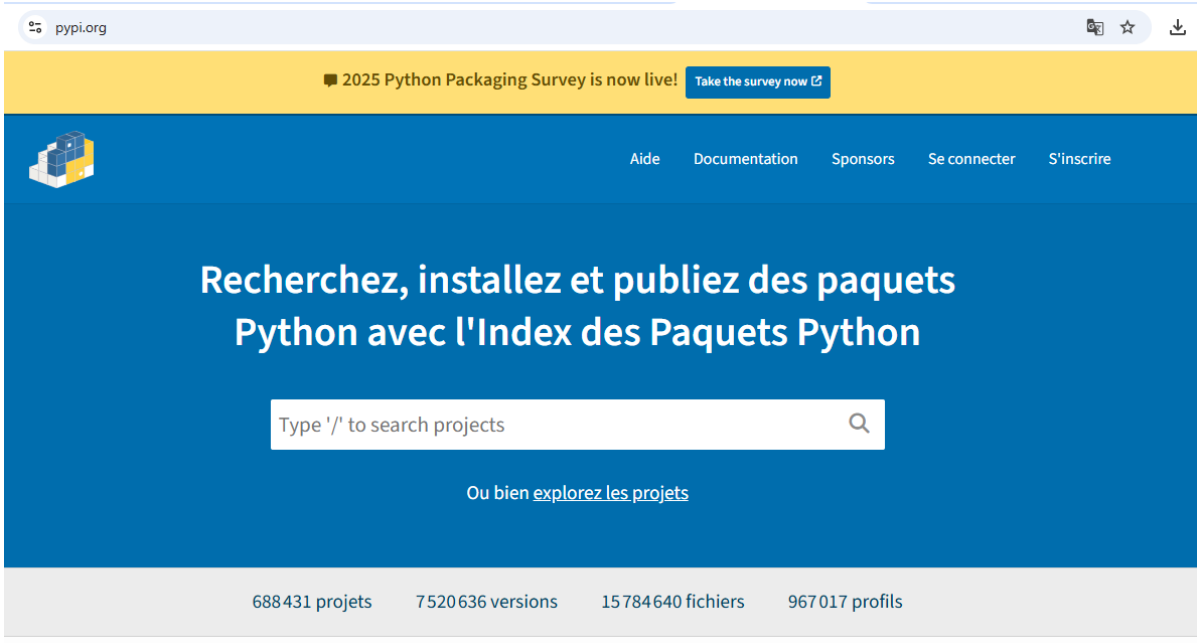
Building a QSAR Model

To understand how Python tools (Pandas, Scikit-learn, Matplotlib) work together to predict chemical properties

2. Brief Description

Python is free and open-source, many tools are also free and open-source.

- **site PyPi (https://pypi.org/)**



The PyPI website hosts thousands upon thousands of these tools, which are called packages.

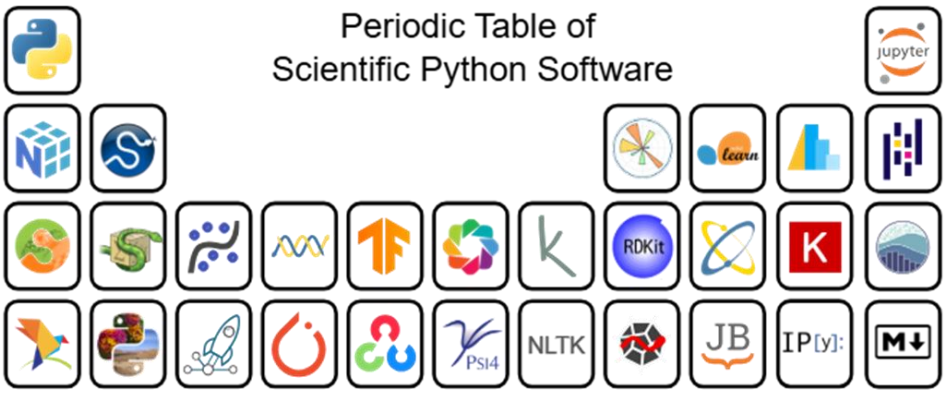
Because they are so easy to install, we have a huge toolbox at our disposal with which we can build just about anything we want.

The Python programming language is most often stored in simple text files called scripts or source code, identifiable by their **.py** extension.

Learning a programming language like Python will require writing lines of code using a text editor.This code can be written on many platforms, and text editors like Notepad or Notepad++ can be used, as well as specialized integrated development environments (IDEs).

-Python Tools

-install the chemical informatics library and import necessary modules



Python Tool	
Pandas (pd)	Organizes the chemical data into a table (DataFrame).
Scikit-learn (sklearn)	Provides the algorithm (Linear Regression) to find the pattern.
	Maps the chemical structure to the activity.
.predict()	Calculates the property of a molecule it has never seen before.
Matplotlib (plt)	Creates the graph to show the correlation.

3. Procedure

-Simple linear regression

-We want to predict the **Toxicity** of a series of alkanes based on a simple structural descriptor: the **Number of Carbon Atoms**

- **Input (X):** Number of Carbons.
- **Output (y):** Measured Toxicity.

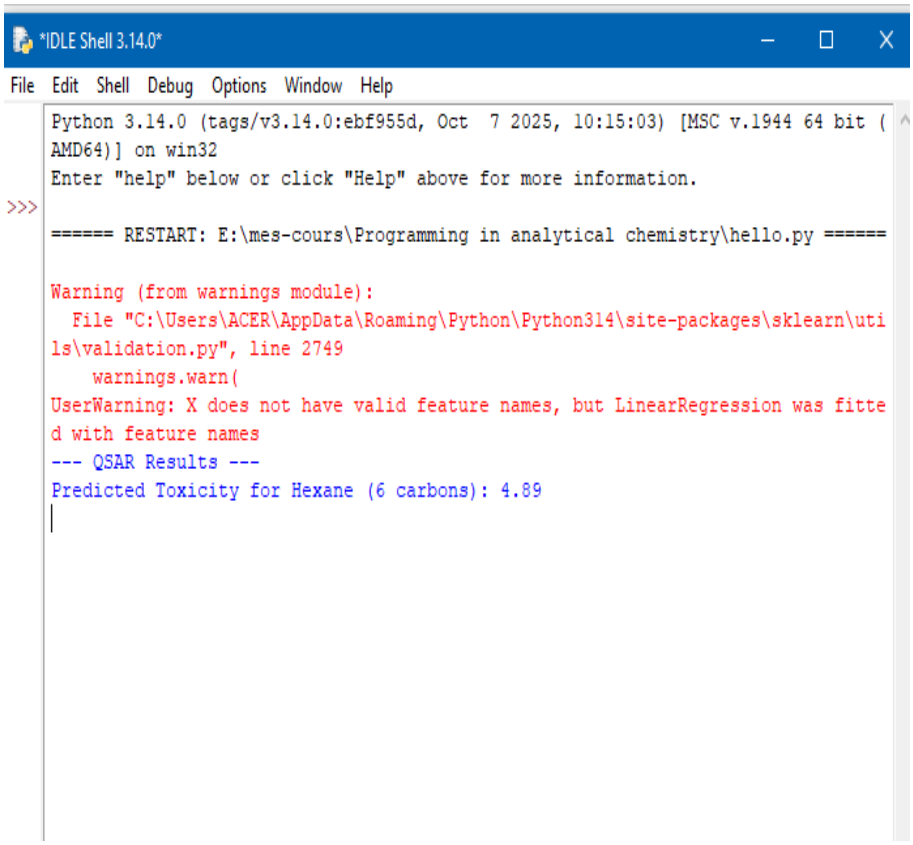
-The Python Code

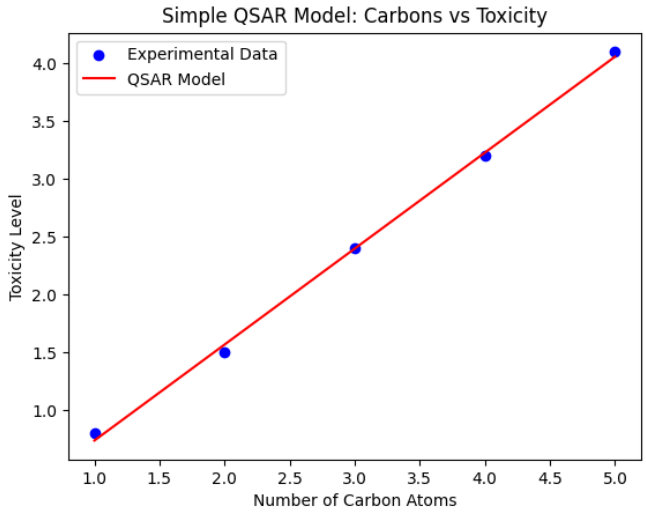
hello.py - E:\mes-cours\Programming in analytical chemistry\hello.py (3.14.0)

File Edit Format Run Options Window Help

```
import pandas as pd
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt

# TOOL 1: Pandas - Used for Data Organization
# We create a dictionary of molecules (Methane to Pentane)
data = {
    'molecule_name': ['Methane', 'Ethane', 'Propane', 'Butane', 'Pentane'],
    'n_carbons': [1, 2, 3, 4, 5],
    'toxicity': [0.8, 1.5, 2.4, 3.2, 4.1]
}
df = pd.DataFrame(data)

# TOOL 2: Scikit-learn - The Machine Learning "Brain"
# Initialize the Linear Regression model
model = LinearRegression()

# TOOL 3: Training (Fitting)
# The model "learns" the mathematical relationship between X and y
X = df[['n_carbons']]
y = df['toxicity']
model.fit(X, y)

# TOOL 4: Prediction
# We ask the model to predict toxicity for a Hexane (6 carbons)
new_molecule = [[6]]
prediction = model.predict(new_molecule)

print(f"--- QSAR Results ---")
print(f"Predicted Toxicity for Hexane (6 carbons): {prediction[0]:.2f}")

# TOOL 5: Matplotlib - Data Visualization
plt.scatter(X, y, color='blue', label='Experimental Data') # Points
plt.plot(X, model.predict(X), color='red', label='QSAR Model') # Line
plt.xlabel("Number of Carbon Atoms")
plt.ylabel("Toxicity Level")
plt.title("Simple QSAR Model: Carbons vs Toxicity")
plt.legend()
plt.show()
```

-QSAR Results





2. Multiple Linear Regression

By adding new chemical properties, we move from simple linear regression to multiple linear regression. This allows the model to learn from several factors simultaneously to make more accurate predictions.

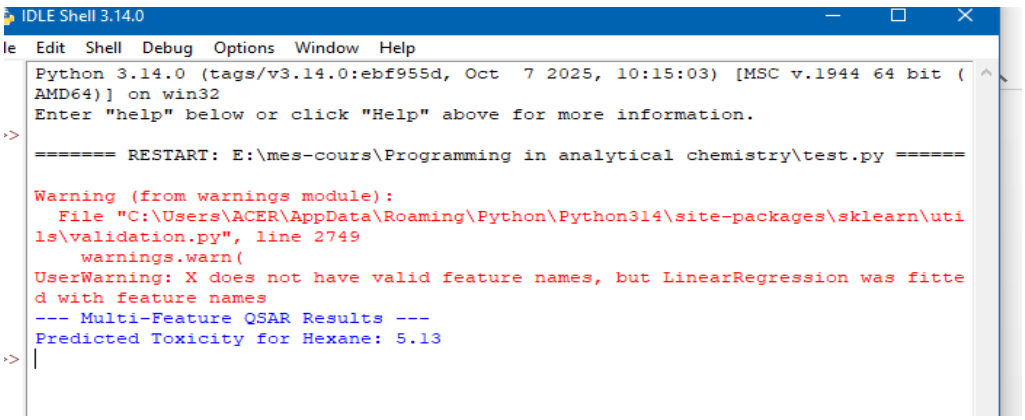
-The Python Code

test.py - E:\mes-cours\Programming in analytical chemistry\test.py (3.14.0)

File Edit Format Run Options Window Help

```
import pandas as pd
from sklearn.linear_model import LinearRegression

# TOOL 1: Pandas - Data Organization with Multiple Features
# We added Molecular Weight and LogP (lipophilicity) as new chemical properties
data = {
    'molecule_name': ['Methane', 'Ethane', 'Propane', 'Butane', 'Pentane'],
    'n_carbons': [1, 2, 3, 4, 5],
    'molecular_weight': [16.04, 30.07, 44.10, 58.12, 72.15],  # New Feature 1
    'logP': [1.09, 1.81, 2.36, 2.89, 3.45],                   # New Feature 2
    'toxicity': [0.8, 1.5, 2.4, 3.2, 4.1]
}

df = pd.DataFrame(data)

# TOOL 2 & 3: Model Initialization and Training (Fitting)
# X now contains a list of multiple input features
X = df[['n_carbons', 'molecular_weight', 'logP']]
y = df['toxicity']

model = LinearRegression()
model.fit(X, y)

# TOOL 4: Prediction for a New Molecule (Hexane)
# Provide values for all 3 features in the exact same order: [n_carbons, molecul
new_molecule = [[6, 86.18, 3.90]]
prediction = model.predict(new_molecule)

print("--- Multi-Feature QSAR Results ---")
print(f"Predicted Toxicity for Hexane: {prediction[0]:.2f}")
```

-QSAR Results



Interaction energy between an amino acid and a water molecule: Practical work 5

Using the program "Avogadro"

1. Purpose:

-Quantify a hydrogen bond type interaction.
https://ressources.unisciel.fr/_reste_a_valider/drug_design/co/TP1.html

We will be using Avogadro, an open-source software for visualizing and editing molecules. The main objective here is to quantify one of the most important molecular interactions: hydrogen bonding.

The publisher's website, in English, is very well illustrated for this purpose: http://avogadro.cc/.

-It provides an intuitive graphical user interface that allows students and researchers to easily construct complex three-dimensional chemical structures, from simple water molecules to intricate amino acids and proteins.

-Its powerful rendering engine enables precise measurement of atomic distances, bond angles, and hydrogen bonding interactions, making it an indispensable tool for quantifying interaction energies and studying molecular dynamics.

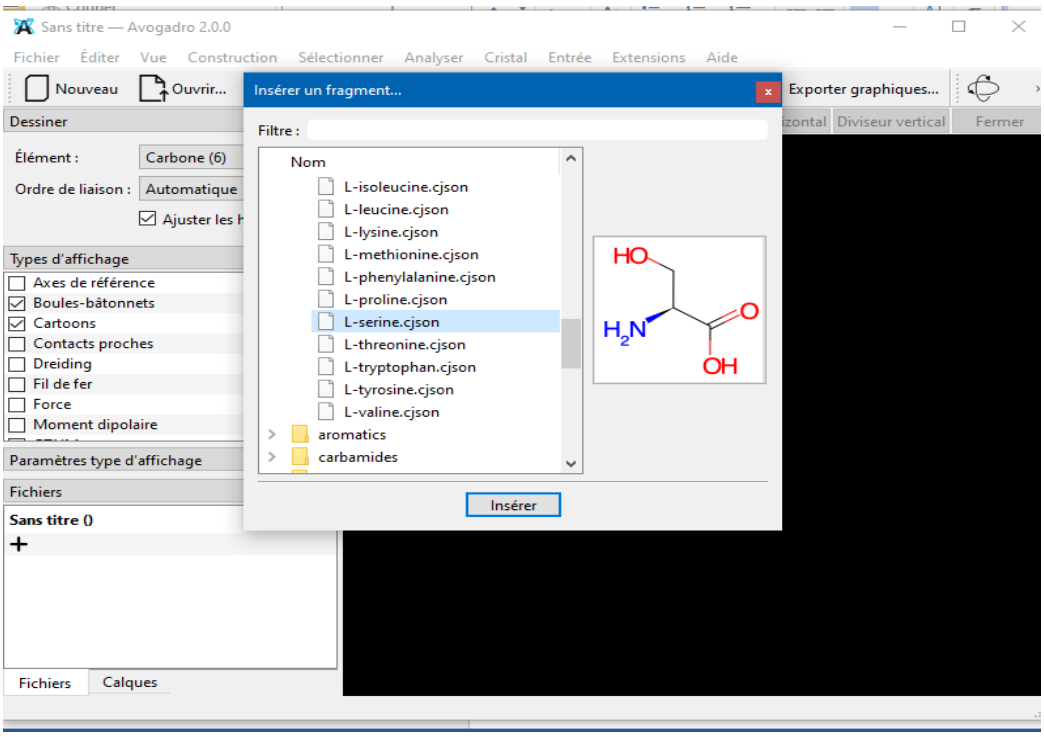


Figure 1. Avogadro 2's main screen.

2.Procedures

1/**Construction of an amino acid – first energy minimization calculation.**

1. Construct the amino acid L-threonine in its neutral form.

- A structure of L-threonine is provided using the Avogadro software. Open the Build menu > Insert > Fragment and then load the L-serine.cml file into the amino-acid subdirectory.

2. Select the MMFF94 force field. Calculate the energy of the unoptimized structure.

3. Optimize the structure of the molecule. To do this, you will first use 1000 steps of steepest descent followed by 5000 steps of conjugate gradient descent.

4. Calculate the energy of the optimized structure.

2/ Analysis of an amino acid-water molecule interaction and quantification of the interaction energy.

1. Construct a water molecule. Optimize its structure and determine the molecule's energy.
2. Construct a molecular system in which a water molecule interacts with the amino acid L-serine.cml.
3. Optimize the molecular system. Determine the energy after optimization.

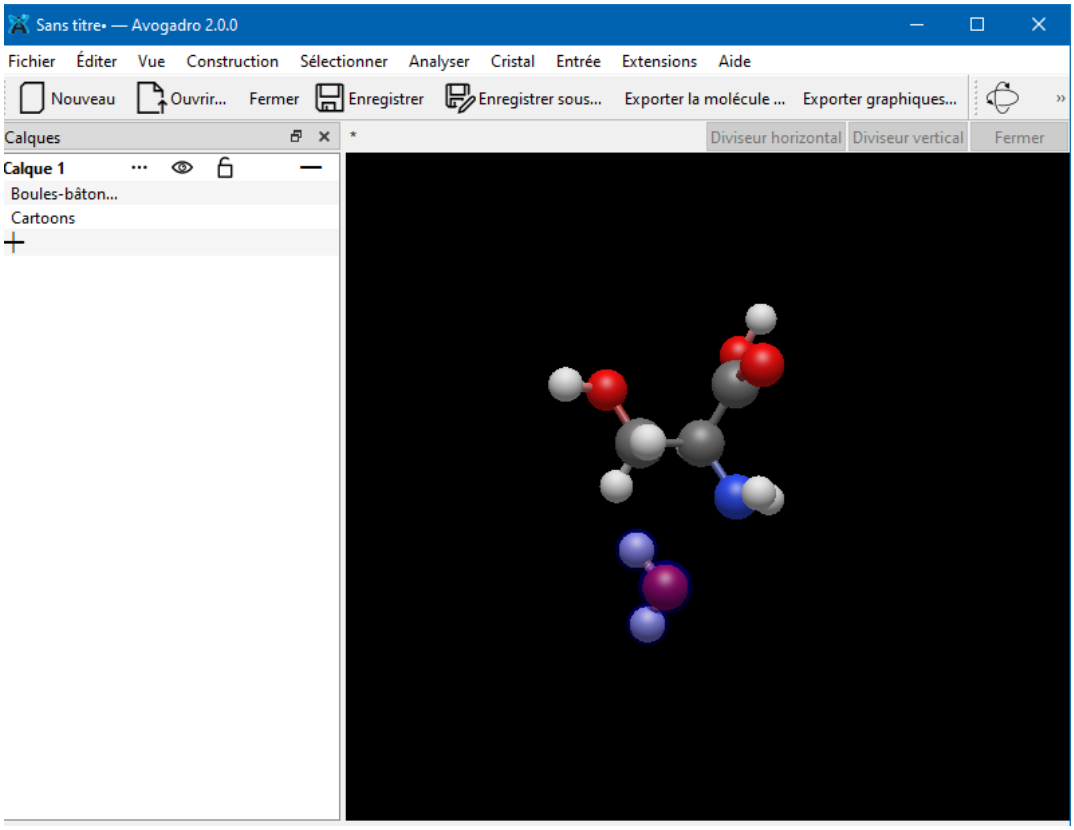


Figure 2. Structure of the L-serine – H_2O complex after geometry optimization.

-Report

-Quantify the interaction energy between the water molecule and the amino acid. Is it a hydrogen bonding type interaction?

To estimate the amino acid-water molecule interaction energy (which we will denote ΔE), we can consider as a first approximation that:

$$\mathbf{\Delta E = E_{cpx} - E_{ser} - E_{H2O}}$$

with:

$\mathbf{E_{cpx}}$ the energy of the optimized L- serine – H2O system

$\mathbf{E_{ser}}$ the energy of the amino acid optimized

$\mathbf{E_{H2O}}$ the energy of the water molecule.

-We are interested in an association reaction between a receptor (R) and a ligand (L) forming a ligand-receptor-ligand (RL) complex such that: R + L = RL

-Assuming that the association is solely due to a hydrogen bond interaction, what are the association and dissociation constants of this equilibrium at 25°C?

Preparing Molecules for DOCKing: Practical work 6

Using the program UCSF ChimeraX

1.Purpose:

Molecular docking programs predict how small molecules (ligands) bind to target macromolecules (receptors). They are widely used in drug discovery to identify lead compounds and optimize therapies. Beginners can explore these concepts using the RCSB PDB and review public domain literature via the National Institutes of Health.

-Preparing the protein and ligand for docking

-Evaluation of a docking protocol to locate the interaction site of carazolol in the β-2 adrenergic receptor.

2. Brief Description

-**Virtual Screening:** Rapidly scanning millions of candidate molecules in chemical databases (like ZINC or DrugBank) to find the best potential inhibitors for a specific disease target.

-**Drug Repurposing:** Testing already approved, existing medications against new diseases to save years in the development and clinical trial process.

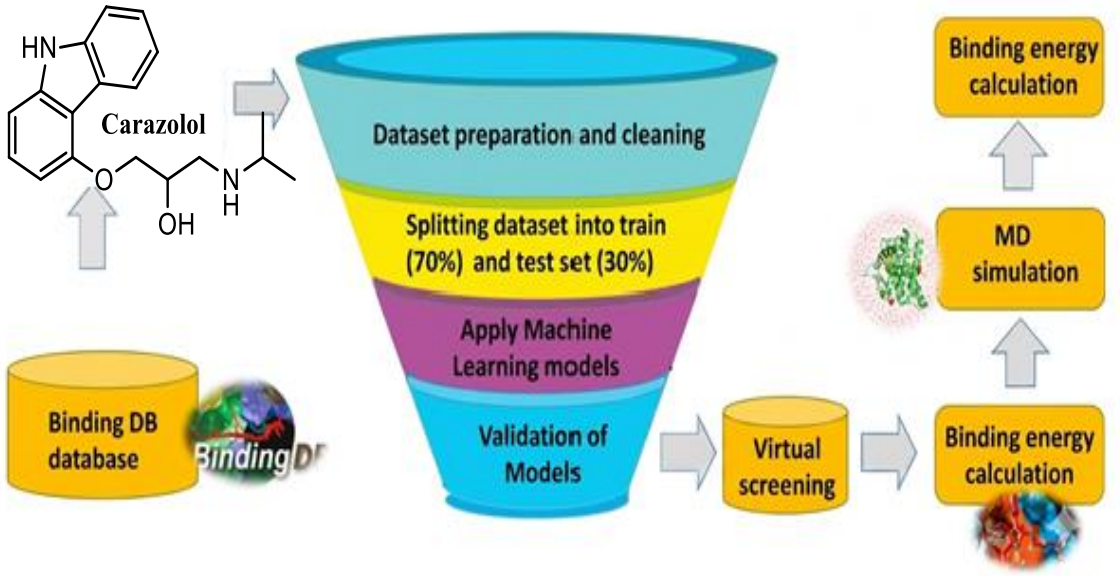
- **Lead Optimization:** Adjusting the chemical structure of a promising hit compound to improve its binding affinity and selectivity.

-**Enzyme and Protein Binding Predictions:** Understanding the exact 3D orientation (the "pose") and the thermodynamics of how a drug fits into the active site of a protein.

-**Prepare the Molecules:** Process receptor and ligand structures into compatible file formats (typically .pdbqt) using tools like AutoDockTools.

-Define the Search Space: Identify the active binding pocket on the receptor and set a 3D grid box where the simulation will take place.

1. **Run the Simulation:** Launch a docking program to test thousands of ligand orientations. The software scores each pose based on the lowest binding free energy.
2. **Analyze the Results:** Use visualization software (like PyMOL or ChimeraX) to review the binding interactions, hydrogen bonding, and hydrophobic contacts.



Assessing molecular docking tools: understanding drug discovery and design

3.Procedures

The software used in this lab, UCSF Chimera, is a visualization tool and will be used to analyze the results. UCSF Chimera is cross-platform and free for non-commercial use.

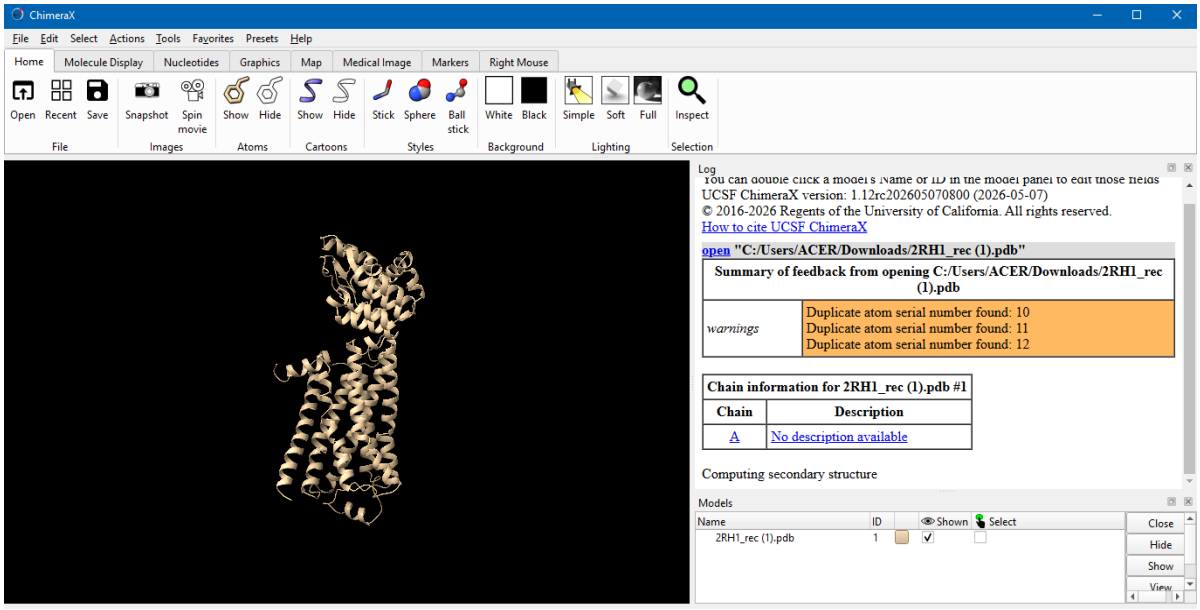
It is not possible to run docking simulations with this software; however, it will serve as an interface to launch Autodock Vina on the server provided by UCSF Chimera or directly on a computer.

A basic step in any docking project is selecting the file that will be used for the structure of the target. We shall use 2RH1_rec.pdb. A visualization of which can be seen below:

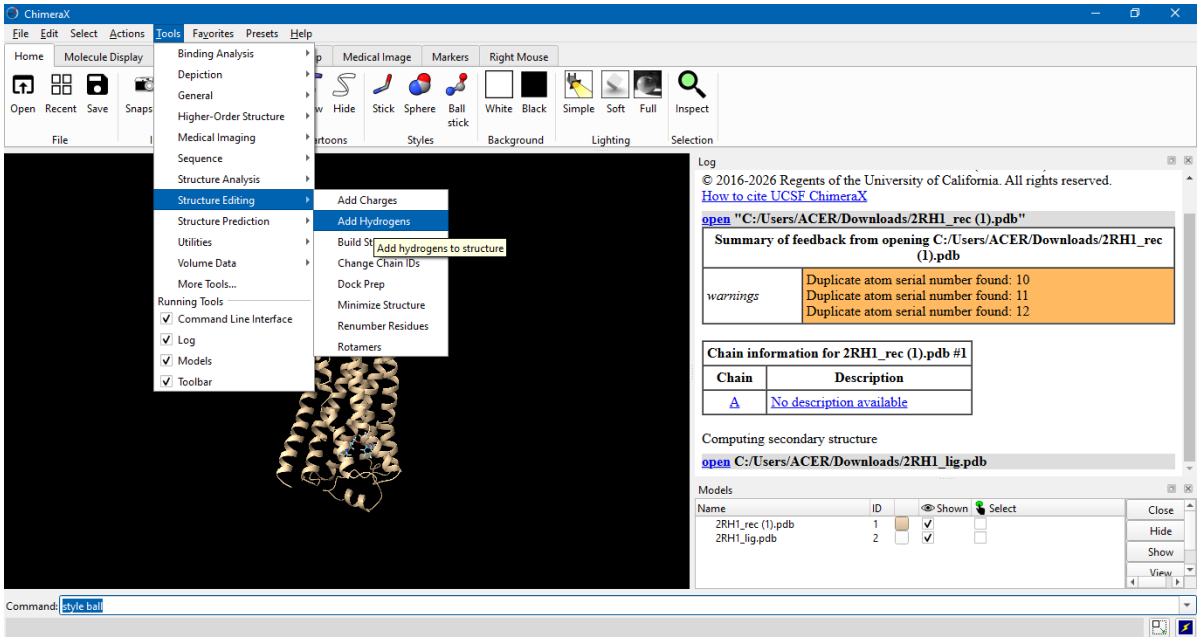
Figure 1. Image generated using Chimera of *β-2 adrenergic receptor*.

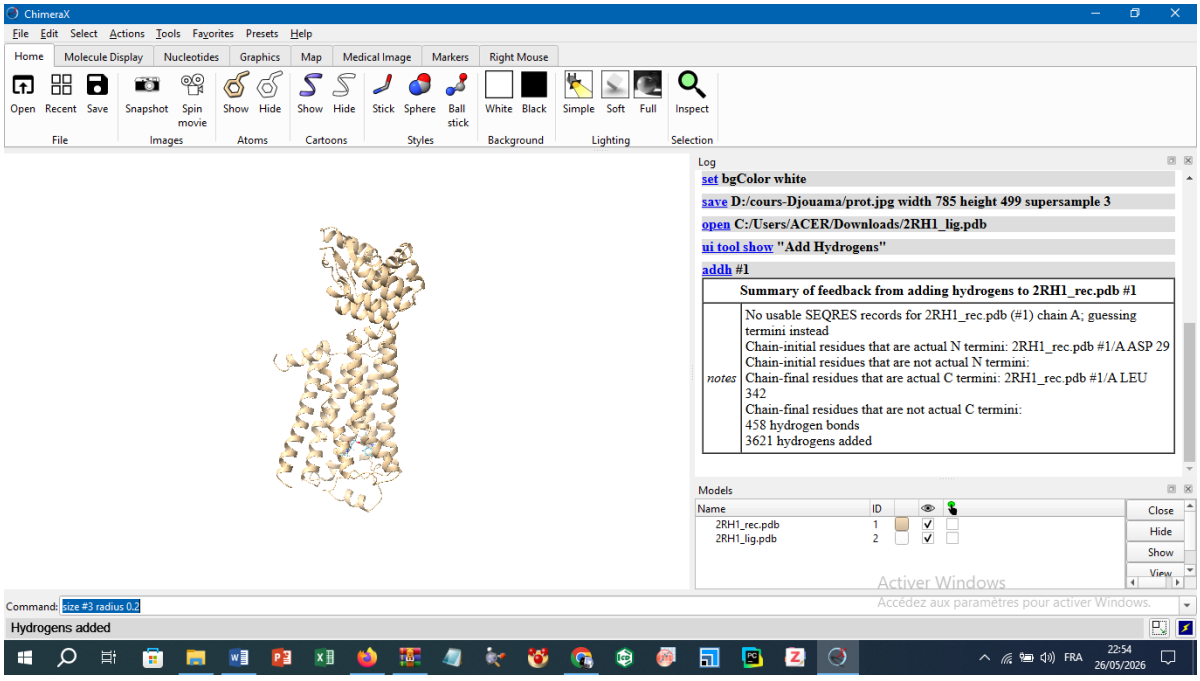
STEP 1: Prepare the receptor file

-Open the **receptor** PDB (2RH1_rec.pdb) file in Chimera



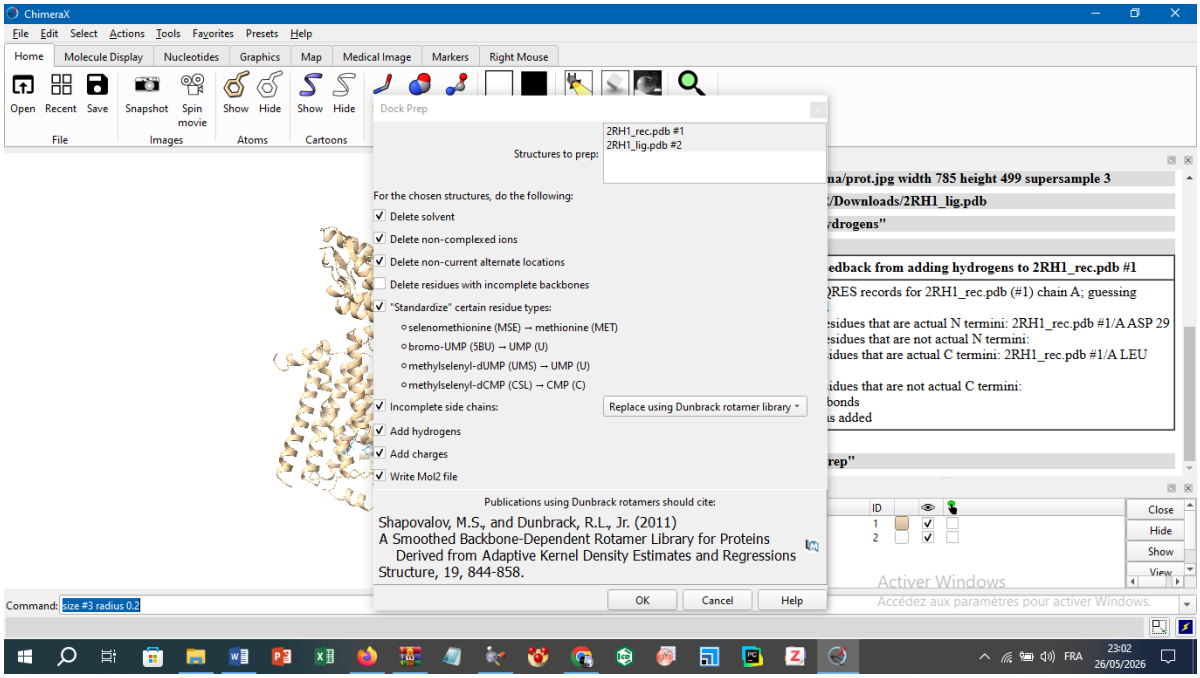
-Select the method for adding hydrogens; in this case we will allow the hydrogen to be optimized by the hydrogen bonding network, and we will allow the method to determine the protonation state





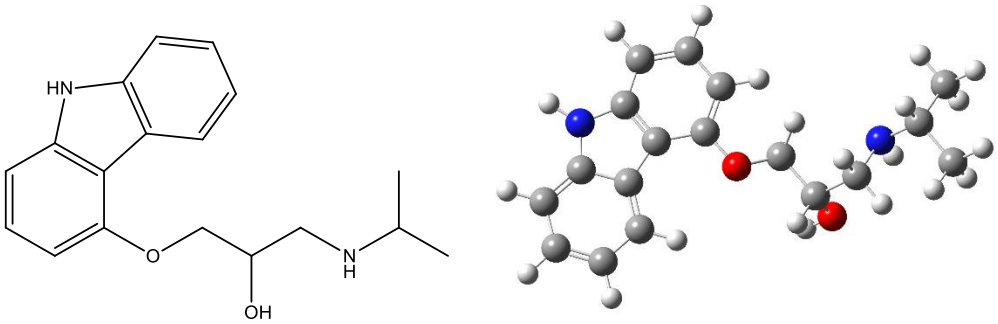
-Save the receptor in mol2 format

The receptor can be saved in mol2 format. The Dock Prep procedure should be run again to incorporate the mutated residues. Make sure the Write Mol2 box is checked at this point.



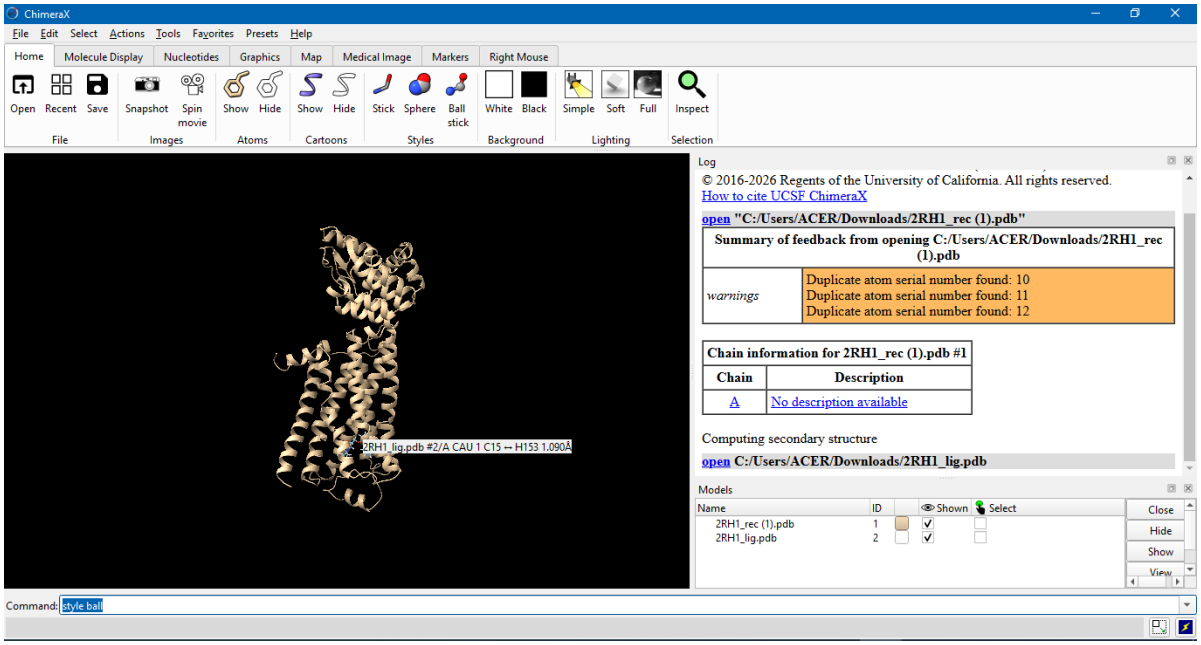
STEP 2: Prepare the ligand file. The ligands (carazolol; (2RH1_lig.pdb)

The structure is derived from the 2RH1 crystal structure and are also available for download at http://www.rcsb.org

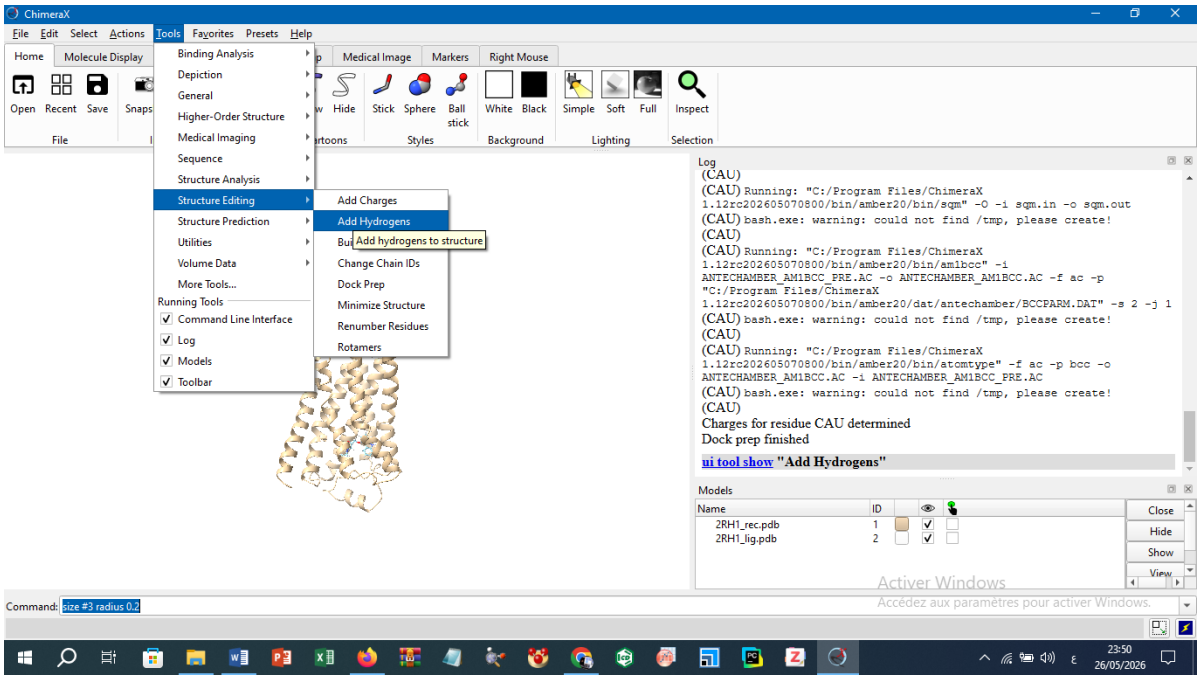


carazolol

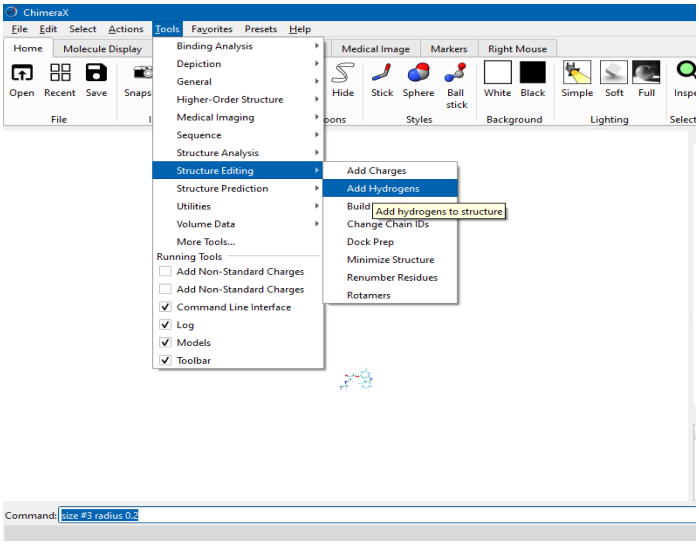
- Open the 2RH1_lig.pdb file in Chimera



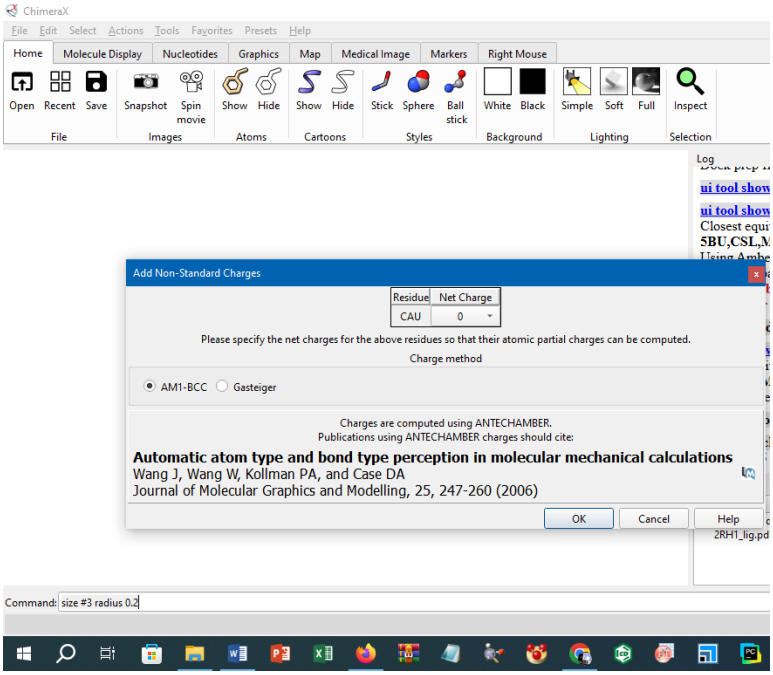
-Add hydrogens



-To complete this option, first activate the Add Charge tool



-To seed the charge calculation، Chimera needs the formal charge of the molecule. Chimera will estimate the value based on the atom types and bonding. Here, Chimera has estimated accurately, so we will calculate AM1-BCC charges.



-Save the molecule in mol2 format.

-Report

1. Why is adding hydrogen atoms (Hydrogens) an essential step that cannot be overlooked during protein preparation?

2. What is the role of partial charges (Partial Charges) in determining the accuracy of the binding site between Carazolol and the receptor?

Reference

[1] École Normale Supérieure (ENS Paris), Rappels de Chimie Théorique (Agrégation), 2019. ens.fr.

[2] B. Arsic, Title of the paper, Rev. Chim., Récupéré de revistadechimie.ro.

[3] K. Yates, Huckel molecular orbital theory, Academic Press, New York, 1978.

[4] S. Sardar, Physical Chemistry Topic: Quantum Chemistry (LCAO and HF-SCF), Récupéré de hal.science.

[5] I. Fleming, Molecular Orbitals and Organic Chemical Reactions, Wiley, 2010.

[6] F. Menezes, E. Slugocka, A. Wieckowska, J. Chem. Educ. 102 (2025) 2144-2150.

[7] ScienceDirect, Molecular Mechanics, Topic page, Récupéré de sciencedirect.com.

[8] ScienceDirect, Computational Chemistry, Book chapter excerpt, Récupéré de sciencedirect.com.

[9] De Gruyter Brill, Glossary of terms used in computational drug design, IUPAC Recommendations (1996).

[10] Compute Canada, Molecular modeling and molecular dynamics theory lesson for novices. github.io.

[11] A.K. Rappe, C.J. Casewit, K.S. Colwell, W.A. Goddard, W.M. Skiff, J. Am. Chem. Soc. 114 (1992) 10024-10035.

[12] D. F. Asli, D. S. Khamouli, P. M. T. Rehman, P. M. F. AlAjmi, Discovery of potent parthenolide derivatives against triple-negative breast cancer using 3D-QSAR, virtual screening, and molecular dynamics, ChemistrySelect 10 (2025).

[13] Z. Wang, W. Zhang, C. Wu, H. Lei, P. Cieplak, Y. Duan, J. Comput. Chem. 27 (2006) 781-790.

[14] University of Illinois, Force Fields (San Francisco Lectures). uiuc.edu.

[15] IIT Kanpur, Chapter 2: Regression - Simple Linear Regression Analysis. iitk.ac.in.

[16] Imarticus Learning, Simple Linear Regression. imarticus.org.

[17] A.B. Rowaiye, et al., Discov. Chem. 2 (2025) 281.

[18] U. Muhammad, A. Uzairu, D. Ebuka Arthur, J. Anal. Pharm. Res. 7 (2018).

[19] In Silico Design, History and basic concepts of molecular modeling. insilicodesign.com.

[20] R. Han, H. Yoon, G. Kim, H. Lee, Y. Lee, Pharmaceuticals 16 (2023) 1259.