SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING
DEPARTMENT OF BIOINFORMATICS

UNIT – 1- SBI1208 – Cheminformatics

Computational chemistry

Computational chemistry is a branch of chemistry that uses computer simulation to assist in solving chemical problems. It uses methods of theoretical chemistry, incorporated into efficient computer programs, to calculate the structures and properties of molecules and solids. It is necessary because, apart from relatively recent results concerning the hydrogen molecular ion (dihydrogen cation, see references therein for more details), the quantum many-body problem cannot be solved analytically, much less in closed form. While computational results normally complement the information obtained by chemical experiments, it can in some cases predict hitherto unobserved chemical phenomena. It is widely used in the design of new drugs and materials.

Examples of such properties are structure (i.e., the expected positions of the constituent atoms), absolute and relative (interaction) energies, electronic charge density distributions, dipoles and higher multipole moments, vibrational frequencies, reactivity, or other spectroscopic quantities, and cross sections for collision with other particles.

The methods used cover both static and dynamic situations. In all cases, the computer time and other resources (such as memory and disk space) increase rapidly with the size of the system being studied. That system can be one molecule, a group of molecules, or a solid. Computational chemistry methods range from very approximate to highly accurate; the latter are usually feasible for small systems only. *Ab initio* methods are based entirely on quantum mechanics and basic physical constants. Other methods are called empirical or semi-empirical because they use additional empirical parameters.

Both *ab initio* and semi-empirical approaches involve approximations. These range from simplified forms of the first-principles equations that are easier or faster to solve, to approximations limiting the size of the system (for example, periodic boundary conditions), to fundamental approximations to the underlying equations that are required to achieve any solution to them at all. For example, most *ab initio* calculations make the Born–Oppenheimer approximation, which greatly simplifies the underlying Schrödinger equation by assuming that the nuclei remain in place during the calculation. In principle, *ab initio* methods eventually converge to the exact solution of the underlying equations as the number of approximations is reduced. In practice, however, it is impossible to eliminate all approximations, and residual error inevitably remains. The goal of computational chemistry is to minimize this residual error while keeping the calculations tractable.

In some cases, the details of electronic structure are less important than the long-time phase space behavior of molecules. This is the case in conformational studies of proteins and protein-ligand binding thermodynamics. Classical approximations to the potential energy surface are used, typically with molecular mechanics force fields, as they are computationally less intensive than electronic calculations, to enable longer simulations of molecular dynamics. Furthermore, cheminformatics uses even more empirical (and computationally cheaper) methods like machine learning based on physicochemical properties. One typical problem in cheminformatics is to predict the binding affinity of drug molecules to a given target. Other problems include predicting binding specificity, off-target effects, toxicity, and pharmacokinetic properties.

The term *theoretical chemistry* may be defined as a mathematical description of chemistry, whereas *computational chemistry* is usually used when a mathematical method is sufficiently well developed that it can be automated for implementation on a computer. In theoretical

chemistry, chemists, physicists, and mathematicians develop algorithms and computer programs to predict atomic and molecular properties and reaction paths for chemical reactions. Computational chemists, in contrast, may simply apply existing computer programs and methodologies to specific chemical questions.

Computational chemistry has two different aspects:

- Computational studies, used to find a starting point for a laboratory synthesis, or to assist in understanding experimental data, such as the position and source of spectroscopic peaks.
- Computational studies, used to predict the possibility of so far entirely unknown molecules or to explore reaction mechanisms not readily studied via experiments.

Thus, computational chemistry can assist the experimental chemist or it can challenge the experimental chemist to find entirely new chemical objects.

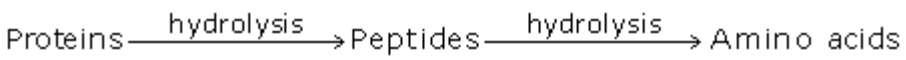
Several major areas may be distinguished within computational chemistry:

- The prediction of the molecular structure of molecules by the use of the simulation of forces, or more accurate quantum chemical methods, to find stationary points on the energy surface as the position of the nuclei is varied.
- Storing and searching for data on chemical entities (see chemical databases).
- Identifying correlations between chemical structures and properties (see *quantitative structure–property relationship* (QSPR) and *quantitative structure–activity relationship* (QSAR)).
- Computational approaches to help in the efficient synthesis of compounds.
- Computational approaches to design molecules that interact in specific ways with other molecules (e.g. drug design and catalysis).

Proteins

Proteins- Properties, Structure, Classification and Functions

- Proteins are the most abundant biological macromolecules, occurring in all cells.
- It is also the most versatile organic molecule of the living systems and occur in great variety; thousands of different kinds, ranging in size from relatively small peptides to large polymers.
- Proteins are the polymers of amino acids covalently linked by the peptide bonds.
- The building blocks of proteins are the twenty naturally occurring amino acids.
- Thus, proteins are the polymers of amino acids.



Properties of Proteins

Solubility in Water

- The relationship of proteins with water is complex.
- The secondary structure of proteins depends largely on the interaction of peptide bonds with water through hydrogen bonds.
- Hydrogen bonds are also formed between protein (alpha and beta structures) and water. The protein-rich static ball is more soluble than the helical structures.

- At the tertiary structure, water causes the orientation of the chains and hydrophilic radicals to the outside of the molecule, while the hydrophobic chains and radicals tend to react with each other within the molecule (hydrophobic effect).

Denaturation and Renaturation

- Proteins can be denatured by agents such as heat and urea that cause unfolding of polypeptide chains without causing hydrolysis of peptide bonds.
- The denaturing agents destroy secondary and tertiary structures, without affecting the primary structure.
- If a denatured protein returns to its native state after the denaturing agent is removed, the process is called renaturation.

Some of the denaturing agents include

Physical agents: Heat, radiation, pH

Chemical agents: Urea solution which forms new hydrogen bonds in the protein, organic solvents, detergents.

Isoelectric point

- The isoelectric point (pI) is the pH at which the number of positive charges equals the number of negative charges, and the overall charge on the amino acid is zero.
- At this point, when subjected to an electric field the proteins do not move either towards anode or cathode, hence this property is used to isolate proteins.

Molecular weight

- The average molecular weight of an amino acid is taken to be 110.
- The total number of amino acids in a protein multiplied by 110 gives the approximate molecular weight of that protein.
- Different proteins have different amino acid composition and hence their molecular weights differ.
- The molecular weights of proteins range from 5000 to 10^9 Daltons.

Posttranslational modifications

- It occurs after the protein has been synthesized on the ribosome.
- Phosphorylation, glycosylation, ADP ribosylation, methylation, hydroxylation, and acetylation affect the charge and the interactions between amino acid residues, altering the three-dimensional configuration and, thus, the function of the protein.

Protein structure

- The linear sequence of amino acid residues in a polypeptide chain determines the three-dimensional configuration of a protein, and the structure of a protein determines its function.
- All proteins contain the elements carbon, hydrogen, oxygen, nitrogen and sulfur some of these may also contain phosphorus, iodine, and traces of metals like ion, copper, zinc and manganese.
- A protein may contain 20 different kinds of amino acids. Each amino acid has an amine group at one end and an acid group at the other and a distinctive side chain.
- The backbone is the same for all amino acids while the side chain differs from one amino acid to the next.

The structure of proteins can be divided into four levels of organization:

1. Primary Structure

- The primary structure of a protein consists of the amino acid sequence along the polypeptide chain.
- Amino acids are joined by peptide bonds.

- Because there are no dissociable protons in peptide bonds, the charges on a polypeptide chain are due only to the N-terminal amino group, the C-terminal carboxyl group, and the side chains on amino acid residues.
- The primary structure determines the further levels of organization of protein molecules.

2. Secondary Structure

- The secondary structure includes various types of local conformations in which the atoms of the side chains are not involved.
- Secondary structures are formed by a regular repeating pattern of hydrogen bond formation between backbone atoms.
- The secondary structure involves α-helices, β-sheets, and other types of folding patterns that occur due to a regular repeating pattern of hydrogen bond formation.
- The secondary structure of protein could be :

1. **Alpha-helix**
2. **Beta-helix**

- The α-helix is a right-handed coiled strand.
- The side-chain substituents of the amino acid groups in an α-helix extend to the outside.
- Hydrogen bonds form between the oxygen of the C=O of each peptide bond in the strand and the hydrogen of the N-H group of the peptide bond four amino acids below it in the helix.
- The side-chain substituents of the amino acids fit in beside the N-H groups.
- The hydrogen bonding in a ß-sheet is between strands (inter-strand) rather than within strands (intra-strand).
- The sheet conformation consists of pairs of strands lying side-by-side.
- The carbonyl oxygens in one strand hydrogen bond with the amino hydrogens of the adjacent strand.
- The two strands can be either parallel or anti-parallel depending on whether the strand directions (N-terminus to C-terminus) are the same or opposite.
- The anti-parallel ß-sheet is more stable due to the more well-aligned hydrogen bonds.

3. Tertiary Structure

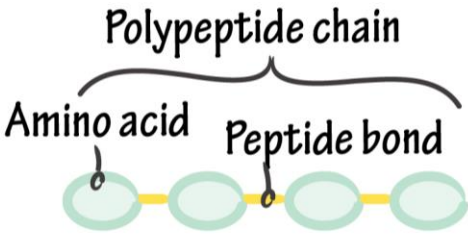
- Tertiary structure of a protein refers to its overall three-dimensional conformation.
- The types of interactions between amino acid residues that produce the three-dimensional shape of a protein include hydrophobic interactions, electrostatic interactions, and hydrogen bonds, all of which are non-covalent.
- Covalent disulfide bonds also occur.
- It is produced by interactions between amino acid residues that may be located at a considerable distance from each other in the primary sequence of the polypeptide chain.
- Hydrophobic amino acid residues tend to collect in the interior of globular proteins, where they exclude water, whereas hydrophilic residues are usually found on the surface, where they interact with water.

4. Quaternary Structure

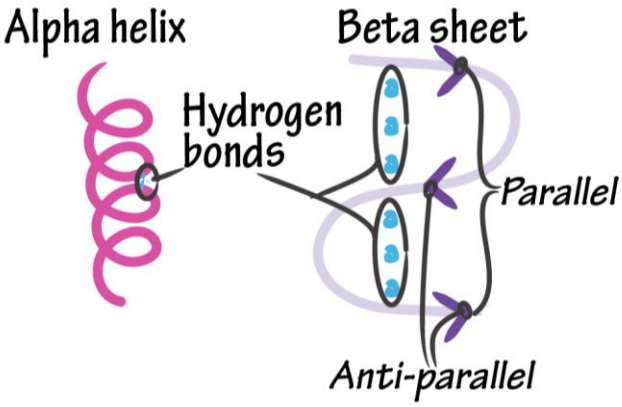
- Quaternary structure refers to the interaction of one or more subunits to form a functional protein, using the same forces that stabilize the tertiary structure.
- It is the spatial arrangement of subunits in a protein that consists of more than one polypeptide chain.

CLASSES OF PROTEIN STRUCTURE

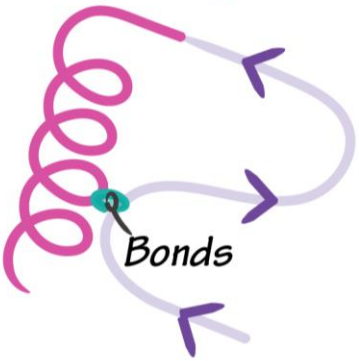
Primary



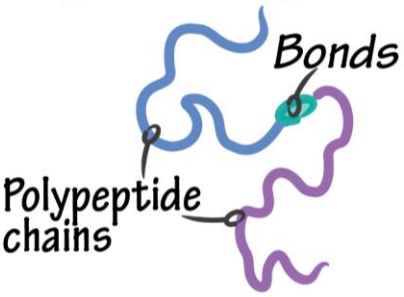
Secondary

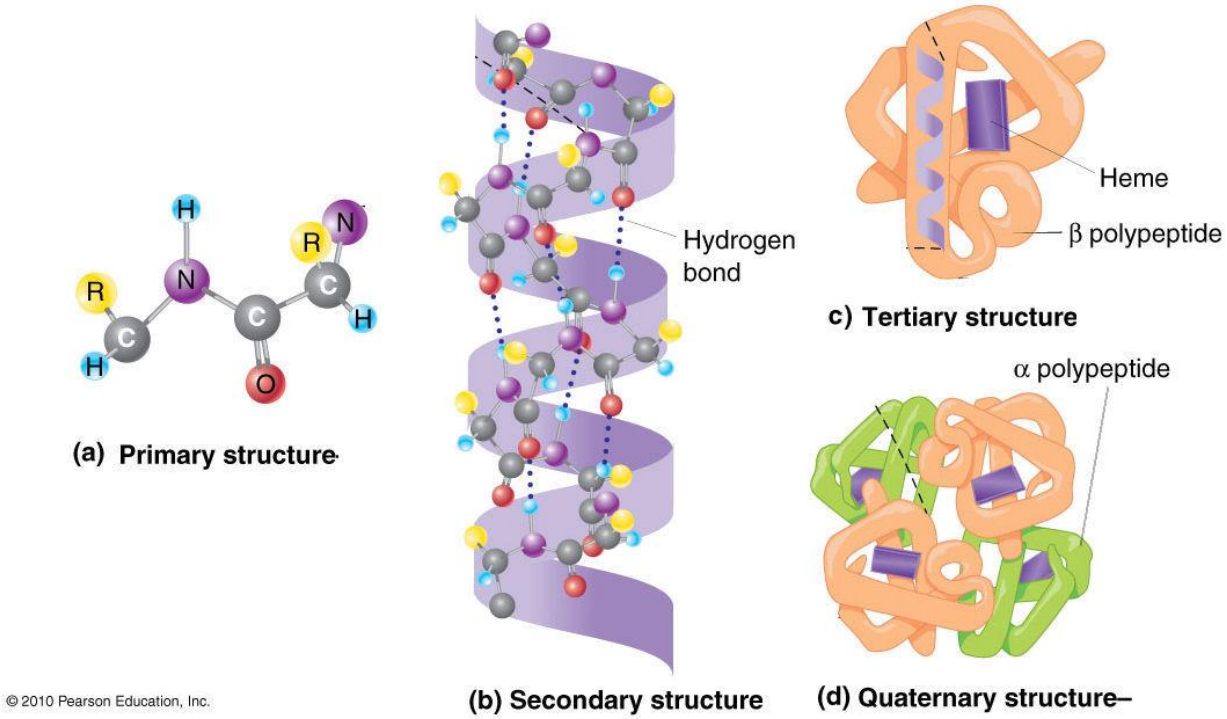


Tertiary

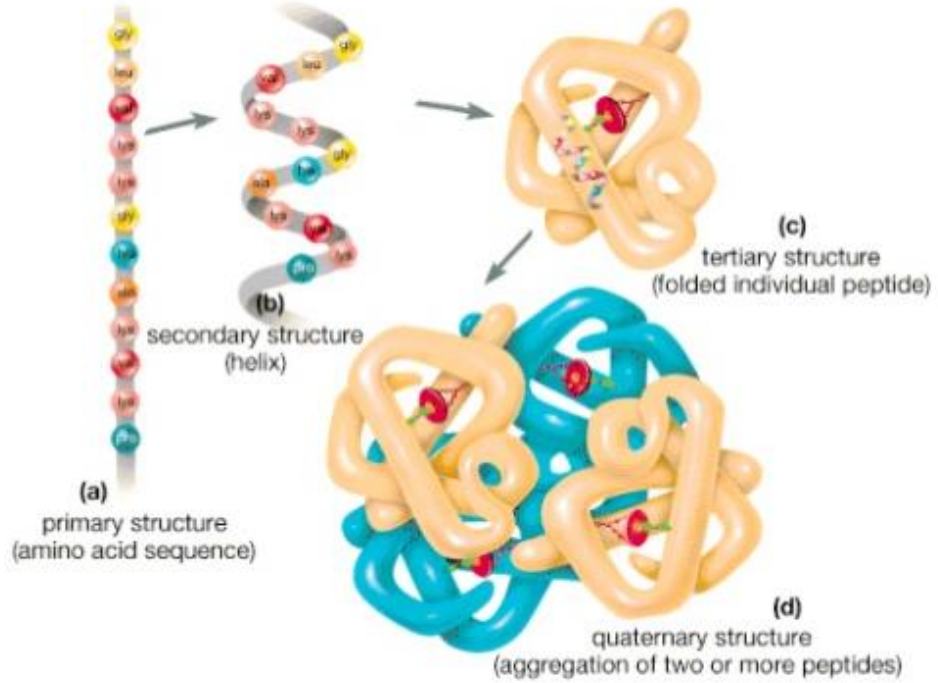


Quaternary





Four Levels of Protein Structure



Functions of proteins

Proteins are vital for the growth and repair, and their functions are endless. They also have enormous diversity of biological function and are the most important final products of the information pathways.

- Proteins, which are composed of amino acids, serve in many roles in the body (e.g., as enzymes, structural components, hormones, and antibodies).
- They act as structural components such as keratin of hair and nail, collagen of bone etc.

- Proteins are the molecular instruments through which genetic information is expressed.
- They execute their activities in the transport of oxygen and carbon dioxide by hemoglobin and special enzymes in the red cells.
- They function in the homostatic control of the volume of the circulating blood and that of the interstitial fluids through the plasma proteins.
- They are involved in blood clotting through thrombin, fibrinogen and other protein factors.
- They act as the defence against infections by means of protein antibodies.
- They perform hereditary transmission by nucleoproteins of the cell nucleus.
- Ovalbumine, glutelin etc. are storage proteins.
- Actin, myosin act as contractile protein important for muscle contraction

Aminoacids

Amino acids Amino acids are the building block of proteins. Amino acids are important organic compounds that contain amine (-NH2) and Carboxyl (-COOH) functional groups, along with a side-chain (R group) that is specific for each amino acid (Figure 1). Twenty different amino acids are commonly found in proteins. All of these 20 common amino acids are α-amino acids except proline and their general structure is shown below. They have a carboxyl group and amino group which are covalently bonded to a α-carbon atom. They differ from each other in their side chain R groups. Since, the remaining structure are same therefore properties of these amino acids are primarily determined by the side chain groups. The nature of these side chain maybe polar, nonpolar (aliphatic), hydrophilic, hydrophobic, acidic, basic and aromatic. These amino acids have been abbreviated using either three letter word or one letter word (Table 1).

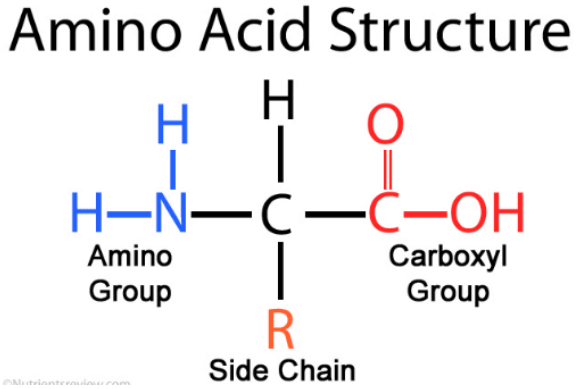


Figure 1: Structure of amino acid containing R side chain.

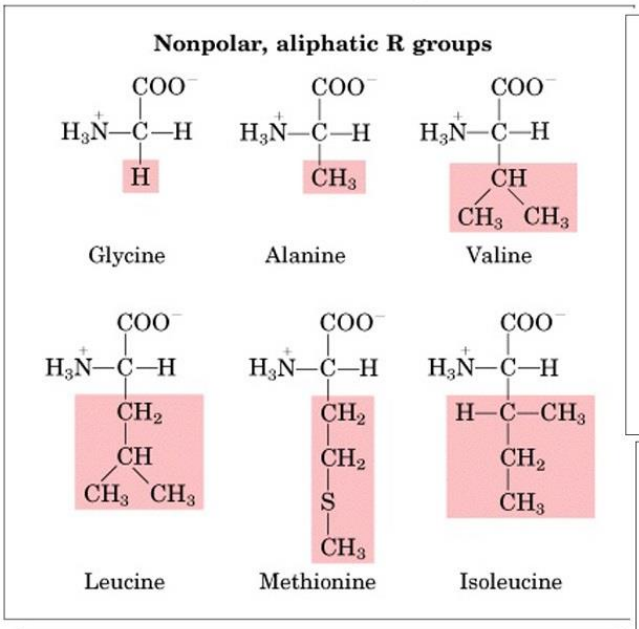
Classification

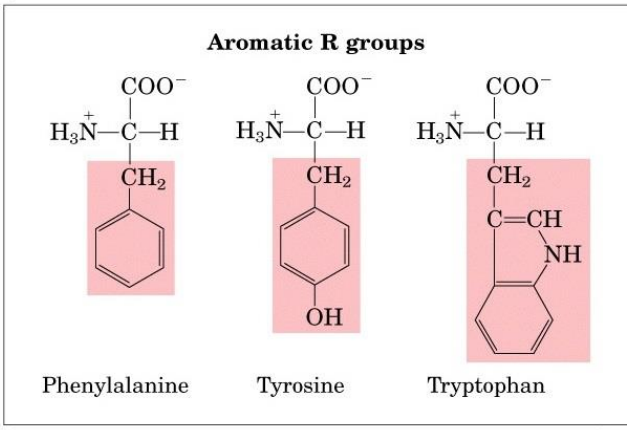
The 20 amino acids have been classified using different criteria by different scientists. For instance, they have been classified as polar, nonpolar, hydrophilic, hydrophobic, acidic, basic, aliphatic and aromatic. Here, we have classified all of these 20 common set of amino acids into six distinct classes. Nonpolar (Aliphatic) amino acids The R side chain in this class of amino acids including alanine, valine, leucine and isoleucine are hydrophobic in nature therefore they stabilize the protein structure through hydrophobic interactions. Glycine is also classified as nonpolar amino acids, but it has very small side chain. Therefore, it does not contribute to hydrophobic interactions. Glycine has the simplest structure. The side chain of proline has a distinctive cyclic structure which is an imino group held in a rigid conformation, therefore it reduces the structural flexibility of particularly that regions of polypeptide chain where it occurs.

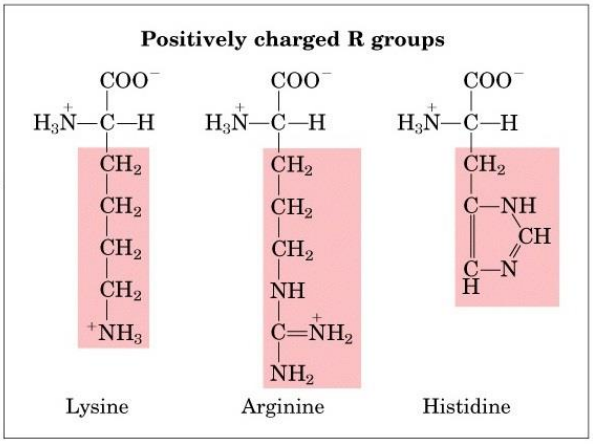
Aromatic amino acids (Phenylalanine, Tyrosine and Tryptophan)
The side chain of aromatic amino acids contains an aromatic ring (Figure 3) which are relatively nonpolar (hydrophobic) in nature. These amino acids can participate in hydrophobic interaction. Tyrosine and tryptophan are much more polar than phenylalanine owing to their hydroxyl and nitrogen indole ring respectively. These amino acids show light absorption in the ultraviolet range due to the presence of conjugated double bond-single bond system.

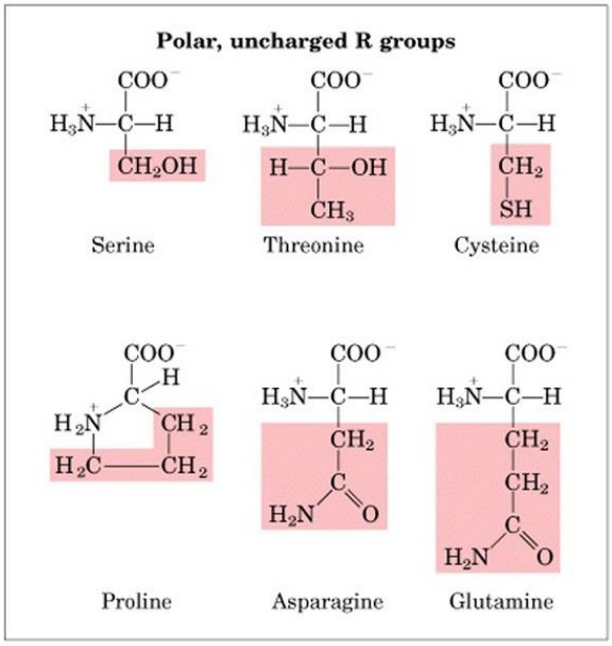
Polar, uncharged amino acids
This class of amino acids includes serine, threonine, cysteine, asparagine and glutamine. The R group of these amino acids are more soluble in water or more hydrophilic than those of nonpolar amino acids because they contain functional groups (OH, SH, $CONH_2$) that form hydrogen bonds with water. The polarity of serine and threonine is contributed by their hydroxyl groups, and that of cysteine and tryptophan by sulfhydryl and indole ring respectively which is weakly hydrogen bonded with oxygen and nitrogen respectively. Furthermore, polarity of asparagine and glutamine is contributed by their amide group.
Acidic amino acids
These amino acids contain two carboxyl groups, one α- carboxyl and other β- or γ-carboxyl group. Since they contain two acidic groups (one α-carboxyl group + one β or γ-carboxyl group) and one basic group (α-amino group), the net charge of these amino acids is therefore acidic and they are negatively charged at physiological pH.

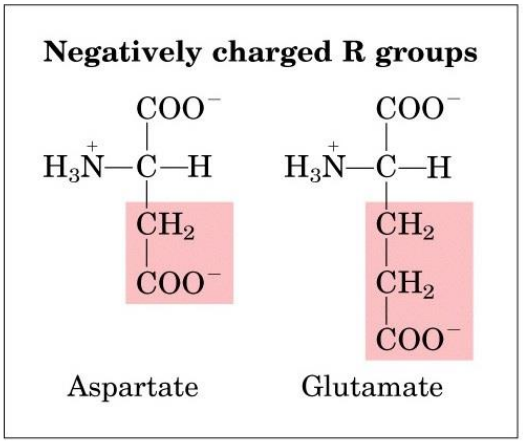
Basic amino acids
The basic amino acid contains an α-amino group and the side chain contains second amino/ imino group (imidazole, ε-amino or guanidine group). These amino acids are histidine, lysine and arginine. Since these amino acids contain two basic groups one acidic group (α-carboxyl group), therefore the net behavior of these amino acids is basic and they are positively charged at physiological pH.


Nonpolar, aliphatic R groups
Glycine
Alanine
Valine
Leucine
Methionine
Isoleucine


Aromatic R groups
Phenylalanine
Tyrosine
Tryptophan


Positively charged R groups
Lysine
Arginine
Histidine


Polar, uncharged R groups
Serine
Threonine
Cysteine
Proline
Asparagine
Glutamine


Negatively charged R groups
Aspartate
Glutamate

Acid-base character of amino acids

For explaining the acid-base character, let us consider neutral (aliphatic) amino acid alanine which is nonpolar that was discovered in 1923. Its carboxyl group can be deprotonated (Figure 7).

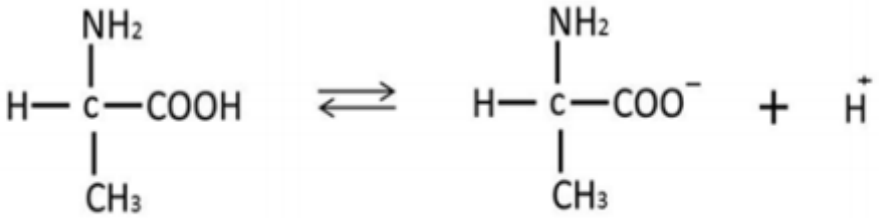


Figure 7: Carboxyl group of amino acid donates a proton.

Similarly, when amino acid accepts a proton (due to the presence of a basic amino group), the amino acid acquires a positive charge (Figure 8).

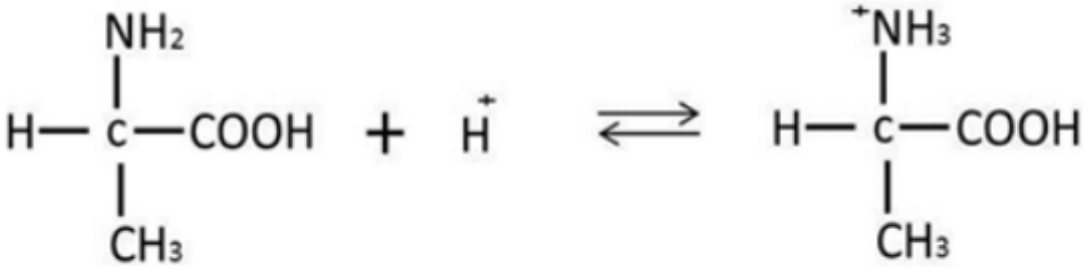


Figure 8: Amino group of amino acid accepts proton.

If we consider the acidity of amino acid, it releases proton which will be taken up by the solvent, water or by the basic amino group available on the amino acid. Since amino group is more basic, it takes the proton donated by carboxyl group. As a result carboxyl group acquires a negative charge whereas a positive charge develops on the amino group. Since this form of amino acid has both positive and negative charges, therefore the net charge of the amino acid is zero. This state of amino acid is known as zwitterionic state (Figure 9).

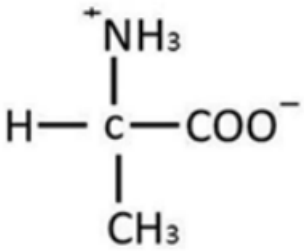


Figure 9: Zwitterionic state of alanine.

In zwitterions form, the carboxylate group acts as a base and the protonated amino group acts as an acid as shown below: -COO- + H+ → -COOH + -NH3 + OH→-NH2 + H2 O Since this type of amino acid is capable of acting as both acid and base, this implies that amino acid can act as buffer

Peptide bonds

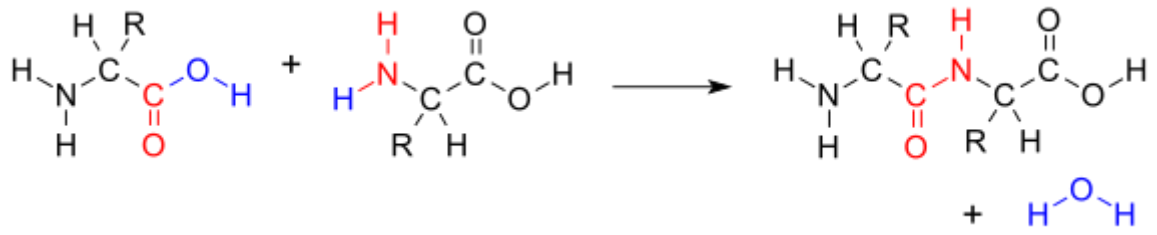
A **peptide bond** is an amide type of covalent chemical bond linking two consecutive alpha-amino acids from C1 (carbon number one) of one alpha-amino acid and N2 (nitrogen number two) of another, along a peptide or protein chain.

It can also be called an **eupeptide bond** to separate it from an isopeptide bond, a different type of amide bond between two amino acids.

Synthesis

When two amino acids form a *dipeptide* through a *peptide bond* it is a type of condensation reaction.[2] In this kind of condensation, two amino acids approach each other, with the non-side chain (C1) carboxylic acid moiety of one coming near the non-side chain (N2) amino moiety of the other. One loses a hydrogen and oxygen from its carboxyl group (COOH) and the other loses a hydrogen from its amino group (NH_2). This reaction produces a molecule of water (H_2O) and two amino acids joined by a peptide bond (-CO-NH-). The two joined amino acids are called a dipeptide.

The amide bond is synthesized when the carboxyl group of one amino acid molecule reacts with the amino group of the other amino acid molecule, causing the release of a molecule of water (H_2O), hence the process is a dehydration synthesis reaction.



The dehydration condensation of two amino acids to form a peptide bond (red) with expulsion of water (blue).

The formation of the peptide bond consumes energy, which, in organisms, is derived from ATP. Peptides and proteins are chains of amino acids held together by peptide bonds (and sometimes by a few isopeptide bonds). Organisms use enzymes to produce nonribosomal peptides, and ribosomes to produce proteins via reactions that differ in details from dehydration synthesis.

Some peptides, like alpha-amanitin, are called ribosomal peptides as they are made by ribosomes,[6] but many are nonribosomal peptides as they are synthesized by specialized enzymes rather than ribosomes. For example, the tripeptide glutathione is synthesized in two steps from free amino acids, by two enzymes: glutamate–cysteine ligase (forms an isopeptide bond, which is not a peptide bond) and glutathione synthetase (forms a peptide bond).

Degradation

A peptide bond can be broken by hydrolysis (the addition of water). In the presence of water they will break down and release 8–16 kilojoule/mol (2–4 kcal/mol) of Gibbs energy This process is extremely slow, with the half-life at 25°C of between 350 and 600 years per bond.

In living organisms, the process is normally catalyzed by enzymes known as peptidases or proteases, although there are reports of peptide bond hydrolysis caused by conformational

strain as the peptide/protein folds into the native structure. This non-enzymatic process is thus not accelerated by transition state stabilization, but rather by ground state destabilization.

Favorable Interactions in Proteins – non-covalent

• Hydrophobic effect – Release of water molecules from the structured solvation layer around the molecule as protein folds increases the net entropy

• Hydrogen bonds – Interaction of N-H and C=O of the peptide bond leads to local regular structures such as α-helices and β-sheets

• London dispersion – Medium-range weak attraction between all atoms contributes significantly to the stability in the interior of the protein

• Electrostatic interactions – Long-range strong interactions between permanently charged groups – Salt-bridges, esp. buried in the hydrophobic environment strongly stabilize the protein

Protein stability

Protein stability is the net balance of forces, which determine whether a protein will be in its native folded conformation or a denatured state. Protein stability normally refers to the physical (thermodynamic) stability, not the chemical stability.

Forces involve in protein stabilization

- Hydrogen Bonding.
- Vander Waals interactions.
- Ionic strengths.
- Disulfide bonds
- Hydrophobicity: the dominant force in protein folding Forces involved in Protein stabilization

Factor affecting protein stability

Temperature.

- Extreme temperature make protein unstable.

pH

- Extreme pH cause unstabilty in protein.

Organic Solvent.

- Unstable the protein

Chaotropic agent.

- Urea and guanidinium hydrochloride.
- Destroy the tertiary structure.

Disulfide bond

- If their disulfides are broken (i.e. reduced) and then carboxymethylated with iodoacetate, the resulting protein is denatured, i.e. unfolded, or mostly unfolded

Ligand binding

- Ligand binding increases the stability of the protein.

Protein denaturation

- A loss of three-dimensional structure sufficient to cause loss of function is called denaturation

- Alterations in the environment (pH, salt concentration, temperature etc.) disrupt bonds and forces of attraction.

Protein folding

- A folding protein follows multiple pathways from high energy and high Entropy to low energy and low entropy.
- Amyloid diseases result from protein misfolding.

Factor of Protein misfolding

- Absence of normal supporting/co factors
- Absence of chaperone protein
- Change in temp and pH

Factors of protein misfolding

- Mutations
- Premature termination of Translation
- Fault in post-translational modifications
- Strong Promoters
- High Inducer concentrations

Reasons for protein misfolding

- Loss of conformation due to stress

Chaperon

- These are protein molecule Assist in protein folding.
- Prevent aggregation.
- Examples GroEL is bacterial chaperone HSP in eukaryotes

Anfinsen experiment –Spontaneous folding

- Ribonuclease is a small protein that contain 8 cysteine linked via four disulfide bond Urea in presence of 2- mercaptoethanol fully denature ribonuclease. When urea and 2 mercaptoethanol are removed the protein spontaneously refolds and correct disulfide bonds are formed. The sequence alone determines the native conformation. Awarded Nobel prize in 1972.

Renaturation

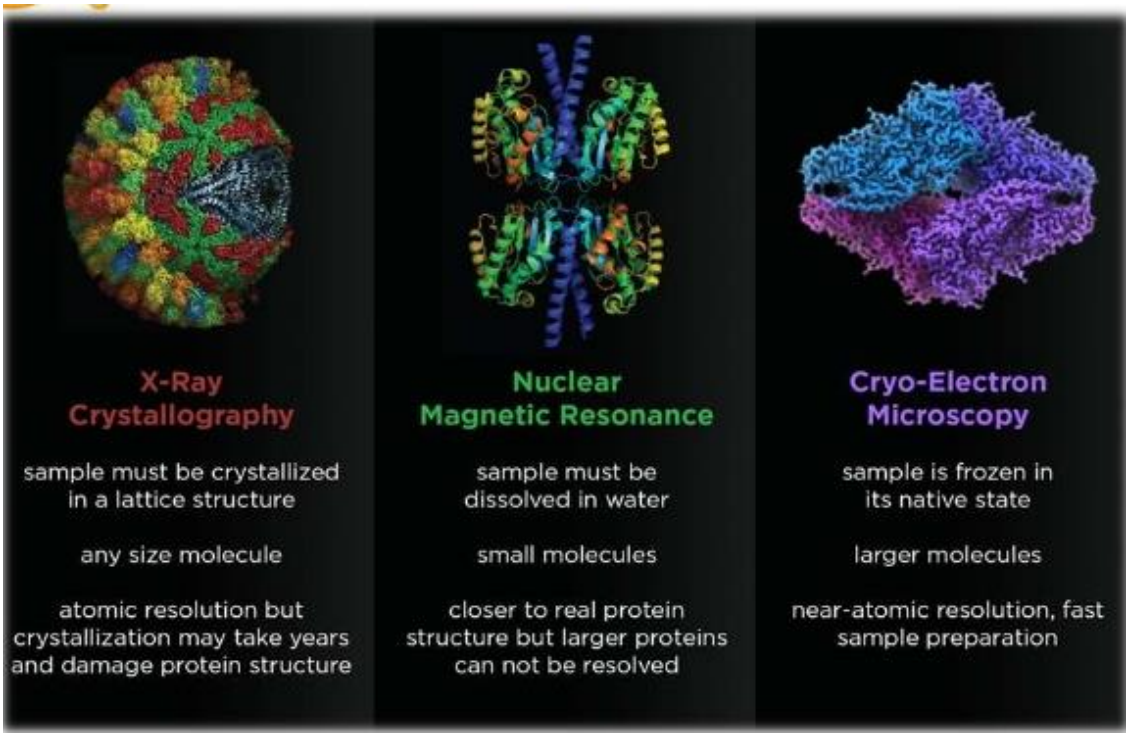
- Native structure and biological activity of some globular proteins can be regained if the denaturing agent will be removed.
- Ribonuclease present a classical example of renaturation.

Protein structure determination

Around 90% of the protein structures available in the Protein Data Bank have been determined by X-ray crystallography. This method allows one to measure the three-dimensional (3-D) density distribution of electrons in the protein, in the crystallized state, and thereby infer the 3-D coordinates of all the atoms to be determined to a certain resolution. Roughly 9% of the known protein structures have been obtained by nuclear magnetic resonance (NMR) techniques. For larger proteincomplexes, cryo-electron microscopy can determine protein structures. The resolution is typically lower than that of X-ray crystallography, or NMR, but the maximum resolution is steadily increasing. This technique

is still a particularly valuable for very large protein complexes such as virus coat proteins and amyloid fibers.

General secondary structure composition can be determined via circular dichroism. Vibrational spectroscopy can also be used to characterize the conformation of peptides, polypeptides, and proteins.[26] Two-dimensional infrared spectroscopy has become a valuable method to investigate the structures of flexible peptides and proteins that cannot be studied with other methods.[27][28] A more qualitative picture of protein structure is often obtained by proteolysis, which is also useful to screen for more crystallizable protein samples. Novel implementations of this approach, including fast parallel proteolysis (FASTpp), can probe the structured fraction and its stability without the need for purification. Once a protein's structure has been experimentally determined, further detailed studies can be done computationally, using molecular dynamic simulations of that structure.[



	Advantages	**Disadvantages**	**Objects**	**Resolution**
X-ray Crystallography	• Well developed • High resolution • Broad molecular weight range • Easy for model building	• Difficult for crystallization • Difficult for diffraction • Solid structure preferred • Static crystalline state structure	• Crystallizable samples • Soluble proteins, membrane proteins, ribosomes, DNA/RNA and protein complexes	High

NMR	• High resolution • 3D structure in solution • Good for dynamic study	• Need for high sample purity • Difficult for sample preparation • Difficult for computational simulation	• MWs below 40–50 kDa • Water soluble samples	High
Cryo-EM	• Easy sample preparation • Structure in native state • Small sample size	• Relatively low resolution • Applicable to samples of high molecular weights only • Highly dependent on EM techniques • Costly EM equipment	• >150 kDa • Virions, membrane proteins, large proteins, ribosomes, complex compounds	Relatively Low (<3.5 Å)

Conformational Change

In biochemistry, a **conformational change** is a change in the shape of a macromolecule, often induced by environmental factors.

A macromolecule is usually flexible and dynamic. It can change its shape in response to changes in its environment or other factors; each possible shape is called a conformation, and a transition between them is called a *conformational change*. Factors that may induce such changes include temperature, pH, voltage, light in chromophores, ion concentration, phosphorylation, or the binding of a ligand. Transitions between these states occur on a variety of length scales (tenths of Å to nm) and time scales (ns to s), and have been linked to functionally relevant phenomena such as allosteric signaling and enzyme catalysis

Protein conformation is of paramount importance in understanding biomolecular interactions. In the simplest scenario, two molecules may interact with no change in their conformation, as in the key-and-lock model. Molecular interactions that involve conformational changes in the interacting molecules are more versatile. In the induced-fit model, two molecules bind optimally with each other only after conformational changes at their interface. Conformational changes may also take place away from the binding interface. This is often the prerequisite for functional activity. For protein like haemoglobin that shows allosteric behaviour, the binding of small molecules at a region of the protein affects its binding affinity with other molecules at a distant region. In membrane receptors binding of ligand at the extracellular region causes changes at the cytoplasmic region, so that an extracellular signal is allowed to alter intracellular activity.

Conformational changes in proteins are made possible by their intrinsic flexibility. These changes may occur with only relatively small expenditure of energy. At the molecular

structural level, conformational changes in single polypeptides are the result of changes in main chain torsional angles and side chain orientations. The overall effect of such changes may be localised with reorientations of a few residues and small torsional changes in the regional main chain. On the other hand torsional changes localised at very few critically placed residues may lead to large changes in tertiary structure. The later type of conformational changes is described as domain motions.

Domain motions

Domain motions have two basic components. Hinge motions may occur within strands, beta-sheets and alpha-helices not constrained by tertiary packing forces. To qualify as fulcrum for hinge-motion, residue must bear very little tertiary structure packing constraints on its main chain. The hinge lies outside the interface between the two domains inter-connected by the hinge. On hinge-opening the motion is perpendicular to the plane of the interface, which is lost after opening. The closed conformation is usually stabilised by a bound ligand. This is necessarily so, for if the closed conformation is strongly held together without a ligand, then the hinge opening will have to cross a high energy barrier. Shear motions occur parallel to the interface between closely packed segments of polypeptides. This type of motion is more severely constrained with additional packing contacts due to interdigitating side chains. A large enough sheared domain motion is due to the combination of a number of shear movements. Proteins that shear often have layered architecture, with shearing that may occur across helix-helix, helix-sheet, helix-loop and sheet-loop interfaces. Helix-helix shearing is the predominant type, usually between crossed helices with interhelical angles of 60^0-90^0.

Hinged-motion occurs in the context of secondary structure interactions. Hinge motion at extended strand involves a few large changes in main chain torsion angles at the hinge connecting two domains, constrained only by the Ramachandran allowance of torsional angles. As the range of (phi, psi) angles is relatively large for extended strand, the hinge angle can change by up to 60^0 with only torsional changes in two residues. In beta-sheets two adjacent strands can move like hinges of a door, with extra constraint of hydrogen bonds that hold the sheet together. To obtain the same hinge angular change, torsional changes at three or more residues are required. Alpha helices are further constrained with their more restrictive hydrogen bonding, thereby in need of more small-amplitude torsional angular changes to bend themselves significantly. Proline-kinked helix may alow larger torsional angular changes. Torsional angular changes may stretch an alpha helix by about 3 Angstroms into a 3_{10} helix. There are also cases where a long helix may split into two smaller helices inter-connected by a short extended strand that was previously in helical conformation.

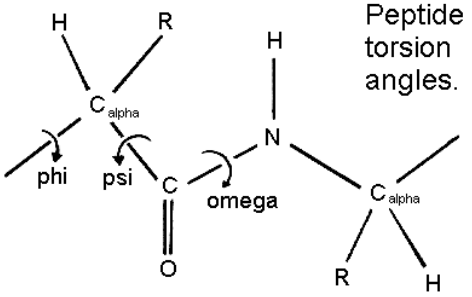
Shear-motion occurs in the context of tertiary structure interactions. Large shear movement that make the interdigitating lock from one state to another is not observed in domain motions, as in subunit interface of allosteric proteins. On the other hand small shear movements that do not require interdigitational repacking are common in domain motions. These shear movements are accommodated by small changes in side chain torsional angles with no significant deformation in main chain torsional configuration of the interface segments. As a consequence the shear-motion causes the segments to shift and rotate relative to each other for up to 2 Angstroms and 15^0 , respectively.

Allosteric transitions

Multimeric proteins have an extra dimensionality to conformational transitions due to their quarternary structure. Haemoglobin is the classic prototype of allosteric proteins with cooperative behaviour. In the case of lamprey haemoglobulin, cooperativity is mediated by reversible dissociation and association of subunits. Packing at subunit interfaces are broken off all together. As for human haemoglobin this is achieved by equilibrium between two alternative quarternary structures of the tetramer. The overall structure changes are due to breaking and formation of electrostatic interactions at the tertiary and quaternary levels, as a result of binding to oxygen or other allosteric effectors. At the interface there is large shear motion that involves repacking in transition between tense and relax states. Cooperativity can also be realised singularly by saving expenditure of entropic energy, as in the binding of dimeric *trp* repressor and immunoglobulins to *operator* and antigen, respectively. After binding of one monomer to a binding site on the target molecule with energy expenditure to pay for the decrease in entropy, the binding of the other monomer to adjacent binding site on the same ligand molecule requires less energy for entropy reduction because the first binding has juxtapose the two interacting molecules, such that the second monomer is already placed in the favourable position to interact with the second binding site without having to increase the order of the bimolecular complexity much more.

Validation or assessment of protein structures - Ramachandran plot

In biochemistry, a **Ramachandran plot** (also known as a **Rama plot**, a **Ramachandran diagram** or a **[φ,ψ] plot**), originally developed in 1963 by G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan,[1] is a way to visualize energetically allowed regions for backbone dihedral angles ψ against φ of amino acid residues in protein structure. The figure on the left illustrates the definition of the φ and ψ backbone dihedral angles[2] (called φ and φ' by Ramachandran). The ω angle at the peptide bond is normally 180°, since the partial-double-bond character keeps the peptide planar.[3] The figure in the top right shows the allowed φ,ψ backbone conformational regions from the Ramachandran et al. 1963 and 1968 hard-sphere calculations: full radius in solid outline, reduced radius in dashed, and relaxed tau (N-Cα-C) angle in dotted lines.[4] Because dihedral angle values are circular and 0° is the same as 360°, the edges of the Ramachandran plot "wrap" right-to-left and bottom-to-top. For instance, the small strip of allowed values along the lower-left edge of the plot are a continuation of the large, extended-chain region at upper left.

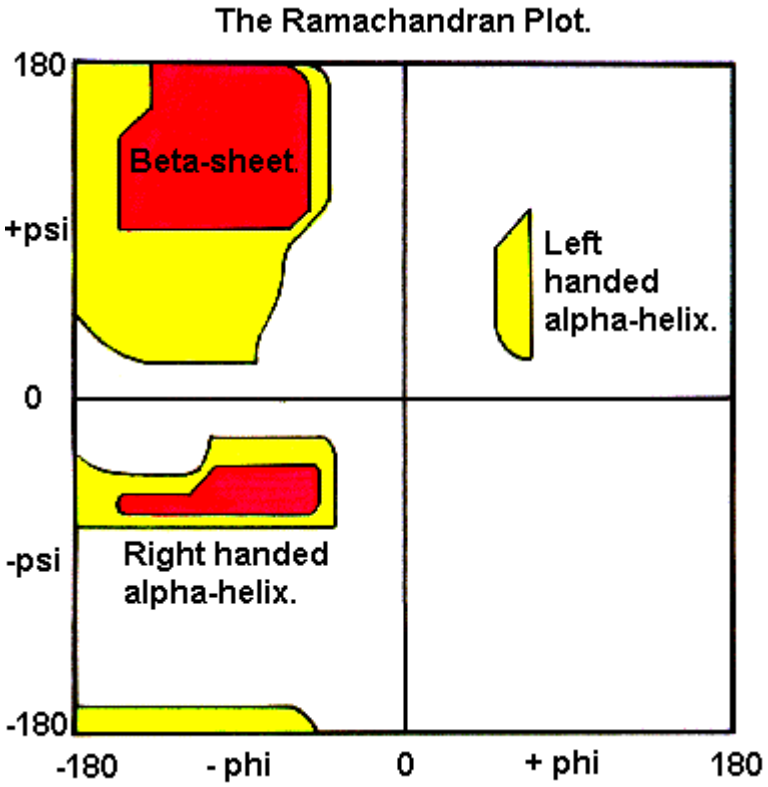


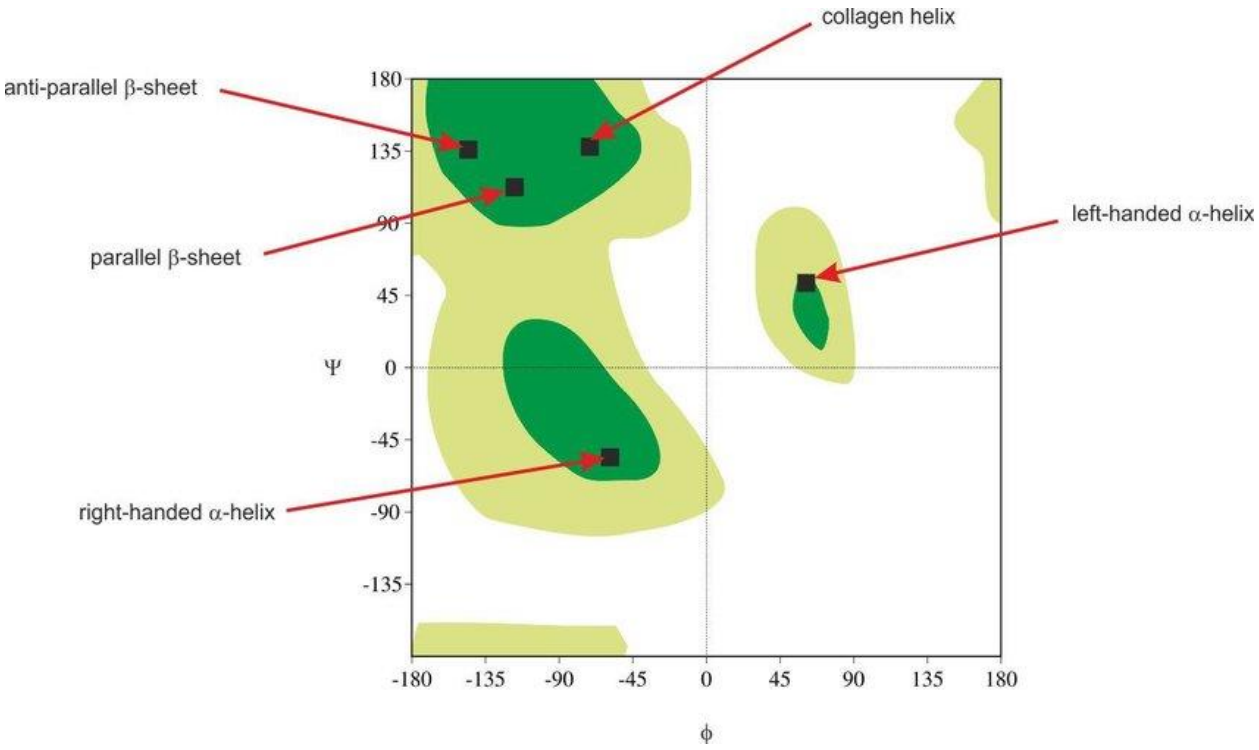
Uses

A Ramachandran plot can be used in two somewhat different ways. One is to show in theory which values, or conformations, of the ψ and φ angles are possible for an amino-acid residue in a protein (as at top right). A second is to show the empirical distribution of datapoints observed in a single structure (as at right, here) in usage for structure validation, or else in a database of many structures (as in the lower 3 plots at left). Either case is usually shown against outlines for the theoretically favored regions

Aminoacid Preferences

One might expect that larger side chains would result in more restrictions and consequently a smaller allowable region in the Ramachandran plot, but the effect of side chains is small.[5] In practice, the major effect seen is that of the presence or absence of the methylene group at Cβ.[5] Glycine has only a hydrogen atom for its side chain, with a much smaller van der Waals radius than the CH_3, CH_2, or CH group that starts the side chain of all other amino acids. Hence it is least restricted, and this is apparent in the Ramachandran plot for glycine (see Gly plot in gallery) for which the allowable area is considerably larger. In contrast, the Ramachandran plot for proline, with its 5-membered-ring side chain connecting Cα to backbone N, shows a limited number of possible combinations of ψ and φ (see Pro plot in gallery). The residue preceding proline ("pre-proline") also has limited combinations compared to the general case.





Protein Databank (PDB):

- PDB is a primary protein structure database. It is a crystallographic database for the three-dimensional structure of large biological molecules, such as proteins.
- In spite of the name, PDB archive the three-dimensional structures of not only proteins but also all biologically important molecules, such as nucleic acid fragments, RNA molecules, large peptides such as antibiotic gramicidin and complexes of protein and nucleic acids.
- The database holds data derived from mainly three sources: Structure determined by X-ray crystallography, NMR experiments, and molecular modeling.

Content

134,146 structures in the PDB have a structure factor file.
10,289 structures have an NMR restraint file.
4,814 structures in the PDB have a chemical shifts file.
4,718 structures in the PDB have a 3DEM map file deposited in EM Data Bank

Most structures are determined by X-ray diffraction, but about 10% of structures are determined by protein NMR. When using X-ray diffraction, approximations of the coordinates of the atoms of the protein are obtained, whereas using NMR, the distance between pairs of atoms of the protein is estimated. The final conformation of the protein is obtained from NMR by solving a distance geometry problem. After 2013, a growing number of proteins are determined by cryo-electron microscopy. Clicking on the numbers in the linked external table displays examples of structures determined by that method.

For PDB structures determined by X-ray diffraction that have a structure factor file, their electron density map may be viewed. The data of such structures is stored on the "electron density server".[17][18]

Historically, the number of structures in the PDB has grown at an approximately exponential rate, with 100 registered structures in 1982, 1,000 structures in 1993, 10,000 in 1999, and 100,000 in 2014

PDB file format

The **Protein Data Bank (pdb) file format** is a textual file format describing the three-dimensional structures of molecules held in the Protein Data Bank. The pdb format accordingly provides for description and annotation of protein and nucleic acid structures including atomic coordinates, secondary structure assignments, as well as atomic connectivity. In addition experimental metadata are stored. PDB format is the legacy file format for the Protein Data Bank which now keeps data on biological macromolecules in the newer mmCIF file format.

A typical PDB file describing a protein consists of hundreds to thousands of lines like the following (taken from a file describing the structure of a synthetic collagen-like peptide):

```
HEADER    EXTRACELLULAR MATRIX                    22-JAN-98   1A3I
TITLE     X-RAY CRYSTALLOGRAPHIC DETERMINATION OF A COLLAGEN-LIKE
TITLE    2 PEPTIDE WITH THE REPEATING SEQUENCE (PRO-PRO-GLY)
...
EXPDTA    X-RAY DIFFRACTION
AUTHOR    R.Z.KRAMER,L.VITAGLIANO,J.BELLA,R.BERISIO,L.MAZZARELLA,
AUTHOR   2 B.BRODSKY,A.ZAGARI,H.M.BERMAN
...
REMARK 350 BIOMOLECULE: 1
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
...
SEQRES   1 A    9  PRO PRO GLY PRO PRO GLY PRO PRO GLY
SEQRES   1 B    6  PRO PRO GLY PRO PRO GLY
SEQRES   1 C    6  PRO PRO GLY PRO PRO GLY
...
ATOM      1  N   PRO A   1       8.316  21.206  21.530  1.00 17.44           N
ATOM      2  CA  PRO A   1       7.608  20.729  20.336  1.00 17.44           C
ATOM      3  C   PRO A   1       8.487  20.707  19.092  1.00 17.44           C
ATOM      4  O   PRO A   1       9.466  21.457  19.005  1.00 17.44           O
ATOM      5  CB  PRO A   1       6.460  21.723  20.211  1.00 22.26           C
...
HETATM  130  C   ACY   401       3.682  22.541  11.236  1.00 21.19           C
HETATM  131  O   ACY   401       2.807  23.097  10.553  1.00 21.19           O
HETATM  132  OXT ACY   401       4.306  23.101  12.291  1.00 21.19           O
...
```

HEADER, TITLE and AUTHOR records

provide information about the researchers who defined the structure; numerous other types of records are available to provide other types of information.

REMARK records

can contain free-form annotation, but they also accommodate standardized information; for example, the REMARK 350 BIOMT records describe how to compute the coordinates of the experimentally observed multimer from those of the explicitly specified ones of a single repeating unit.

SEQRES records

give the sequences of the three peptide chains (named A, B and C), which are very short in this example but usually span multiple lines.

ATOM records

describe the coordinates of the atoms that are part of the protein. For example, the first ATOM line above describes the alpha-N atom of the first residue of peptide chain A, which is a proline residue; the first three floating point numbers are its x, y and z coordinates and are in units of Ångströms.[3] The next three columns are the occupancy, temperature factor, and the element name, respectively.

HETATM records

describe coordinates of hetero-atoms, that is those atoms which are not part of the protein molecule.

Drug Discovery
An introduction to the drug discovery process

The drug discovery process underpins the entire pharmaceutical industry, encompassing the early stages of research from target discovery and validation, right through to the identification of a drug candidate or lead compound. Initial identification of small therapeutic candidates comes about via a variety of streams. Research can lead to new insights into disease processes that highlight novel pathways for which drugs can be developed to intervene. Alternatively, companies conduct large scale trial and error based programs in order to identify molecular compounds that may be of interest. This is the process most often performed during initial lead discovery, with a view to take novel compounds right the way through to preclinical and clinical trials. Thoroughly calculated risk analysis at this point can increase the chances of success when investments into a lead are made.

Step 1 – Target identification and Validation
Target identification and validation kicks off the whole drug discovery process. Naturally occurring cellular or modular structures that appear to play an important role in pathogenicity or disease progression are normally targets for therapeutics. A good target needs to be efficacious, safe and be accessible by the drug molecule/meet clinical needs of the prospective patient.

Following identification of the drug target, a systematic validation approach should be adhered to for the mode of action of lead candidate to be assessed for efficacy. The approach itself depends on the therapeutic area, but has a set of general principles that include disease association, preclinical evidence in key cells, preclinical evidence in intact systems (i.e. transgenic animals), and literature survey and competitor information.

Step 2 – Hit identification and Validation
The obvious next step is to identify whether the small molecule leads have the desired effect against the identified targets. There are a number of approaches by which hits can be identified, including high-throughput screening, knowledge-based approaches, and virtual screening. Validation of hits is required following initial screening, and again there are a few options to choose from.

Step 3 – Moving from a hit to a lead
After a number of hit series have been established, the aim at this point is the refinement of each hit series in order to produce more selective compounds. Multiple series should be worked on in tandem, as it is likely that some hit series will fail, often due to particular characteristics of the series. Focusing on multiple structurally different sets of hit series will help to offset this possibility.

Step 4 – Lead Optimization
At this stage, the aim is to maintain the desired properties of lead compounds while improving on possible deficiencies of their structures, with a view to produce a preclinical drug candidate. This stage can be used to find out whether your drug metabolizes in the right area of the body, or whether there are currently any side effects that are cause for concern. For this process, an integrated approach is recommended. The combination of specialists in computational chemistry, medical chemistry, drug metabolism, and other areas can provide unique insights into this late stage of the process.

Step 5 – Late Lead Optimization
Before progression to preclinical and clinical trials, late stage optimization, in which further pharmacological safety of a lead compound is assessed, is a vital step. If this stage is overlooked, problems in efficacy, pharmacokinetics, and safety are more likely to occur later on in drug development. Safety optimization is a core stage, the aims are to identify and progress the leads with the best overall safety profile, remove the most toxic leads, and establish a well-characterized hazard and translational risk profile to enable further *in vitro* tests.

Drug development
Drug development is the process of bringing a new pharmaceutical drug to the market once a lead compound has been identified through the process of drug discovery. It includes preclinical research on microorganisms and animals, filing for regulatory status, such as via the United States Food and Drug Administration for an investigational new drug to initiate clinical trials on humans, and may include the step of obtaining regulatory approval with a new drug application to market the drug. The entire process – from concept through preclinical testing in the laboratory to clinical trial development, including Phase I–III trials – to approved vaccine or drug typically takes more than a decade
Broadly, the process of drug development can be divided into preclinical and clinical work.
Step 1: Drug discovery and target validation
The first step in the drug development process involves discovery work. This is where drug development companies choose a molecule, such as a gene or protein, to target with a drug. This is also where the drug developer will confirm that the molecule is indeed involved with the disease in question. After testing multiple drug molecules, the drug development company will choose those that have promise. Keep in mind it's not uncommon for drug developers to have more than a handful of promising lead compounds at this stage.

Step 2: Preclinical testing
The next step in the drug development process is preclinical testing, which in itself is divided into two subcomponents: *in vitro* and *in vivo* testing. *In vitro* testing examines the drug molecules' interactions in test tubes and within the lab setting. *In vivo* testing involves testing the drug molecules on animal models and in other living cell cultures. Although efficacy is beginning to be established here, safety is paramount as the FDA will not let preclinical studies move into human trials without extensive data on safety. This is the stage where researchers will whittle thousands of drug molecule candidates down to between one and five. By the time preclinical testing has concluded, many years have often passed.

Step 3: Investigational New Drug application filing
The third step involves submitting an Investigational New Drug application to the FDA prior to beginning human clinical trials. This is the point where the FDA will scrutinize the results from the preclinical testing, look at side effects and other safety features of an experimental drug, examine the drugs' chemical structure and how it's believed to work, and take its first look at the manufacturing process of the drug. If the FDA approves a drug developers' IND then it can move onto human trials. An IND approval is also the point at which a patented drugs' 20-year exclusivity period begins.

Step 4: Phase 1 clinical studies
The first phase of human clinical testing involves a relatively small group of healthy people, usually a dozen to a few dozen, and it'll focus entirely on safety. This stage of study involves looking at how a drug is absorbed and eliminated from the body, as well as what side effects

it may cause and whether or not it's producing the desired effect. Phase 1 clinical studies are also where maximum tolerated doses are established. It really is all about safety, although it's not uncommon for drug developers to tout early signs of efficacy in phase 1. If everything looks promising the study moves to phase 2, or midstage trials.

Step 5: Phase 2 clinical studies
The two big changes between early stage and mid-stage trials are that the patient pool widens from a few dozen to perhaps 100 or more patients, and the patients being treated are no longer healthy volunteers but people being afflicted by the disease in question. Safety remains a big focus of phase 2 studies, with short-term side effects being closely monitored, although an increasing emphasis will begin to be placed on whether or not a drug is working as expected and if it's improving the condition or not. Phase 2 studies also establish which dose (if multiple doses were tested, as is often the case) performed most optimally. If the experimental drug continues to look promising it'll move onto late-stage studies.

Step 6: Phase 3 clinical studies
In phase 3 studies, safety remains a priority, but this is where efficacy also plays a big role. Phase 3 studies are designed by drug developers but approved by the FDA with guidelines for a clearly defined primary endpoint to determine the success or failure of a tested drug. Phase 3 trials involve even more patients, perhaps a few hundred to maybe thousands, and they are by far the longest and costliest of all components of the drug development process. This is also the stage where drug developers will begin to think about how they're going to ramp up production if the phase 3 results are promising. Assuming an experimental drug meets its primary endpoint and is demonstrated to be safe, the next step is to file for its approval.

Step 7: New Drug Application filing
The seventh step in the drug development process is simple: filing a New Drug Application with the FDA. Unfortunately this isn't just a single page that says "please look at our drug!" An NDA can be tens of thousands or perhaps 100,000 or more pages long, and it contains all research and safety data examined during each of the six prior steps. Still, this stage isn't the point where the FDA has to make a decision to approve or deny the drug; it's merely a stepping stone that says it promises to review the application over the next 10 months. If the NDA is accepted a PDUFA, or Prescription Drug User Fee Act, date is set 10 months down the road (for a standard application) whereby the FDA is expected to make its decision. Keep in mind the FDA can postpone this decision or even rule early should it choose.

Step 8: PDUFA date and decision
More often than not, the FDA will wait until the PDUFA date to release its decision. Essentially the FDA has three choices: it can approve a drug; it can outright deny a drug (which is pretty rare from what I've witnessed in 15 years), or it can request additional information by sending a complete response letter, or CRL. A CRL simply states what was lacking that prevented the drug from being approved and offers suggestions as to how to remedy the situation. Often times it requires drug developers to run additional studies or perhaps alter their manufacturing process to appease the FDA. If approved by the FDA, the drug becomes immediately available for commercial production.

Step 9: Phase 4 clinical studies
"Technically" an approved drug can make it to your medicine cabinet after step eight, but that doesn't mean the drug developer is off the hook yet. Even after approval, it's not uncommon for the FDA to request long-term safety studies be undertaken whereby drug developers are required to submit regular reports detailing any adverse events with the drug to the FDA. Even following approval, safety remains the top priority of the FDA.

From start to finish, the entire drug development process (steps 1 through 8) usually spans about 10 to 15 years, leaving drug developers with around a decade or less of patent exclusivity on branded drugs once they make it to market. This should help provide some insight into why prescription drug prices are so high, why drug companies may seem like they're taking "forever" in developing the next cure for a terrible disease, and why so few drugs actually earn a spot in your medicine cabinet.

Drug design

Drug design, sometimes referred to as **rational drug design** or more simply rational design, is the inventive process of finding new medications based on the knowledge of a biological target.[1] The drug is most commonly an organic small molecule that activates or inhibits the function of a biomolecule such as a protein, which in turn results in a therapeutic benefit to the patient. In the most basic sense, drug design involves the design of small molecules that are complementary in shape and charge to the biomolecular target with which they interact and therefore will bind to it. Drug design frequently but not necessarily relies on computer modeling techniques.[2] This type of modeling is often referred to as **computer-aided drug design**. Finally, drug design that relies on the knowledge of the three-dimensional structure of the biomolecular target is known as **structure-based drug design**.

The phrase “drug design” is to some extent a misnomer. What is really meant by drug design is ligand design (i.e., design of a small molecule that will bind tightly to its target).[3] Although modeling techniques for prediction of binding affinity are reasonably successful, there are many other properties, such as bioavailability, metabolic half-life, lack of side effects, etc., that first must be optimized before a ligand can become a safe and efficacious drug. These other characteristics are often difficult to optimize using rational drug design techniques.

Background

Typically a drug target is a key molecule involved in a particular metabolic or signaling pathway that is specific to a disease condition or pathology or to the infectivity or survival of a microbial pathogen. Some approaches attempt to inhibit the functioning of the pathway in the diseased state by causing a key molecule to stop functioning. Drugs may be designed that bind to the active region and inhibit this key molecule. Another approach may be to enhance the normal pathway by promoting specific molecules in the normal pathways that may have been affected in the diseased state. In addition, these drugs should also be designed so as not to affect any other important “off-target” molecules or antitargets that may be similar in appearance to the target molecule, since drug interactions with off-target molecules may lead to undesirable side effects. Sequence homology is often used to identify such risks.

Most commonly, drugs are organic small molecules produced through chemical synthesis, but biopolymer-based drugs (also known as biologics) produced through biological processes are becoming increasingly more common. In addition, mRNA-based gene silencing technologies may have therapeutic applications.

Types

There are two major types of drug design. The first is referred to as **ligand-based drug design** and the second, **structure-based drug design**.

Ligand-based

Ligand-based drug design (or **indirect drug design**) relies on knowledge of other molecules that bind to the biological target of interest. These other molecules may be used to derive a pharmacophore model that defines the minimum necessary structural characteristics a molecule must possess in order to bind to the target.[4] In other words, a model of the biological target may be built based on the knowledge of what binds to it, and this model in turn may be used to design new molecular entities that interact with the target. Alternatively, a quantitative structure-activity relationship (QSAR), in which a correlation between calculated properties of molecules and their experimentally determined biological activity, may be derived. These QSAR relationships in turn may be used to predict the activity of new analogs.

Structure-based

Structure-based drug design (or **direct drug design**) relies on knowledge of the three dimensional structure of the biological target obtained through methods such as x-ray crystallography or NMR spectroscopy.[5] If an experimental structure of a target is not available, it may be possible to create a homology model of the target based on the experimental structure of a related protein. Using the structure of the biological target, candidate drugs that are predicted to bind with high affinity and selectivity to the target may be designed using interactive graphics and the intuition of a medicinal chemist. Alternatively various automated computational procedures may be used to suggest new drug candidates.

As **experimental methods** such as X-ray crystallography and NMR develop, the amount of information concerning 3D structures of biomolecular targets has increased dramatically. In parallel, information about the structural dynamics and electronic properties about ligands has also increased. This has encouraged the rapid development of the structure-based drug design. Current methods for structure-based drug design can be divided roughly into two categories. The first category is about "finding" ligands for a given receptor, which is usually referred as database searching. In this case, a large number of potential ligand molecules are screened to find those fitting the binding pocket of the receptor. This method is usually referred as ligand-based drug design. The key advantage of database searching is that it saves synthetic effort to obtain new lead compounds. Another category of structure-based drug design methods is about "building" ligands, which is usually referred as receptor-based drug design. In this case, ligand molecules are built up within the constraints of the binding pocket by assembling small pieces in a stepwise manner. These pieces can be either individual atoms or molecular fragments. The key advantage of such a method is that novel structures, not contained in any database, can be suggested.[6][7][8]

Active site identification

Active site identification is the first step in this program. It analyzes the protein to find the binding pocket, derives key interaction sites within the binding pocket, and then prepares the necessary data for Ligand fragment link. The basic inputs for this step are the 3D structure of the protein and a pre-docked ligand in PDB format, as well as their atomic properties. Both ligand and protein atoms need to be classified and their atomic properties should be defined, basically, into four atomic types:

- **hydrophobic atom**: All carbons in hydrocarbon chains or in aromatic groups.
- **H-bond donor**: Oxygen and nitrogen atoms bonded to hydrogen atom(s).
- **H-bond acceptor**: Oxygen and sp2 or sp hybridized nitrogen atoms with lone electron pair(s).

- **Polar atom**: Oxygen and nitrogen atoms that are neither H-bond donor nor H-bond acceptor, sulfur, phosphorus, halogen, metal, and carbon atoms bonded to hetero-atom(s).

The space inside the ligand binding region would be studied with virtual probe atoms of the four types above so the chemical environment of all spots in the ligand binding region can be known. Hence we are clear what kind of chemical fragments can be put into their corresponding spots in the ligand binding region of the receptor.

Ligand fragment link

When we want to plant "seeds" into different regions defined by the previous section, we need a fragments database to choose fragments from. The term "fragment" is used here to describe the building blocks used in the construction process. The rationale of this algorithm lies in the fact that organic structures can be decomposed into basic chemical fragments. Although the diversity of organic structures is infinite, the number of basic fragments is rather limited.

Before the first fragment, i.e. the seed, is put into the binding pocket, and other fragments can be added one by one, it is useful to identify potential problems. First, the possibility for the fragment combinations is huge. A small perturbation of the previous fragment conformation would cause great difference in the following construction process. At the same time, in order to find the lowest binding energy on the Potential energy surface (PES) between planted fragments and receptor pocket, the scoring function calculation would be done for every step of conformation change of the fragments derived from every type of possible fragments combination. Since this requires a large amount of computation, one may think using other possible strategies to let the program works more efficiently. When a ligand is inserted into the pocket site of a receptor, conformation favor for these groups on the ligand that can bind tightly with receptor should be taken priority. Therefore it allows us to put several seeds at the same time into the regions that have significant interactions with the seeds and adjust their favorite conformation first, and then connect those seeds into a continuous ligand in a manner that make the rest part of the ligand having the lowest energy. The conformations of the pre-placed seeds ensuring the binding affinity decide the manner that ligand would be grown. This strategy reduces calculation burden for the fragment construction efficiently. On the other hand, it reduces the possibility of the combination of fragments, which reduces the number of possible ligands that can be derived from the program. These two strategies above are well used in most structure-based drug design programs. They are described as "**Grow**" and "**Link**". The two strategies are always combined in order to make the construction result more reliable.[6][7][9]

Scoring method

Structure-based drug design attempts to use the structure of proteins as a basis for designing new ligands by applying accepted principles of molecular recognition. The basic assumption underlying structure-based drug design is that a good ligand molecule should bind tightly to its target. Thus, one of the most important principles for designing or obtaining potential new ligands is to predict the binding affinity of a certain ligand to its target and use it as a criterion for selection.

One early method was developed by Böhm to develop a general-purposed empirical scoring function in order to describe the binding energy. The following "Master Equation" was derived:

$$\Delta G_{\mathrm{bind}} = -RT \ln K_{\mathrm{d}}$$

$$K_{\mathrm{d}} = \frac{[\mathrm{Receptor}][\mathrm{Acceptor}]}{[\mathrm{Complex}]}$$

$$\Delta G_{\mathrm{bind}} = \Delta G_{\mathrm{desolvation}} + \Delta G_{\mathrm{motion}} + \Delta G_{\mathrm{configuration}} + \Delta G_{\mathrm{interaction}}$$

where:

- desolvation – enthalpic penalty for removing the ligand from solvent
- motion – entropic penalty for reducing the degrees of freedom when a ligand binds to its receptor
- configuration – conformational strain energy required to put the ligand in its "active" conformation
- interaction – enthalpic gain for "resolvating" the ligand with its receptor

The basic idea is that the overall binding free energy can be decomposed into independent components that are known to be important for the binding process. Each component reflects a certain kind of free energy alteration during the binding process between a ligand and its target receptor. The Master Equation is the linear combination of these components. According to Gibbs free energy equation, the relation between dissociation equilibrium constant, K_d, and the components of free energy was built.

Rational drug discovery

In contrast to traditional methods of drug discovery, which rely on trial-and-error testing of chemical substances on cultured cells or animals, and matching the apparent effects to treatments, rational drug design begins with a hypothesis that modulation of a specific biological target may have therapeutic value. In order for a biomolecule to be selected as a drug target, two essential pieces of information are required. The first is evidence that modulation of the target will have therapeutic value. This knowledge may come from, for example, disease linkage studies that show an association between mutations in the biological target and certain disease states. The second is that the target is "drugable". This means that it is capable of binding to a small molecule and that its activity can be modulated by the small molecule.

Once a suitable target has been identified, the target is normally cloned and expressed. The expressed target is then used to establish a screening assay. In addition, the three-dimensional structure of the target may be determined.

The search for small molecules that bind to the target is begun by screening libraries of potential drug compounds. This may be done by using the screening assay (a "wet screen"). In addition, if the structure of the target is available, a virtual screen may be performed of candidate drugs. Ideally the candidate drug compounds should be "drug-like", that is they should possess properties that are predicted to lead to oral bioavailability, adequate chemical and metabolic stability, and minimal toxic effects. Several methods are available to estimate druglikeness such as Lipinski's Rule of Five and a range of scoring methods such as Lipophilic efficiency. Several methods for predicting drug metabolism have been proposed in the scientific literature, and a recent example is SPORCalc. Due to the complexity of the drug design process, two terms of interest are still serendipity and bounded rationality. Those challenges are caused by the large chemical space describing potential new drugs without side-effects.

Computer-aided drug design

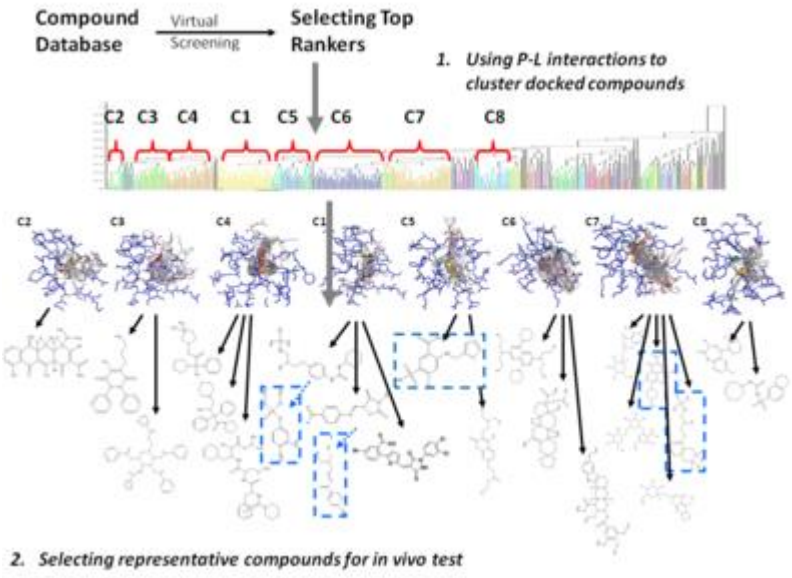
Computer-aided drug design uses computational chemistry to discover, enhance, or study drugs and related biologically active molecules. The most fundamental goal is to predict whether a given molecule will bind to a target and if so how strongly. Molecular mechanics or molecular dynamics are most often used to predict the conformation of the small molecule and to model conformational changes in the biological target that may occur when the small molecule binds to it. Semi-empirical, ab initio quantum chemistry methods, or density functional theory are often used to provide optimized parameters for the molecular mechanics calculations and also provide an estimate of the electronic properties (electrostatic potential, polarizability, etc.) of the drug candidate that will influence binding affinity.

Molecular mechanics methods may also be used to provide semi-quantitative prediction of the binding affinity. Also, knowledge-based scoring function may be used to provide binding affinity estimates. These methods use linear regression, machine learning, neural nets or other statistical techniques to derive predictive binding affinity equations by fitting experimental affinities to computationally derived interaction energies between the small molecule and the target.

Ideally the computational method should be able to predict affinity before a compound is synthesized and hence in theory only one compound needs to be synthesized. The reality however is that present computational methods are imperfect and provide at best only qualitatively accurate estimates of affinity. Therefore in practice it still takes several iterations of design, synthesis, and testing before an optimal molecule is discovered. On the other hand, computational methods have accelerated discovery by reducing the number of iterations required and in addition have often provided more novel small molecule structures.

Drug design with the help of computers may be used at any of the following stages of drug discovery:

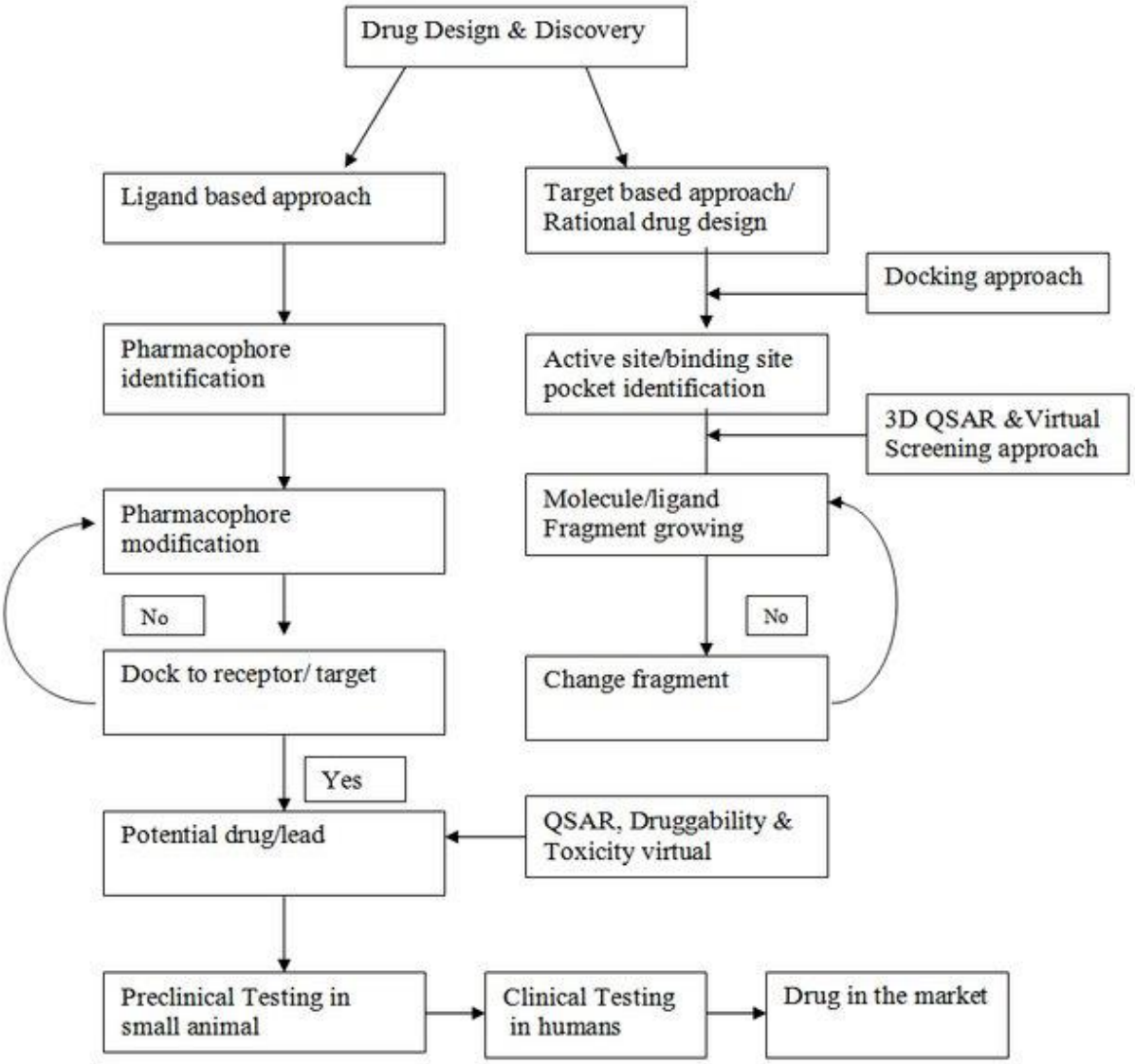
1. hit identification using virtual screening (structure- or ligand-based design)
2. hit-to-lead optimization of affinity and selectivity (structure-based design, QSAR, etc.)
3. lead optimization optimization of other pharmaceutical properties while maintaining affinity



Flowchart of a Usual Clustering Analysis for Structure-Based Drug Design

In order to overcome the insufficient prediction of binding affinity calculated by recent scoring functions, the protein-ligand interaction and compound 3D structure information are used to analysis. For structure-based drug design, several post-screening analysis focusing on protein-ligand interaction has been developed for improving enrichment and effectively mining potential candidates:

- Consensus scoring
  - Selecting candidates by voting of multiple scoring functions
  - May lose the relationship between protein-ligand structural information and scoring criterion
- Geometric analysis
  - Comparing protein-ligand interactions by visually inspecting individual structures
  - Becoming intractable when the number of complexes to be analyzed increasing
- Cluster analysis
  - Represent and cluster candidates according to protein-ligand 3D information
  - Needs meaningful representation of protein-ligand interactions.



Molecular modeling in drug discovery

The term molecular modelling expanded over the last decade from the tools to visulalize three dimensional structures and to simulate, predict and analyse the properties and the behaviour of the molecules on an atomic level to data mining and platform to organize many compounds and their properties into database and to perform virtual drug screening via 3D database screening for novel drug compounds.

Molecular modelling allow the scientist to use computers to visualize molecules means representing molecular structures numerically and simulating their behavior with the equations of quantum and classical physics to discover new lead compounds for drugs or to refine existing drugs Insilco.

Goal:

To develop a sufficient accurate model of the system so that physical experiment may not

be necessary

The definition currently accepted of what molecular modeling is, can be stated as this: "Molecular modeling is anything that requires the use of a computer to paint, describe or evaluate any aspect of the properties of the structure of a molecule" (Pensak, 1989). Methods used in the molecular modeling arena regard automatic structure generation, analysis of three-dimensional (3D) databases, construction of protein models by techniques based on sequence homology, diversity analysis, docking of ligands or continuum methods.
Thus, today molecular modelling is regarded as a field concerned with the use of all sort of different strategies to model and to deduce information of a system at the atomic level. On the other hand, this discipline includes all methodologies used in computational chemistry, like computation of the energy of a molecular system, energy minimization, Monte Carlo methods or molecular dynamics. In other words, it is possible to conclude that computational chemistry is the nucleus of molecular modeling.

Applications

Molecular modelling methods are now routinely used to investigate the structure, dynamics, surface properties and thermodynamics of inorganic, biological and polymeric systems.
The types of biological activity that have been investigated using molecular modelling include protein folding, enzyme catalysis, protein stability, conformational changes associated with biomolecular function, and molecular recognition of proteins, DNA, and membrane complexes.

Why models are used?

a) to help with analysis and interpretation of experimental data
b) to uncover new laws and formulate new theories
c) to help solve problems and hint solutions before doing experiments
d) to help design new experiments
e) to predict properties and quantities that are difficult or even impossible to observe experimentally
Simulations and computer "experiments" can be designed to mimic reality, however, are always based on assumptions, approximations and simplifications (i.e. models).

Important characteristics of models are:

a) Level of simplification: very simple to very complex
b) Generality: general or specific, i.e. relate only to specific systems or problems

c) Limitations: one must always be aware of the range of applicability and limits of accuracy of any model.
d) Cost and efficiency: CPU time, memory, disk space

Computable quantities:

a) molecular structures: closely tied to energy (best structure - one for which the energy is minimum)
b) energy: potential energy surfaces (PES) - extremely important! PES dictate essentially everything about the molecule or system
c) molecular properties that can be compared to/used to interpret experiments: thermodynamics, kinetics, spectra (IR, UV, NMR)
d) properties that are not experimental observables: bond order, aromaticity, molecular orbitals

Cheminformatics

With the increase of computational power, machine learning has found many applications in different fields of science. One of them is chemistry, where scientists apply machine learning models to predict various molecule's properties such as its solubility and toxicity [1] or use it for drug discovery.

Cheminformatics (also known as **chemoinformatics**) refers to use of physical chemistry theory with computer and information science techniques—so called "*in silico*" techniques—in application to a range of descriptive and prescriptive problems in the field of chemistry, including in its applications to biology and related molecular fields. Such *in silico* techniques are used, for example, by pharmaceutical companies and in academic settings to aid and inform the process of drug discovery, for instance in the design of well-defined combinatorial libraries of synthetic compounds, or to assist in structure-based drug design. The methods can also be used in chemical and allied industries, and such fields as environmental science and pharmacology, where chemical processes are involved or studied.

It is also referred as Chemoinformatics/Chemiinformatics/Chemical information/Chemical informatics has been recognised in recent years as a distinct discipline in computational molecular sciences. Cheminformatics is also known as interface science as it combines Physics, Chemistry, Biology, Mathematics, Biochemistry, Statistics and informatics.

The primary focus of cheminformatics is to analyse/simulate/modelling/manipulate chemical information which can be represented either in 2D structure or in 3D structure. Industry sectors such as, agrochemicals, food and pharmaceutical are distinct areas where cheminformatics plays significant role in the recent history of molecular sciences.

Cheminformatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization, and use of chemical information [3]. According to F.K Brown "The use of information technology and management has become a critical part of the drug discovery process. Cheminformatics is the mixing of information resources to transform data into information and information into knowledge which is collectively referred as inductive learning as shown in Fig.1. for the intended purpose of making better decisions faster in the areas of drug lead identification and organization" [1]. From J. Gasteiger and T. Engel perception, cheminformatics can be viewed as "The application of informatics methods to solve chemical problems" [2]. M. Hann and R. Green coined cheminformatics as "a new name for an old problem" [5].

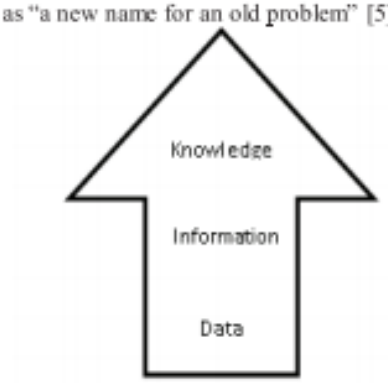


Fig. 1. Cheminformatics transformation

Cheminformatics has mainly dealt with small molecules, whereas bioinformatics addresses genes, proteins, and other larger chemical compounds (shown in Fig. 2). Chem and Bioinformatics complements each other for bimolecular process, like structure and function of proteins, the binding of a ligand to its binding site, the conversion of a substrate within its enzyme receptor, and the catalysis of a biochemical reaction by an enzyme.

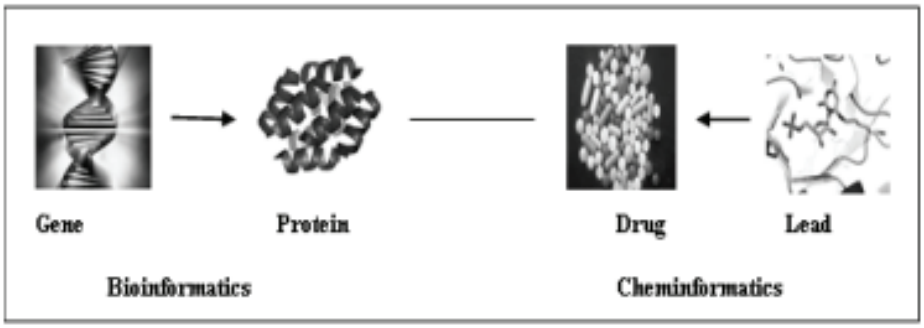


Fig. 2. The Cooperation of Bioinformatics and Cheminformatics

Different tools and methods are available to represent chemical structure, database to store chemical data, to perform searching process, Quality Structure- Activity Relationship(QSAR), Quality Structure-

Need and importance of Chemnformatics

Cheminformatics plays a key role to maintain and access enormous amount of chemical data, produced by chemist (more than 45 million chemical compounds are known and the number may increase in million every year,) by using a proper database. Also, the field of chemistry needs a novel technique for knowledge extraction from data to model complex relationships between the structure of the chemical compound and biological activity or the influence of reaction condition on chemical reactivity [6][16]. Cheminformatics has wider range of application and Fig. 3. shows influence if cheminformatics in some specific research areas.

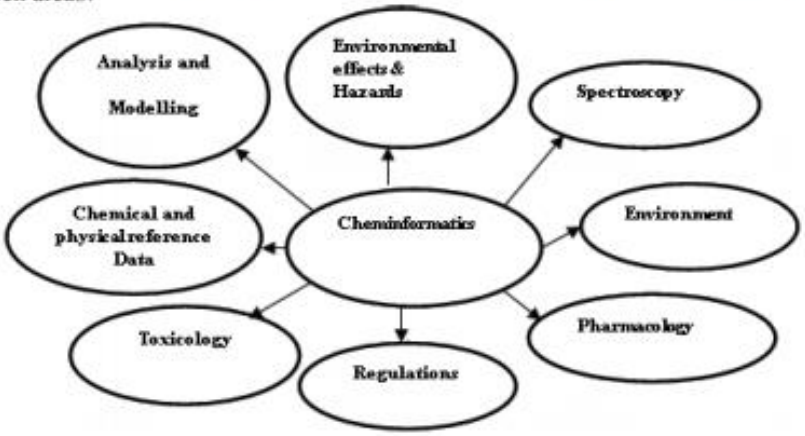


Fig. 3. Need for Cheminformatics

Three major aspects of Cheminformatics are;

i) Information Acquisition, is a process of generating and collecting data empirically (experimentation) or from theory (molecular simulation)
ii) Information Management deals with storage and retrieval of information and
iii) Information use, which includes Data Analysis, correlation, and application to problems in the chemical and biochemical sciences [24]

Scope of chemo-informatics

- In Drug designing
  - Provide virtual structures libraries
  - Provide Graphical interphase for drug designing
  - Docking
  - Drug discovery
  - Drug development
  - QSAR studies
- In Clinical research
  - Interaction studies
  - Provide molecular visualization
  - Provide predicted score
  - Structural analysis
- In Synthetic chemistry
  - chemical synthesis
  - designing new molecules
  - Similarity searching
- In Pharma-industries
  - Activity prediction
  - Geno-toxicity prediction
  - Mutagenicity prediction
  - Carcinogenicity prediction
  - Interaction sutidies
- In Pharmacogenomics

- Drug genome interaction studies
- Genome level drug developments
- Pharmacophore predictions(also used in drug designing)

- In systems biology
  - Target designing
  - Interaction studies
  - Interaction prediction of interaction values
- In nanotechnology
  - Nanoparticle libraries generation
  - Nanoparticle studies
  - Designing, prediction

Applications

- Storage and retrieval
  - The primary application of chemo- informatics is in the storage, indexing and search of information relating to compounds. include: Unstructured data Digital libraries Information retrieval Information extraction
- File formats
  - The in silico representation of chemical structures uses specialized formats such as the: ϖ SDF:2D & 3D ϖ Mol:2D ϖ Mol2:3D ϖ Xml
- Virtual libraries
  - Chemical data can pertain to real or virtual molecules Virtual libraries of classes of compounds: drugs, natural products, diversity-oriented synthetic products
- Virtual screening
  - virtual screening involves computationally screening in silico libraries of compounds virtual screening on the basis of: biological activity against a given target, Docking, similarity, etc.

- Quantitative structure-activity relationship (QSAR)
  - This is the calculation of quantitative structure-activity relationship and quantitative structure property relationship values, used to predict the activity of compounds from their structures Use descriptors

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING
DEPARTMENT OF BIOINFORMATICS

UNIT – 3- SBI1208 – Cheminformatics

MOLECULAR MODELING

The term molecular modelling expanded over the last decade from the tools to visulalize three dimensional structures and to simulate, predict and analyse the properties and the behaviour of the molecules on an atomic level to data mining and platform to organize many compounds and their properties into database and to perform virtual drug screening via 3D database screening for novel drug compounds.

Molecular modelling allow the scientist to use computers to visualize molecules means representing molecular structures numerically and simulating their behavior with the equations of quantum and classical physics to discover new lead compounds for drugs or to refine existing drugs Insilco.

Goal:

To develop a sufficient accurate model of the system so that physical experiment may not be necessary

The definition currently accepted of what molecular modeling is, can be stated as this: "Molecular modeling is anything that requires the use of a computer to paint, describe or evaluate any aspect of the properties of the structure of a molecule" (Pensak, 1989). Methods used in the molecular modeling arena regard automatic structure generation, analysis of three-dimensional (3D) databases, construction of protein models by techniques based on sequence homology, diversity analysis, docking of ligands or continuum methods.

Thus, today molecular modelling is regarded as a field concerned with the use of all sort of different strategies to model and to deduce information of a system at the atomic level. On the other hand, this discipline includes all methodologies used in computational chemistry, like computation of the energy of a molecular system, energy minimization, Monte Carlo methods or molecular dynamics. In other words, it is possible to conclude that computational chemistry is the nucleus of molecular modeling.

Applications

Molecular modelling methods are now routinely used to investigate the structure, dynamics, surface properties and thermodynamics of inorganic, biological and polymeric systems.

The types of biological activity that have been investigated using molecular modelling include protein folding, enzyme catalysis, protein stability, conformational changes associated with biomolecular function, and molecular recognition of proteins, DNA, and membrane complexes.

Why models are used?

- a) to help with analysis and interpretation of experimental data
- b) to uncover new laws and formulate new theories
- c) to help solve problems and hint solutions before doing experiments
- d) to help design new experiments

- e) to predict properties and quantities that are difficult or even impossible to observe experimentally

Simulations and computer "experiments" can be designed to mimic reality, however, are always based on assumptions, approximations and simplifications (i.e. models).

Important characteristics of models are:

- a) Level of simplification: very simple to very complex
- b) Generality: general or specific, i.e. relate only to specific systems or problems
- c) Limitations: one must always be aware of the range of applicability and limits of accuracy of any model.
- d) Cost and efficiency: CPU time, memory, disk space

Computable quantities:

- a) molecular structures: closely tied to energy (best structure - one for which the energy is minimum)
- b) energy: potential energy surfaces (PES) - extremely important! PES dictate essentially everything about the molecule or system
- c) molecular properties that can be compared to/used to interpret experiments: thermodynamics, kinetics, spectra (IR, UV, NMR)
- d) properties that are not experimental observables: bond order, aromaticity, molecular orbitals

Three stages of Molecular Modeling

1. Model is selected to describe the intra and inter mol. Interactions in the system
   – Two common models
   – Quantum mechanics
   – Molecular mechanics

These models enable the energy of any arrangement if the atoms and mol to be calculated and allow the modeller to determine how the energy of the system varies as the positions of the atoms and molecular changes

2. Calculation itself such as energy minimization, molecular dynamics or Monte carlo simulations or conformational search
3. Calculation must be analyzed not only to calculate properties but also to check that it has been performed properly

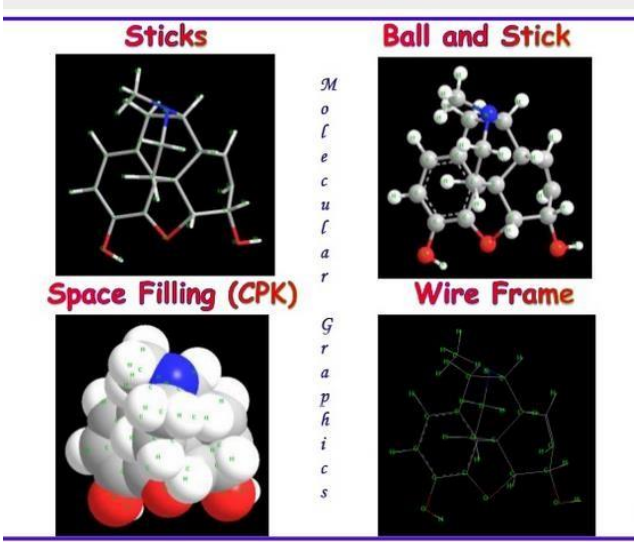
Molecular Visualisation

Once 3D coordinates are available, they can be visualised, an important aid to

interpretation of molecular modelling:

- **Wireframe**, **Ball and Stick** and **Spacefill** for small and medium sized molecules
- **Ribbon** for protein, nucleotide and carbohydrate structures to render the tertiary molecular structures, **Polyhedral modes** for eg ionic lattices.
- **Isosurfaces**, which are generated from the sizes of atoms, and onto which can be colour coded further properties such as MOs, charges etc.
- **Animation** to view molecular vibrations and the time dependent properties of molecules such as (intrinsic) reaction coordinates, protein folding dynamics, etc.

• **Integration and Scripting**. Programs such as Jmol or ChemDoodle allow seamless integration of models as part of lecture courses, electronic journals, podcasts, iPads, etc and increasingly elaborate scripting of the models to illustrate scientific points.



1.2 Coordinate Systems

It is obviously important to be able to specify the positions of the atoms and/or molecules in the system to a modelling program*. There are two common ways in which this can be done. The most straightforward approach is to specify the Cartesian (x, y, z) coordinates of all the atoms present. The alternative is to use *internal coordinates*, in which the position of each atom is described relative to other atoms in the system. Internal coordinates are usually written as a Z-matrix. The Z-matrix contains one line for each atom in the system. A sample Z-matrix for the staggered conformation of ethane (see Figure 1.1) is

*For a system containing a large number of independent molecules it is common to use the term 'configuration' to refer to each arrangement; this use of the word 'configuration' is not to be confused with its standard chemical meaning as a different bonding arrangement of the atoms in a molecule

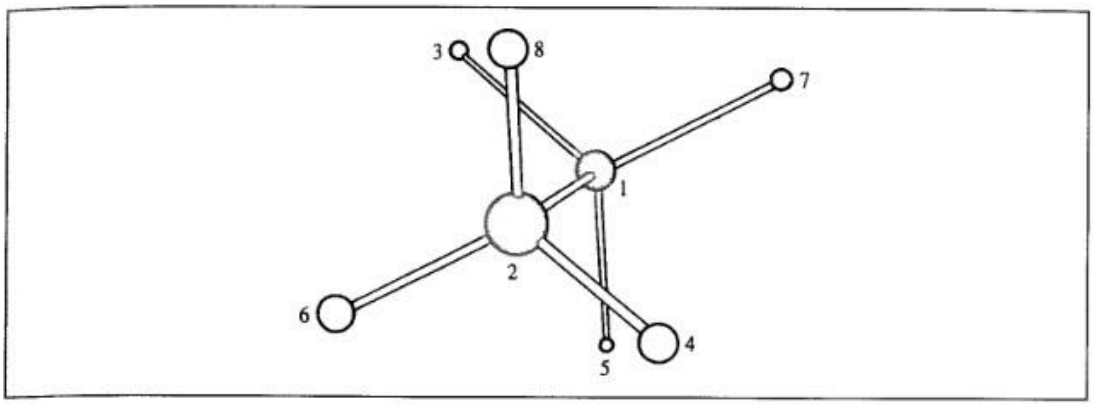


Fig. 1 1 The staggered conformation of ethane

as follows:

1	C						
2	C	1.54	1				
3	H	1.0	1	109.5	2		
4	H	1.0	2	109.5	1	180.0	3
5	H	1.0	1	109.5	2	60.0	4
6	H	1.0	2	109.5	1	−60.0	5
7	H	1.0	1	109.5	2	180.0	6
8	H	1.0	2	109.5	1	60.0	7

In the first line of the Z-matrix we define atom 1, which is a carbon atom. Atom number 2 is also a carbon atom that is a distance of 1.54 Å from atom 1 (columns 3 and 4). Atom 3 is a hydrogen atom that is bonded to atom 1 with a bond length of 1.0 Å. The angle formed by atoms 2–1–3 is 109.5°, information that is specified in columns 5 and 6. The fourth atom is a hydrogen, a distance of 1.0 Å from atom 2, the angle 4–2–1 is 109.5°, and the torsion angle (defined in Figure 1.2) for atoms 4–2–1–3 is 180°. Thus for all except the first three atoms, each atom has three internal coordinates: the distance of the atom from one of the atoms previously defined, the angle formed by the atom and two of the previous atoms, and the torsion angle defined by the atom and three of the previous atoms. Fewer internal coordinates are required for the first three atoms because the first atom can be placed anywhere in space (and so it has no internal coordinates); for the second atom it is only necessary to specify its distance from the first atom and then for the third atom only a distance and an angle are required.

It is always possible to convert internal to Cartesian coordinates and vice versa. However, one coordinate system is usually preferred for a given application. Internal coordinates can usefully describe the relationship between the atoms in a single molecule, but Cartesian coordinates may be more appropriate when describing a collection of discrete molecules. Internal coordinates are commonly used as input to quantum mechanics programs, whereas calculations using molecular mechanics are usually done in Cartesian coordinates. The total number of coordinates that must be specified in the internal coordinate system is six fewer

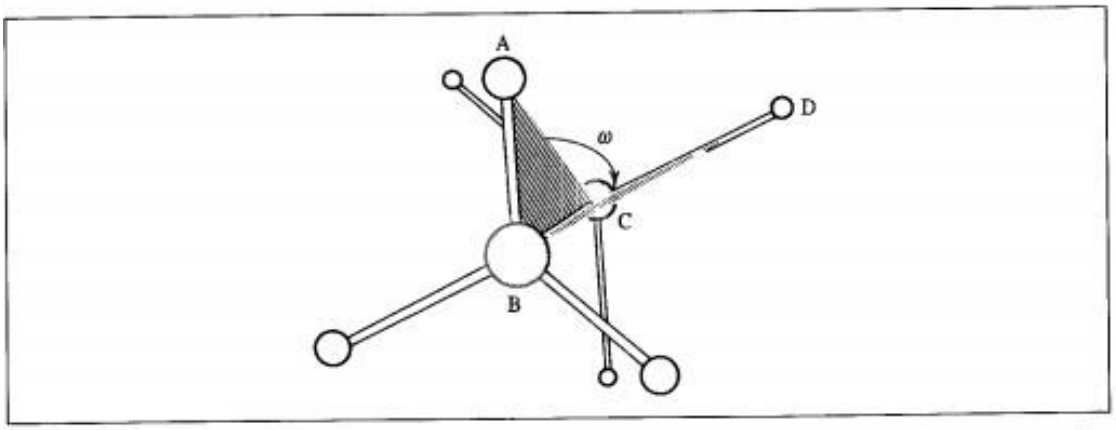


Fig. 1.2 A torsion angle A–B–C–D is defined as the angle between the planes A, B, C and B, C, D A torsion angle can vary through 360° although the range −180° to +180° is most commonly used We shall adopt the IUPAC definition of a torsion angle in which an eclipsed conformation corresponds to a torsion angle of 0° and a trans *or* anti *conformation to a torsion angle of 180°. The reader should note that this may not correspond to some of the definitions used in the literature, where the* trans *arrangement is defined as a torsion angle of 0° If one looks along the bond B–C, then the torsion angle is the angle through which it is necessary to rotate the bond AB in a clockwise sense in order to superimpose the two planes, as shown*

than the number of Cartesian coordinates for a non-linear molecule. This is because we are at liberty to arbitrarily translate and rotate the system within Cartesian space without changing the relative positions of the atoms

POTENTIAL ENERGY SURFACE

• A potential energy surface (PES) describes the energy of a system, especially a collection of atoms, in terms of certain parameters, normally the positions of the atoms. The surface might define the energy as a function of one or more coordinates; if there is only one coordinate, the surface is called a *potential energy curve*.

• The PES concept finds application in fields such as chemistry and physics, especially in the theoretical sub-branches of these subjects. It can be used to theoretically explore properties of structures composed of atoms, for example, finding the minimum energy shape of a molecule or computing the rates of a chemical reaction

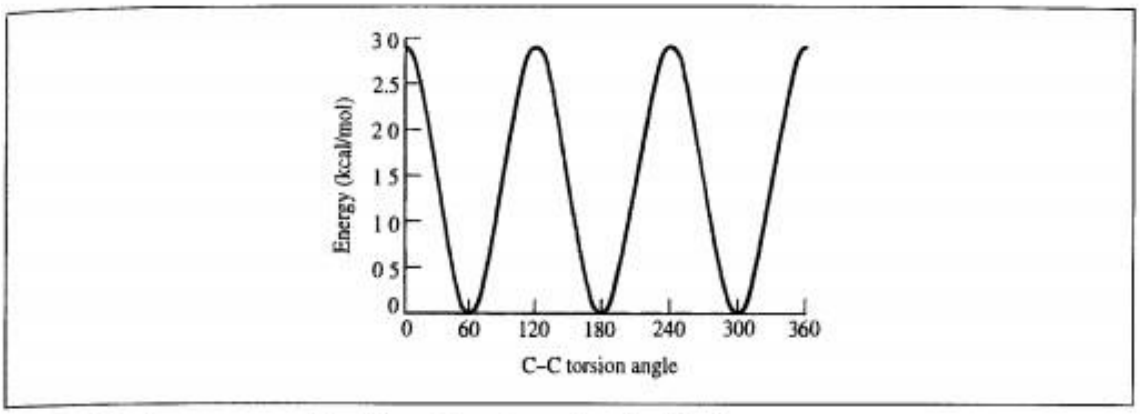


Fig. 1 3 Variation in energy with rotation of the carbon–carbon bond in ethane

Changes in the energy of a system can be considered as movements on a multidimensional 'surface' called the *energy surface*. We shall be particularly interested in stationary points on the energy surface, where the first derivative of the energy is zero with respect to the internal or Cartesian coordinates. At a stationary point the forces on all the atoms are zero. Minimum points are one type of stationary point; these correspond to stable structures. Methods for locating stationary points will be discussed in more detail in Chapter 5, together with a more detailed consideration of the concept of the energy surface.

The Molecular Modeling Toolbox

Molecular Mechanics Methods

Molecules modeled as spheres (atoms) connected by springs (bonds)

- Fast, $>10^6$ atoms
- Limited flexibility due to lack of electron treatment Typical applications
- Simulating biomolecules in explicit solvent/membrane
- Geometry optimization
- Conformational search

Quantum Mechanical Methods

- Molecules represented using electron structure (Schrödinger equation)
- Computationally expensive , <10-100 atoms, depending on method
- Highly flexible – any property can in principle be calculated Typical applications
- Chemical reactions
- Spectra
- Accurate (gas phase) structures, energies

Chapter 3 we then build upon this chapter and consider more advanced concepts. Quantum mechanics does, of course, predate the first computers by many years, and it is a tribute to the pioneers in the field that so many of the methods in common use today are based upon their efforts. The early applications were restricted to atomic, diatomic or highly symmetrical systems which could be solved by hand. The development of quantum mechanical techniques that are more generally applicable and that can be implemented on a computer (thereby eliminating the need for much laborious hand calculation) means that quantum mechanics can now be used to perform calculations on molecular systems of real, practical interest. Quantum mechanics explicitly represents the electrons in a calculation, and so it is possible to derive properties that depend upon the electronic distribution and, in particular, to investigate chemical reactions in which bonds are broken and formed. These qualities,

Fundamentals of Quantum mechanics

Light- energy- photons/quanta- wave –particle-duality

Schrodinger -Every quantum particle is characterized by wave function Developed a differential equation which describes the evolution of Predicts analytically and precisely the probability of events/outcome (TIME)

- Represents electrons in a calculation
- Derive the properties that depend on electronic distribution – particularly the chemical reactions in which bonds are broken and formed

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2}+V(x)\psi = E\psi \qquad \text{or} \qquad \hat{H}\psi = E\psi$$

- **H – Hamiltonian operator**
- E – energy of the system
- □ - wave function
- But SE can be used only for very small mol such as H and He

So approximations must be used in order to extend the utility of the method to polyatomic systems

The starting point for any discussion of quantum mechanics is, of course, the Schrödinger equation. The full, time-dependent form of this equation is

$$\left\{-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)+\mathscr{V}\right\}\Psi(\mathbf{r},t)=i\hbar\frac{\partial\Psi(\mathbf{r},t)}{\partial t} \tag{2.1}$$

Equation (2.1) refers to a single particle (e.g. an electron) of mass m which is moving through space (given by a position vector $\mathbf{r}=x\mathbf{i}+y\mathbf{j}+z\mathbf{k}$) and time ($t$) under the influence of an external field $\mathscr{V}$ (which might be the electrostatic potential due to the nuclei of a molecule). $\hbar$ is Planck's constant divided by 2π and i is the square root of -1. Ψ is the *wavefunction* which characterises the particle's motion; it is from the wavefunction that we can derive various properties of the particle. When the external potential $\mathscr{V}$ is independent of time then the wavefunction can be written as the product of a spatial part and a time part: $\Psi(\mathbf{r},t)=\psi(\mathbf{r})T(t)$. We shall only consider situations where the potential is independent of time, which enables the time-dependent Schrödinger equation to be written in the more familiar, time-independent form:

$$\left\{-\frac{\hbar^2}{2m}\nabla^2+\mathscr{V}\right\}\Psi(\mathbf{r})=E\Psi(\mathbf{r}) \tag{2.2}$$

Here, E is the energy of the particle and we have used the abbreviation ∇^2 (pronounced 'del-squared').

$$\nabla^2=\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2} \tag{2.3}$$

It is usual to abbreviate the left-hand side of Equation (2.1) to $\mathscr{H}\Psi$, where $\mathscr{H}$ is the *Hamiltonian operator*:

$$\mathscr{H}=-\frac{\hbar^2}{2m}\nabla^2+\mathscr{V} \tag{2.4}$$

This reduces the Schrödinger equation to $\mathscr{H}\Psi=E\Psi$. To solve the Schrödinger equation it is necessary to find values of E and functions Ψ such that, when the wavefunction is operated upon by the Hamiltonian, it returns the wavefunction multiplied by the energy. The Schrödinger equation falls into the category of equations known as partial differential eigenvalue equations in which an operator acts on a function (the eigenfunction) and returns the

Time-Dependent Schrodinger Wave Equation

Eigenvalue equation

$$i\hbar\frac{\partial}{\partial t}\Psi(x,t) = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\Psi(x,t) + V(x)\Psi(x,t)$$

PHYSICS NOTATION

Total E term

K.E. term

P.E. term

$$\Psi(x,t) = e^{-iEt/\hbar}\psi(x)$$

Time-Independent Schrodinger Wave Equation

$$E\psi(x) = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\psi(x) + V(x)\psi(x)$$

2.2 One-electron Atoms

In an atom that contains a single electron, the potential energy depends upon the distance between the electron and the nucleus as given by the Coulomb equation. The Hamiltonian thus takes the following form:

$$\mathscr{H} = -\frac{\hbar^2}{2m}\nabla^2 - \frac{Ze^2}{4\pi\varepsilon_0 r} \tag{2.16}$$

In atomic units the Hamiltonian is:

$$\mathscr{H} = -\frac{1}{2}\nabla^2 - \frac{Z}{r} \tag{2.17}$$

For the hydrogen atom, the nuclear charge, Z, equals $+1$. r is the distance of the electron from the nucleus. The helium cation, He^+, is also a one-electron atom but has a nuclear charge of $+2$. As atoms have spherical symmetry it is more convenient to transform the Schrödinger equation to polar coordinates r, θ and ϕ, where r is the distance from the nucleus (located at the origin), θ is the angle to the z axis and ϕ is the angle from the x axis in the xy plane (Figure 2.1). The solutions can be written as the product of a radial function $R(r)$, which depends only on r, and an angular function $Y(\theta, \phi)$ called a *spherical harmonic*, which

2.3.1 The Born–Oppenheimer Approximation

It was stated above that the Schrödinger equation cannot be solved exactly for any molecular systems However, it is possible to solve the equation exactly for the simplest molecular species, H_2^+ (and isotopically equivalent species such as HD^+), when the motion of the electrons is decoupled from the motion of the nuclei in accordance with the Born–Oppenheimer approximation. The masses of the nuclei are much greater than the masses of the electrons (the resting mass of the lightest nucleus, the proton, is 1836 times heavier than the resting mass of the electron). This means that the electrons can adjust almost instantaneously to any changes in the positions of the nuclei. The electronic wavefunction thus depends only on the positions of the nuclei and not on their momenta. Under the Born–Oppenheimer approximation the total wavefunction for the molecule can be written in the following form:

$$\Psi_{tot}(\text{nuclei, electrons}) = \Psi(\text{electrons})\Psi(\text{nuclei}) \qquad (2.31)$$

The total energy equals the sum of the nuclear energy (the electrostatic repulsion between the positively charged nuclei) and the electronic energy. The electronic energy comprises

Molecular Mechanics Force Field

The "mechanical" molecular model was developed out of a need to describe molecular structures and properties in as practical a manner as possible. The range of applicability of molecular mechanics includes:

- Molecules containing thousands of atoms.
- Organics, oligonucleotides, peptides, and saccharides (metallo-organics and inorganic
- Vacuum, implicit, or explicit solvent environments.
- Ground state only.
- Thermodynamic and kinetic (via molecular dynamics) properties.

The great computational speed of molecular mechanics allows for its use in procedures such as molecular dynamics, conformational energy searching, and docking. All the procedures require large numbers of energy evaluations.

Molecular mechanics methods are based on the following principles:

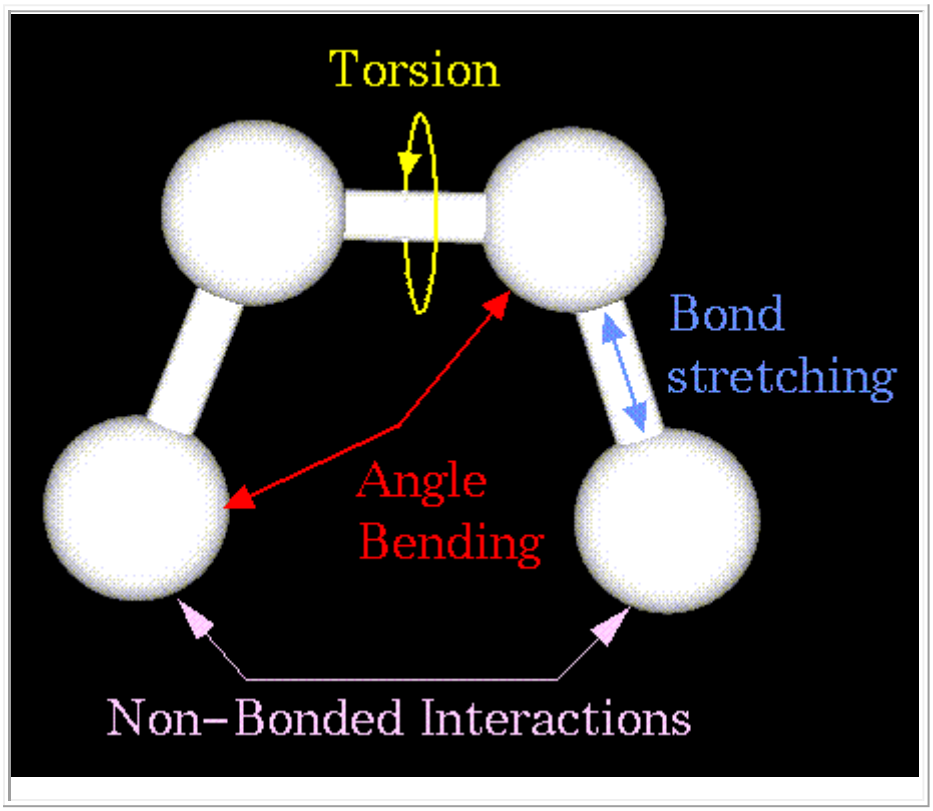
- Nuclei and electrons are lumped into atom-like particles.

- Atom-like particles are spherical (radii obtained from measurements or theory) and have a net charge (obtained from theory).
- Interactions are based on springs and classical potentials.
- Interactions must be preassigned to specific sets of atoms.

Interactions determine the **spatial distribution** of atom-like particles and their energies.

To define a force field one must specify not only the functional form but also the parameters (i.e.the various constants). Two force fields may use an identical functional form yet have very different parameters. A force field should be considered as a single entity; it is not strictly correct to divide the energy into its individual components, let alone to take some of the parameters from one forcefield and mix them with parameters from another force field. The forcefields used in molecular modelling are primarily designed to reproduce structural properties but they can also be used to predict other properties, such as molecular spectra. However, molecular mechanics force fields can rarely predict spectra with great accuracy (although the more recent molecular. mechanics force fields are much better in this regard). A force field is generally designed to predict certain properties and will be parametrised accordingly. While it is useful to try to predict other quantities which have not been included in the parametrisation process it is not necessarily a failing if a force field is unable to do so. Transferability of the functional form and parameters is an important feature of a forcefield. Transferability means that the same set of parameters can be used to model a series of related molecules, rather than having to define a new set of parameters for each individual molecule. A concept that is common to most force fields is that of an atom type. When preparing the input for a quantum mechanics calculation it is usually necessary to specify the atomic numbers of the nuclei present, together with the geometry of the system and the overall charge and spin multiplicity. For a force field the overall charge and spin multiplicity are not explicitly required, but it is usually necessary to assign an atom type to each atom in the system. The atom type is more than just the atomic number of an atom; it usually contains information about its hybridisation state and sometimes the local environment. For example, it is necessary in most force fields to distinguish between sp3 - hybridised carbon atoms (which adopt a tetrahedral geometry), sp2-hybridised carbons (which are trigonal) and sp-hybridised carbons (which are linear).

The mechanical molecular model considers atoms as spheres and bonds as springs. The mathematics of spring deformation can be used to describe the ability of bonds to stretch, bend, and twist:



Non-bonded atoms (greater than two bonds apart) interact through van der Waals attraction, steric repulsion, and electrostatic attraction/repulsion. These properties are easiest to describe mathematically when atoms are considered as spheres of characteristic radii.

The object of molecular mechanics is to predict the energy associated with a given conformation of a molecule. However, molecular mechanics energies have no meaning as absolute quantities. Only differences in energy between two or more conformations have meaning. A simple molecular mechanics energy equation is given by:

Energy = Stretching Energy + Bending Energy + Torsion Energy + Non-Bonded Interaction Energy

- A **force field** refers to the form and parameters of mathematical functions used to describe the potential energy of a system of particles (typically molecules and atoms).
- calculates the molecular system's potential energy (E) in a given conformation as a sum of individual energy terms.
- where the components of the covalent and noncovalent contributions are given by the following summations:

$$E_{\text{noncovalent}} = E_{\text{electrostatic}} + E_{\text{van der Waals}}$$

- where the components of the covalent and noncovalent contributions are given by the following summations

$$E_{\text{covalent}} = E_{\text{bond}} + E_{\text{angle}} + E_{\text{dihedral}}$$

$$E_{\text{noncovalent}} = E_{\text{electrostatic}} + E_{\text{van der Waals}}$$

- FF is a mathematical function which returns the energy of the system as a function of the conformation of the system.

These equations together with the data (parameters) required to describe the behavior of different kinds of atoms and bonds, is called a force-field. Many different kinds of force-fields have been developed over the years. Some include additional energy terms that describe other kinds of deformations. Some force-fields account for coupling between bending and stretching in adjacent bonds in order to improve the accuracy of the mechanical model.

$$\mathscr{V}(\mathbf{r}^N) = \sum_{\text{bonds}} \frac{k_i}{2}(l_i - l_{i,0})^2 + \sum_{\text{angles}} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2 + \sum_{\text{torsions}} \frac{V_n}{2}(1 + \cos(n\omega - \gamma))$$
$$+ \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right)$$

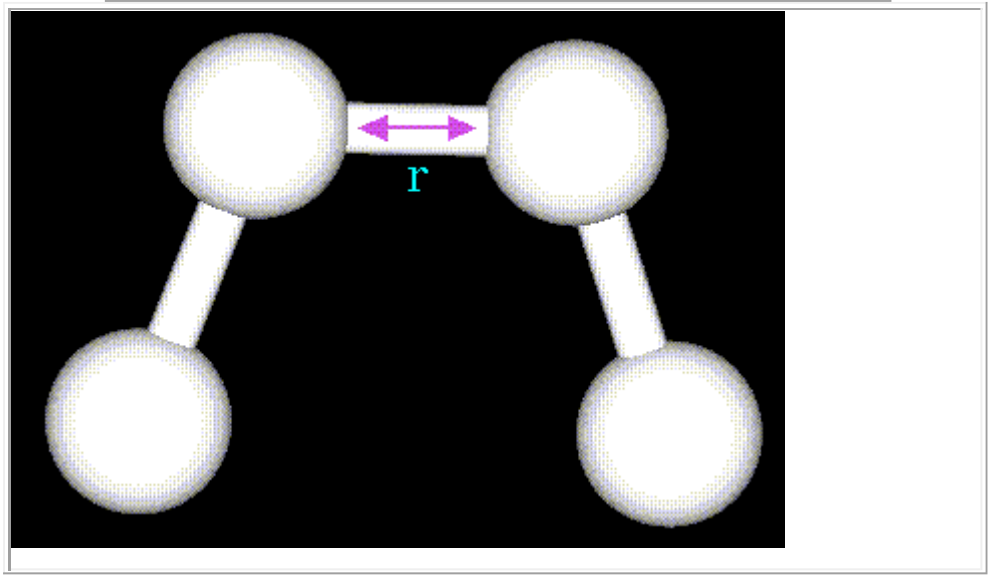
$\mathscr{V}(\mathbf{r}^N)$ Potential energy as a function of position r of N particles

- Reproduce the structural properties such as molecular spectra
- Transferability

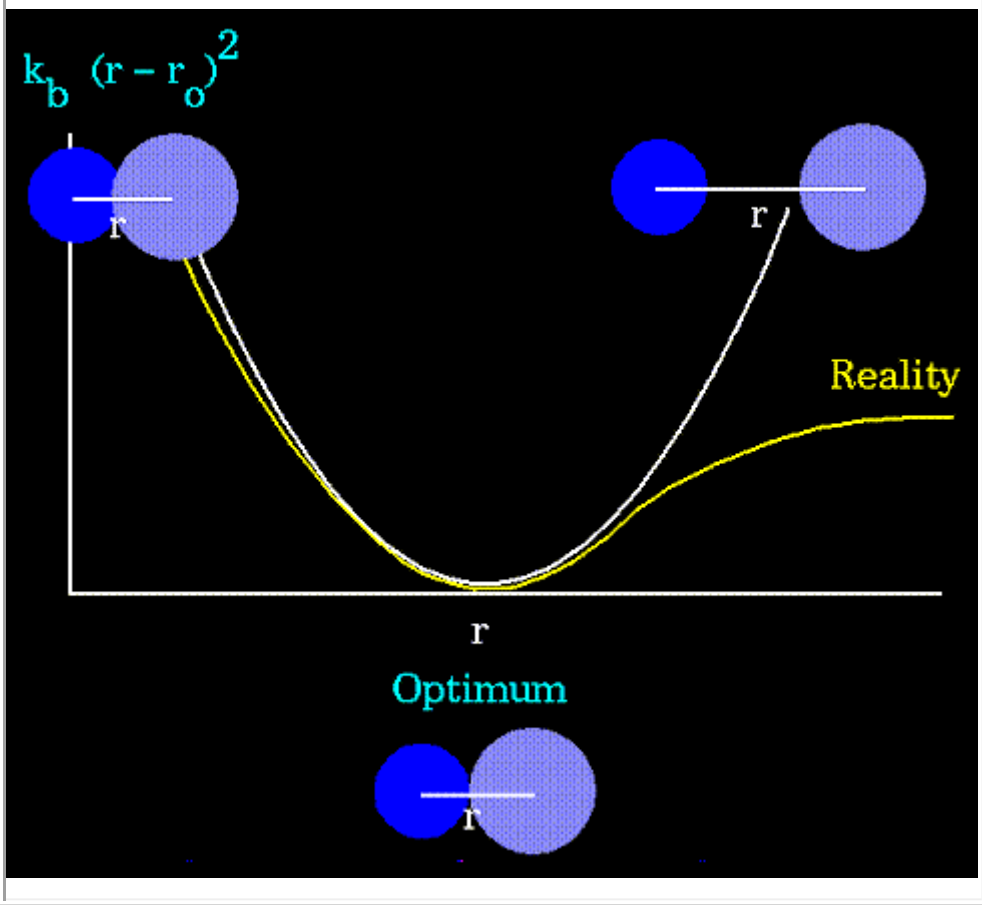
The mathematical form of the energy terms varies from force-field to force-field. The more common forms will be described.

Stretching Energy

$$E = \sum_{bonds} k_b (r - r_o)^2$$



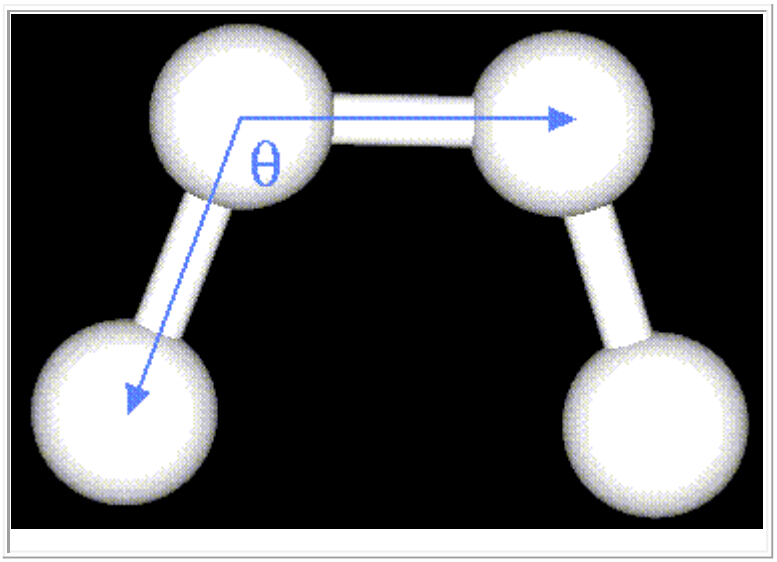
The stretching energy equation is based on Hooke's law. The "k*b*" parameter controls the stiffness of the bond spring, while "r*o*" defines its equilibrium length. Unique "k*b*" and "r*o*" parameters are assigned to each pair of bonded atoms based on their types (e.g. C-C, C-H, O-C, etc.). This equation estimates the energy associated with vibration about the equilibrium bond length. This is the equation of a parabola, as can be seen in the following plot:



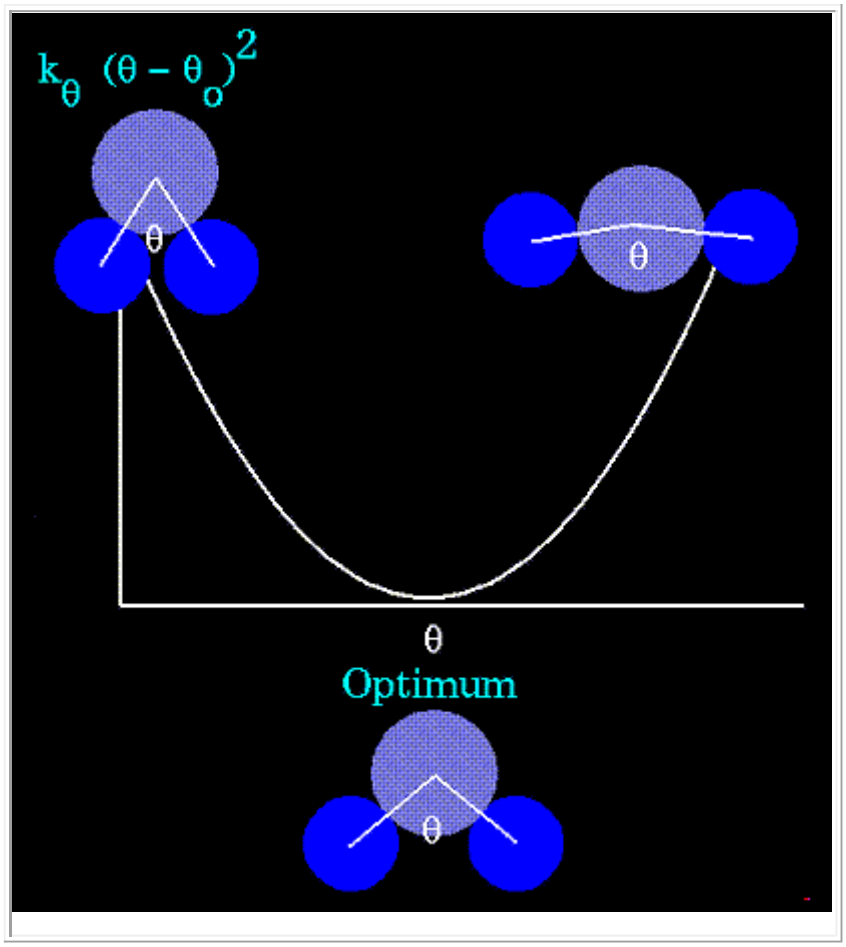
Notice that the model tends to break down as a bond is stretched toward the point of dissociation.

Bending Energy

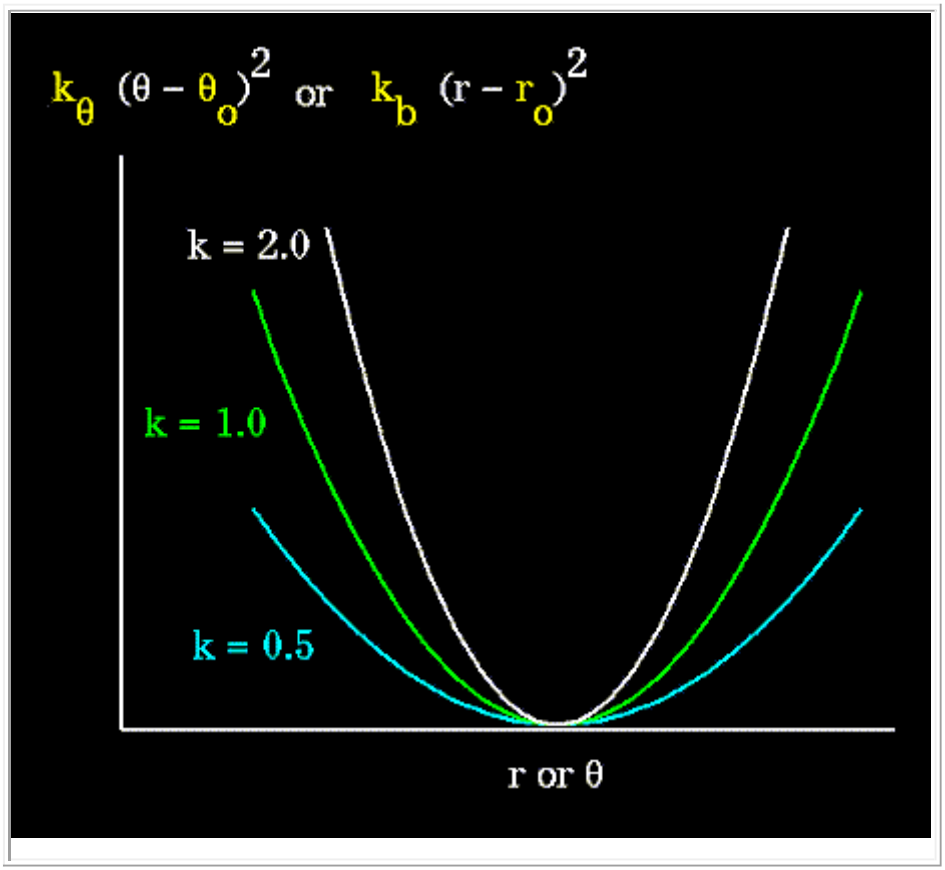
$$E = \sum_{\text{angles}} k_\theta (\theta - \theta_o)^2$$



The bending energy equation is also based on Hooke's law. The "k*theta*" parameter controls the stiffness of the angle spring, while "theta*o*" defines its equilibrium angle. This equation estimates the energy associated with vibration about the equilibrium bond angle:

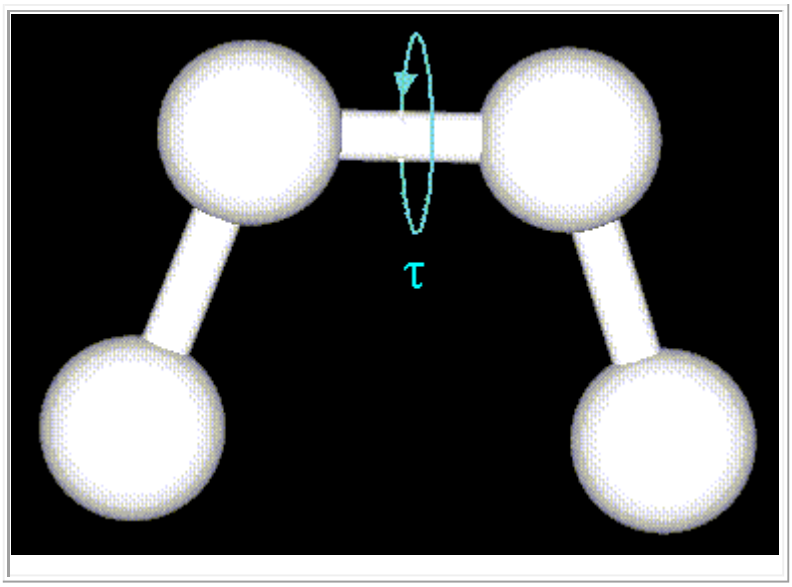


Unique parameters for angle bending are assigned to each bonded triplet of atoms based on their types (e.g. C-C-C, C-O-C, C-C-H, etc.). The effect of the "k*b*" and "k*theta*" parameters is to broaden or steepen the slope of the parabola. The larger the value of "k", the more energy is required to deform an angle (or bond) from its equilibrium value. Shallow potentials are achieved for "k" values between 0.0 and 1.0. The Hookeian potential is shown in the following plot for three values of "k":

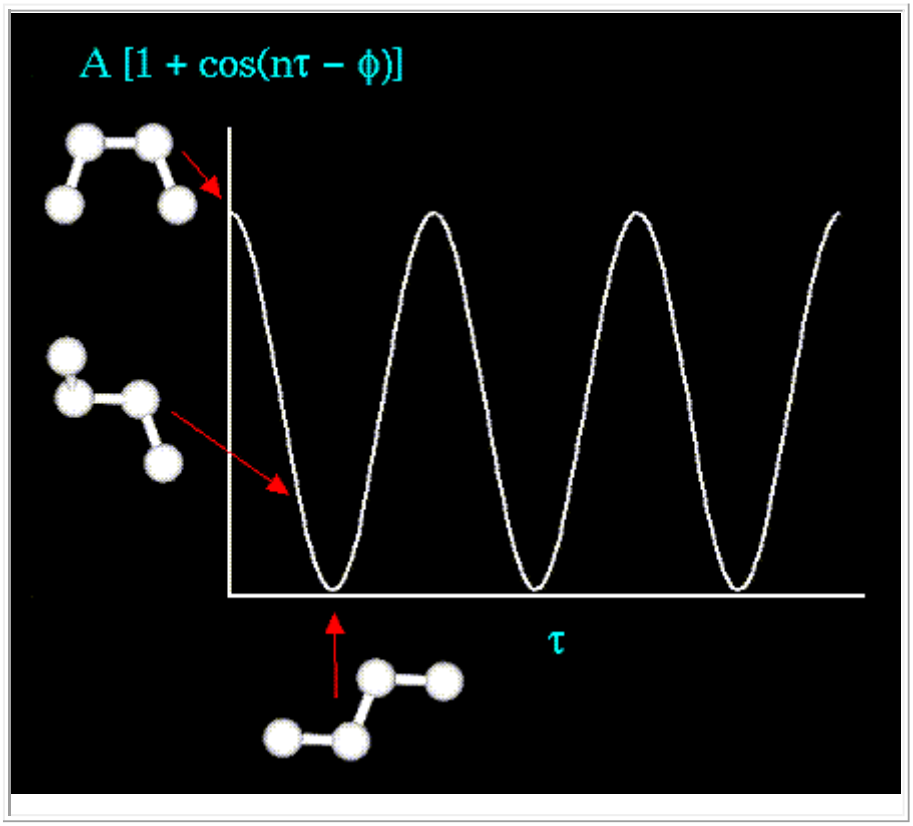


Torsion Energy

$$E = \sum_{\text{torsions}} A\,[1 + \cos(n\tau - \phi)]$$



The torsion energy is modeled by a simple periodic function, as can be seen in the following plot:



The torsion energy in molecular mechanics is primarily used to correct the remaining energy terms rather than to represent a physical process. The torsional energy represents the amount of energy that must be added to or subtracted from the Stretching Energy + Bending Energy

+ Non-Bonded Interaction Energy terms to make the total energy agree with experiment or rigorous quantum mechanical calculation for a model dihedral angle (ethane, for example might be used a a model for any H-C-C-H bond).

Cross terms

The presence of cross terms in a forcefield reflects coupling between the internal coordinates. For example, as a bond angle is decreased it is found that the adjacent bonds stretch to reduce the interaction between the 1,3 atoms, as illustrated in Figure.

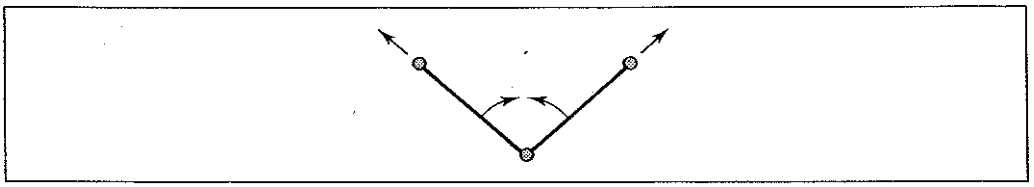
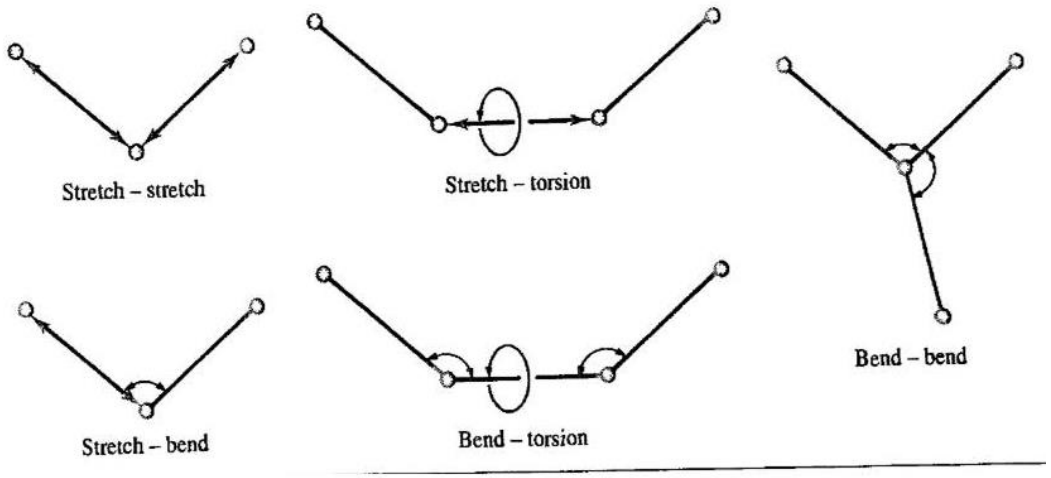


Fig. 4.12: Coupling between the stretching of the bonds as an angle closes.

One should in principle include cross terms between all contributions to a force field. However, only a few cross terms are generally found to be necessary in order to reproduce structural properties accurately; more may be needed to reproduce other properties such as vibrational frequencies, which are more sensitive to the presence of such terms. In general, any interactions involving motions that are far apart in a molecule can usually be set to zero. Most cross terms are functions of two internal coordinates, such as stretch-stretch, stretch-bend and stretch-torsion terms, but cross terms involving more than two internal coordinates such as the bend- bend-torsion have also been used.

Cross terms



Various functional forms are possible for the cross terms. For example, the stretch-s

$$v(l_1, l_2) = \frac{k_{l_1,l_2}}{2}[(l_1 - l_{1,0})(l_2 - l_{2,0})] \tag{4.13}$$

The stretching of the two bonds adjoining an angle could be modelled using an equation of the following form (as in MM2, MM3 and MM4):

$$v(l_1, l_2, \theta) = \frac{k_{l_1,l_2,\theta}}{2}[(l_1 - l_{1,0}) + (l_2 - l_{2,0})](\theta - \theta_0) \tag{4.14}$$

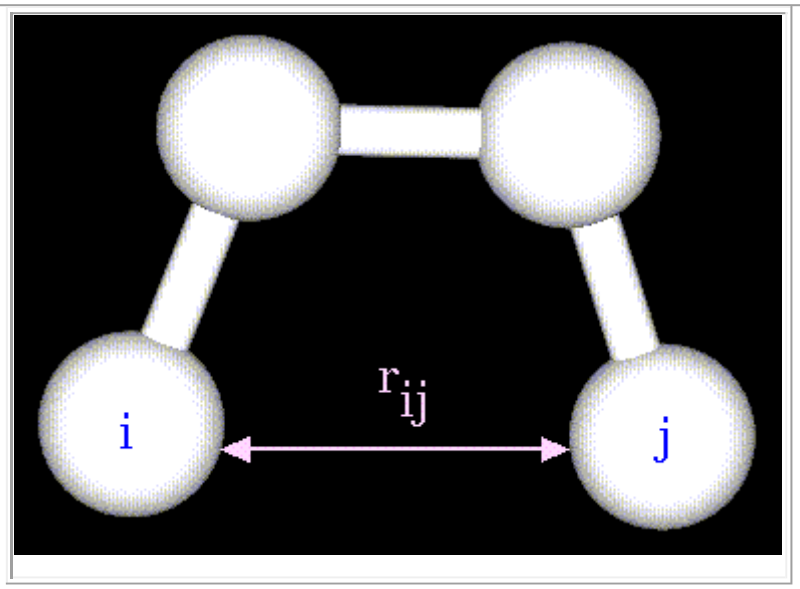
Non-Bonded Energy

Independent molecules and atoms interact through non-bonded forces, which also play an important role in determining the structure of individual molecular species. The non-bonded interactions do not depend upon a specific bonding relationship between atoms. They are 'through-space' interactions and are usually modelled as a function of some inversepower of the distance. The non-bonded terms in a forcefield are usually considered in two groups, one comprising electrostatic interactions and the other van der Waals interactions.

The non-bonded energy represents the pair-wise sum of the energies of all possible interacting non-bonded atoms i and j:

$$E = \sum_i \sum_j \frac{-A_{ij}}{r_{ij}^6} + \frac{B_{ij}}{r_{ij}^{12}} + \sum_i \sum_j \frac{q_i q_j}{r_{ij}}$$

van der Waals term Electrostatic term



The non-bonded energy accounts for repulsion, van der Waals attraction, and electrostatic interactions.

Van der Waals attraction occurs at short range, and rapidly dies off as the interacting atoms move apart by a few Angstroms. Repulsion occurs when the distance between interacting atoms becomes even slightly less than the sum of their contact radii. Repulsion is modeled by an equation that is designed to rapidly blow up at close distances. The energy term that describes attraction/repulsion provides for a smooth transition between these two regimes. These effects are often modeled using a 6-12 equation, as shown in the following plot:

Electrostatic interactions

Electrostatic interactions also arise from changes in the charge distribution of a molecule or atom caused by an external field, a process called polarisation. The primary effect of the external electric field (which in our case will be caused by neighbouring molecules) is to induce a dipole in the molecule. The magnitude of the induced dipole moment μind is proportional to the electric field E, with the constant of proportionality being the polarisability a:

$$\boldsymbol{\mu}_{\text{ind}} = \alpha \mathbf{E} \tag{4.51}$$

The energy of interaction between a dipole $\boldsymbol{\mu}_{\text{ind}}$ and an electric field $\mathbf{E}$ (the induction energy) is determined by calculating the work done in charging the field from zero to E, using the following integral:

$$v(\alpha, E) = -\int_0^E d\mathbf{E}\,\boldsymbol{\mu}_{\text{ind}} = -\int_0^E d\mathbf{E}\,\alpha\mathbf{E} = -\tfrac{1}{2}\alpha E^2 \tag{4.52}$$

In strong electric fields contributions to the induced dipole moment that are proportional to E^2 or E^3 can also be important, and higher-order moments such as quadrupoles can also be induced. We will not be concerned with such contributions.

The electrostatic contribution is modeled using a Coulombic potential. The electrostatic energy is a function of the charge on the non-bonded atoms, their interatomic distance, and a molecular dielectric expression that accounts for the attenuation of electrostatic interaction by the environment (e.g. solvent or the molecule itself). Often, the molecular dielectric is set to a constant value between 1.0 and 5.0. A linearly varying distance-dependent dielectric (i.e. 1/r) is sometimes used to account for the increase in environmental bulk as the separation distance between interacting atoms increases.

ENERGY MINIMISATION

In the field of computational chemistry, **energy minimization** (also called **energy optimization**, **geometry minimization**, or **geometry optimization**) is the process of finding an arrangement in space of a collection of atoms where, according to some computational model of chemical bonding, the net inter-atomic force on each atom is acceptably close to zero and the position on the potential energy surface (PES) is a stationary point. The collection of atoms might be a single molecule, an ion, a condensed phase, a transition state or even a collection of any of these. The computational model of chemical bonding might, for example, be quantum mechanics.

As an example, when optimizing the geometry of a water molecule, one aims to obtain the hydrogen-oxygen bond lengths and the hydrogen-oxygen-hydrogen bond angle which minimize the forces that would otherwise be pulling atoms together or pushing them apart.

The motivation for performing a geometry optimization is the physical significance of the obtained structure: optimized structures often correspond to a substance as it is found in nature and the geometry of such a structure can be used in a variety of experimental and theoretical investigations in the fields of chemical structure, thermodynamics, chemical kinetics, spectroscopy and others.

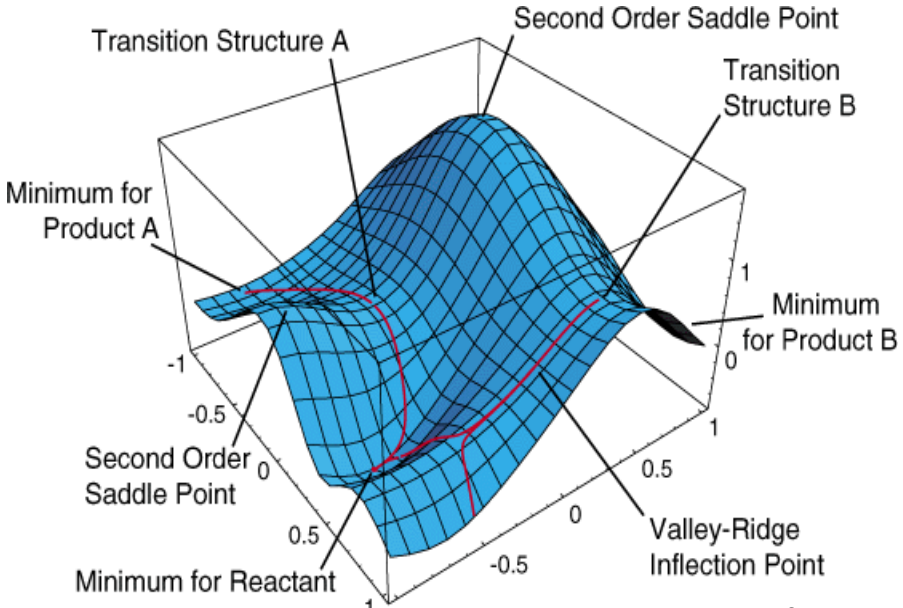
Typically, but not always, the process seeks to find the geometry of a particular arrangement of the atoms that represents a local or global energy minimum. Instead of searching for global energy minimum, it might be desirable to optimize to a transition state, that is, a saddle point on the potential energy surface. Additionally, certain coordinates (such as a chemical bond length) might be fixed during the optimization.

- Energy minimization methods can precisely locate minimum energy conformations by mathematically "homing in" on the energy function minima (one at a time).
- The goal of energy minimization is to find a route (consisting of variation of the intramolecular degrees of freedom) from an initial conformation to the nearest minimum energy conformation using the smallest number of calculations possible.
- The way in which the energy varies with the coordinates is usually referred to as PES or hyper surface
- Energy of any conformation is a function of its internal or cartesian coordinates

- N atoms – energy is a function of 3N-6 internal coordinates or 3N cartesian coordinates
- Changes in the energy are a function of its nuclear coordinates.

Potential energy surface

- Changes in the energy of a system can be considered as movements on a multidimensional surface called energy surface.
- Changes in the energy funtion of its nuclear coordinates.
- Movement of the nuclei influences change in energy
- Mathematical function that gives the energy of a molecule as a function of its geometry
- Energy is plotted on the vertical axis, geometric coordinates (e.g bond lengths, valence angles, etc.) are plotted on the horizontal axes
- A PES can be thought of it as a hilly landscape, with valleys, mountain passes and peaks
- Real PES have many dimensions, but key feature can be represented by a 3 dimensional PES



- Equilibrium molecular structures correspond to the positions of the minima in the valleys on a PES
- Energetics of reactions can be calculated from the energies or altitudes of the minima for reactants and products
- A reaction path connects reactants and products through a mountain pass
- A transition structure is the highest point on the lowest energy path

- Reaction rates can be obtained from the height and profile of the potential energy surface around the transition structure
- The shape of the valley around a minimum determines the vibrational spectrum
- Each electronic state of a molecule has a separate potential energy surface, and the separation between these surfaces yields the electronic spectrum
- Properties of molecules such as dipole moment, polarizability, NMR shielding, etc. depend on the response of the energy to applied electric and magnetic fields

- Minima, lowest – global energy minima
- Minimization algorithms
- Highest point in the pathway between 2 minima is saddle point represents the transition state
- Minima and saddle points are stationary states on PES where the first derivative of energy function is 0
- $E = f(x)$
- E is a function of coordinates either cartesian or internal
- At minimum the first derivatives are zero and the second derivatives are all positive

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING
DEPARTMENT OF BIOINFORMATICS

UNIT – 4- SBI1208 – Cheminformatics

Computer simulation

Computer simulation is the process of mathematical modelling, performed on a computer, which is designed to predict the behaviour of or the outcome of a real-world or physical system. Since they allow to check the reliability of chosen mathematical models, computer simulations have become a useful tool for the mathematical modeling of many natural systems in physics (computationalphysics), astrophysics, climatology, chemistry, biology and manufacturing, as well as human systems in economics, psychology, social science, health care and engineering. Simulation of a system is represented as the running of the system's model. It can be used to explore and gain new insights into new technology and to estimate the performance of systems too complex for analytical solutions.

A computer model is the algorithms and equations used to capture the behavior of the system being modeled. By contrast, computer simulation is the actual running of the program that contains these equations or algorithms. Simulation, therefore, is the process of running a model. Thus one would not "build a simulation"; instead, one would "build a model", and then either "run the model" or equivalently "run a simulation"

Benefits

- Gain greater understanding of a process
- Identify problem areas or bottlenecks in processes
- Evaluate effect of systems or process changes such as demand, resources, supply, and constraints
- Identify actions needed upstream or downstream relative to a given operation, organization, or activity to either improve or mitigate processes or events
- Evaluate impact of changes in policy prior to implementation

Types

- Discrete Models – Changes to the system occur at specific times
- Continuous Models – The state of the system changes continuously over time
- Mixed Models – Contains both discrete and continuous elements

Types of Data/Information Needed to Develop a Simulation Model:

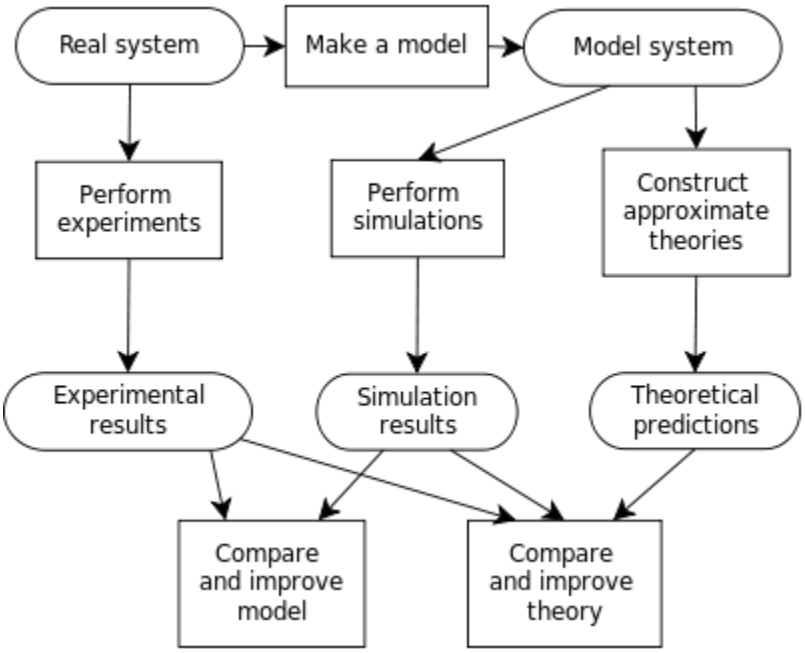
- The overall process flow and its associated resources
- What is being produced, served, or acted upon by the process (entities)
- Frequency at which the entities arrive in the process
- How long do individual steps in the process take
- Probability distributions that characterize real life uncertainties and variations in the process

- Computer simulation is the use of a computer to represent the dynamic responses of one system by the behavior of another system modeled after it.
- A simulation uses a mathematical description, or model, of a real system in the form of a computer program.

- This model is composed of equations that duplicate the functional relationships within the real system.
- When the program is run, the resulting mathematical dynamics form an analog of the behavior of the real system, with the results presented in the form of data.
- A simulation can also take the form of a computer-graphics image that represents dynamic processes in an animated sequence.
- Computer simulations have become a useful part of mathematical modeling of many natural systems in physics, astrophysics, chemistry, biology, climatology, psychology, social science, etc

USES

- Computer simulations are used to study the dynamic behavior of objects or systems in response to conditions that cannot be easily or safely applied in real life.
- Simulations are especially useful in enabling observers to measure and predict how the functioning of an entire system may be affected by altering individual components within that system.
- Simulations have great military applications also. Many uses for a computer simulation can be found within various scientific fields of study such as meteorology, physical sciences, etc



Process of building a computer model, and the interplay between experiment, simulation, and theory.

Advantages of Computer Simulation

- To test a system without having to create the system for real (Building real-life systems can be expensive, and take a long time)
- To predict what might happen to a system in the future (An accurate model allows us to go forward in virtual time to see what the system will be doing in the future)
- To train people to use a system without putting them at risk (Learning to fly an airplane is very difficult and mistake will be made. In a real plane mistakes could be fatal!)
- To investigate a system in great detail (A model of a system can be zoomed in/out or rotated. Time can be stopped, rewound, etc.)

22

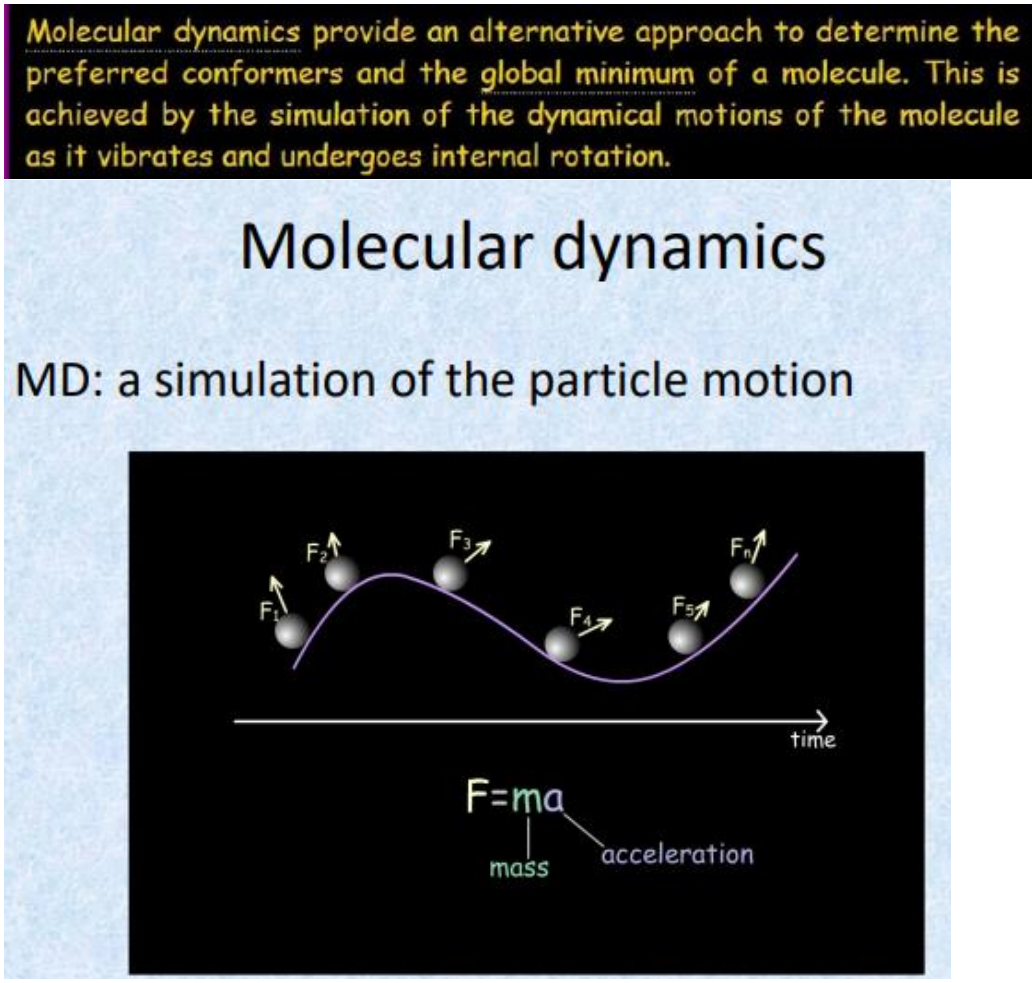
Disadvantages of Computer Simulation

- It depends upon the skills of the model maker
- Time consuming and high cost
- Sometimes, it does not give the best solution for system

Basic Simulation Techniques

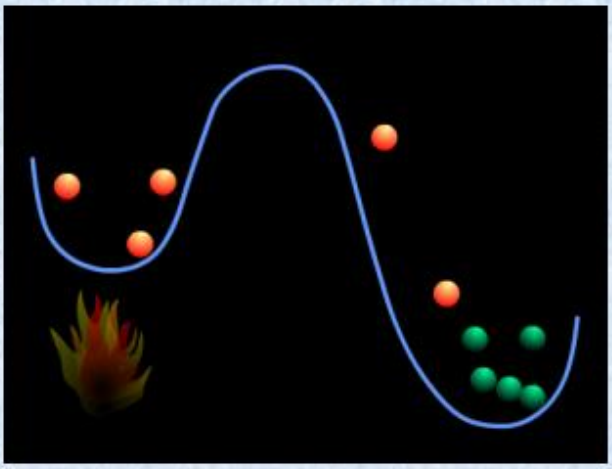
To explore the energy landscape described by the molecular mechanics force field, *i.e.* to sample molecular conformations, a simulation is required. This is also the route to relate the microscopic movements and positions of the atoms to the macroscopic or thermodynamic quantities that can be measured experimentally. There are two major simulation methods to sample biomolecular systems: molecular dynamics (MD) and Monte Carlo (MC)

Molecular dynamics (MD) is a computer simulation method for analyzing the physical movements of atoms and molecules. The atoms and molecules are allowed to interact for a fixed period of time, giving a view of the dynamic "evolution" of the system.



Molecular dynamics

The motion (determined by the temperature) allows conformational changes



Molecular dynamics

- Calculates the time dependent behaviour of a molecular system
- Provides detailed information on the fluctuations and conformational changes of macromolecules
- Routinely used to investigate the structure, dynamics and thermodynamics of biological molecules
- Used in the determination of structures from x-ray and NMR experiments

In a molecular dynamics (MD) simulation it is possible to explore the macroscopic properties of a system

The connection between microscopic simulation and macroscopic properties is made through statistical mechanics

Allows to study both thermodynamic properties and time dependent (kinetic) phenomenon

A MD simulation is practically carried out through the application of the Newton law:

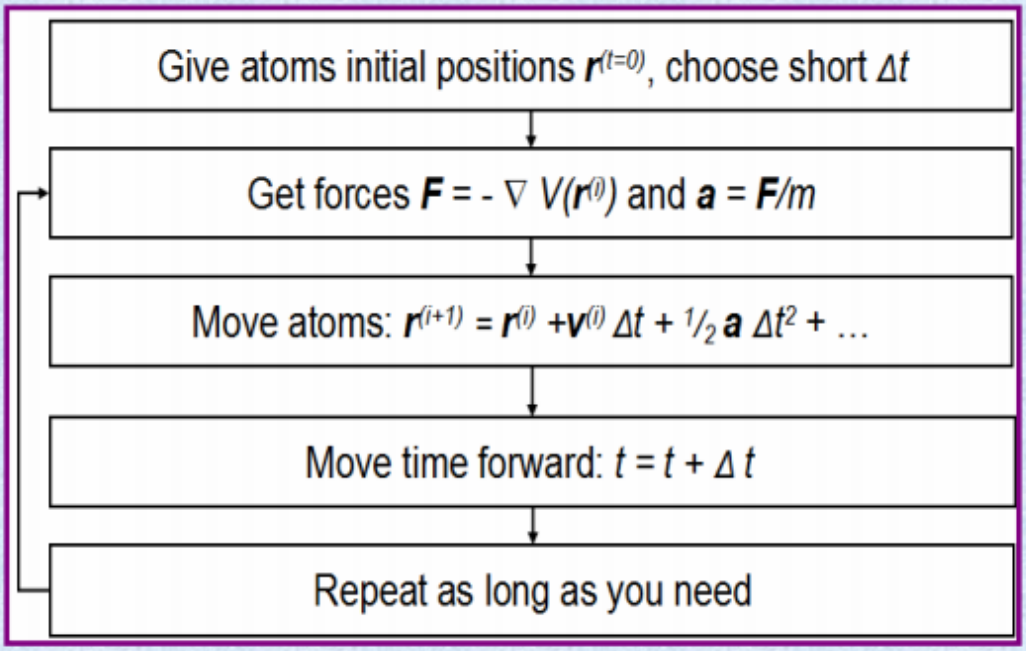
$$f = m \times a$$

The motion of each particle of the system is calculated from a

a is calculated from f

f is calculated from the potential V

A very simplified MD run



Molecular dynamics

- The potential **V** can be calculated at different accuracy level (from MM to QM)
- In biology the potential **V** is generally obtained by a MM force field
- This is a classical treatment allowing the calculation of conformational changes but usually it is not able to reproduce chemical reactions

Δt cannot be longer than the fastest atomic motion, therefore:

$$\Delta t = 10^{-15}$$

consequently a simulation of a microsecond needs one billion steps

Molecular dynamics

Temperature is directly correlated with kynetic energy:

$$K = \frac{3}{2} N k_B T$$

Generally a "free" evolution of the system is not allowed. Constraints on temperature and/or pressure are imposed in order to reproduce a particular ensemble.

Molecular dynamics

Environment simulation

The solvent can be simulated in an implicit and in an explicit manner.

Implicit solvent (in most cases the *continuum* approximation is used): fast calculation but poor results

Explicit solvent (*periodic boundary conditions* are generally used): accurate results but time consuming

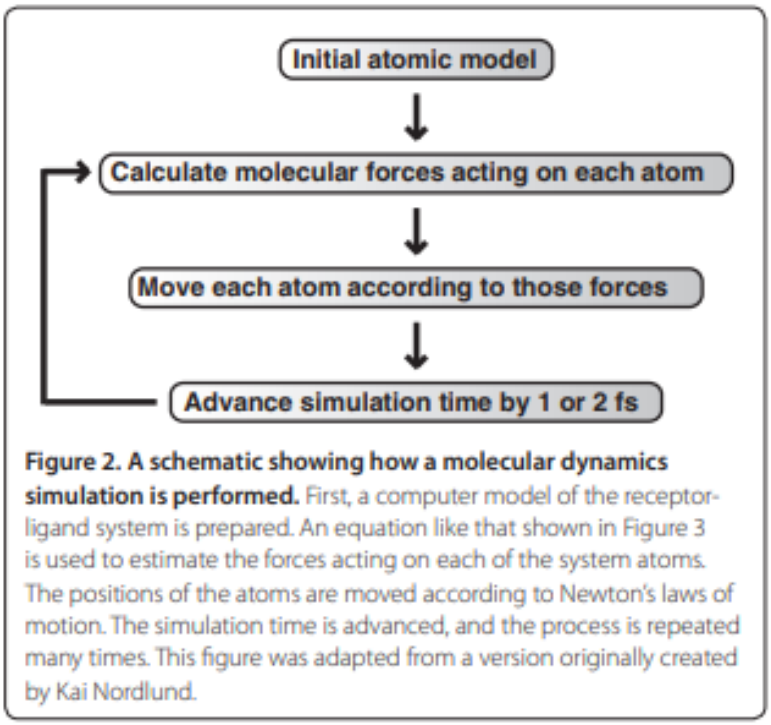


Figure 2. A schematic showing how a molecular dynamics simulation is performed. First, a computer model of the receptor-ligand system is prepared. An equation like that shown in Figure 3 is used to estimate the forces acting on each of the system atoms. The positions of the atoms are moved according to Newton's laws of motion. The simulation time is advanced, and the process is repeated many times. This figure was adapted from a version originally created by Kai Nordlund.

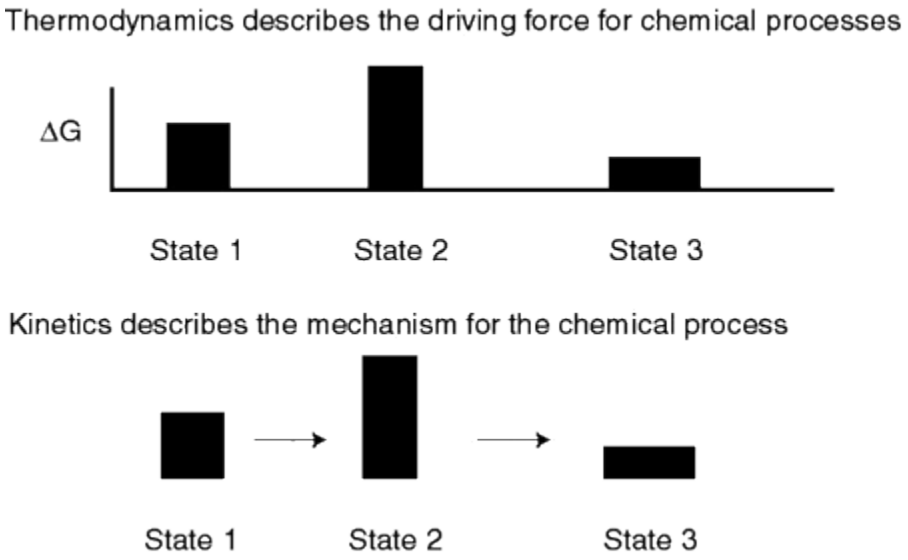
Analyses in MD Simulation

The common output from MD simulations includes positions, velocities, potential energies. Some other useful information can also be analyzed from the trajectory file.

1. Conformational analysis: analyzing conformational changes of proteins (stability, folding or unfolding), nucleic acids or polymers in different solutions or temperatures. 16

2. Hydrogen bonds, coordination bonds analysis: analyzing number and occupancy of hydrogen bonds or coordination bonds of selected groups.
3. Chemical shift analysis: predicting the chemical shift of each atom in a molecule in nuclear magnetic resonance (NMR) spectroscopy.
4. pKa value analysis: predicting protonation states of residues on proteins or polymers in aqueous solution at various pH values.
5. Protein-ligand docking: searching the binding site for ligands on the surface of a protein based on geometric complementary and scoring functions for drug design or protein purification.
6. Interaction energy analysis: calculating VdW and electrostatic interaction energies between two selected groups.
7. Water dynamics analysis: calculating residence time, the self-diffusion coefficient, or molecular orientations in selected regions.
8. Free energy analysis: calculating relative free energies between different states such as solvation free energy and binding free energy.
9. Mechanistic analysis: constructing the free energy surface with defined reaction coordinates to investigate biological processes such as the ion channel, as well as the reaction mechanisms of the enzyme or catalysis

In a molecular dynamics simulation, one often wishes to explore the macroscopic properties of a system through microscopic simulations, for example, to calculate changes in the binding free energy of a particular drug candidate, or to examine the energetics and mechanisms of conformational change. The connection between microscopic simulations and macroscopic properties is made via statistical mechanics which provides the rigorous mathematical expressions that relate macroscopic properties to the distribution and motion of the atoms and molecules of the N-body system; molecular dynamics simulations provide the means to solve the equation of motion of the particles and evaluate these mathematical formulas. With molecular dynamics simulations, one can study both thermodynamic properties and/or time dependent (kinetic) phenomenon.



Statistical mechanics is the branch of physical sciences that studies macroscopic systems from a molecular point of view. The goal is to understand and to predict macroscopic phenomena from the properties of individual molecules making up the system. The system could range from a

collection of solvent molecules to a solvated protein-DNA complex. In order to connect the macroscopic system to the microscopic system, time independent statistical averages are often introduced.

Definitions

The thermodynamic state of a system is usually defined by a small set of parameters, for example, the temperature, T, the pressure, P, and the number of particles, N. Other thermodynamic properties may be derived from the equations of state and other fundamental thermodynamic equations.

The mechanical or microscopic state of a system is defined by the atomic positions, q, and momenta, p; these can also be considered as coordinates in a multidimensional space called phase space. For a system of N particles, this space has 6N dimensions. A single point in phase space, denoted by G, describes the state of the system. An ensemble is a collection of points in phase space satisfying the conditions of a particular thermodynamic state. A molecular dynamics simulations generates a sequence of points in phase space as a function of time; these points belong to the same ensemble, and they correspond to the different conformations of the system and their respective momenta. Several different ensembles are described below.

An ensemble is a collection of all possible systems which have different microscopic states but have an identical macroscopic or thermodynamic state.

There exist different ensembles with different characteristics.

Microcanonical ensemble (NVE) : The thermodynamic state characterized by a fixed number of atoms, N, a fixed volume, V, and a fixed energy, E. This corresponds to an isolated system.

Canonical Ensemble (NVT): This is a collection of all systems whose thermodynamic state is characterized by a fixed number of atoms, N, a fixed volume, V, and a fixed temperature, T.

Isobaric-Isothermal Ensemble (NPT): This ensemble is characterized by a fixed number of atoms, N, a fixed pressure, P, and a fixed temperature, T.

Grand canonical Ensemble (mVT): The thermodynamic state for this ensemble is characterized by a fixed chemical potential, m, a fixed volume, V, and a fixed temperature, T.

Calculating Averages from a Molecular Dynamics Simulation

An experiment is usually made on a macroscopic sample that contains an extremely large number of atoms or molecules sampling an enormous number of conformations. In statistical mechanics, averages corresponding to experimental observables are defined in terms of ensemble averages; one justification for this is that there has been good agreement with experiment. An ensemble average is average taken over a large number of replicas of the system considered simultaneously.

6.1.1 Time Averages, Ensemble Averages and Some Historical Background

Suppose we wish to determine experimentally the value of a property of a system such as the pressure or the heat capacity. In general, such properties will depend upon the positions and momenta of the N particles that comprise the system. The instantaneous value of the property A can thus be written as $A(\mathbf{p}^N(t)\ \mathbf{r}^N(t))$, where $\mathbf{p}^N(t)$ and $\mathbf{r}^N(t)$ represent the N momenta and positions respectively at time t (i.e. $A(\mathbf{p}^N(t),\mathbf{r}^N(t)) \equiv A(p_{1x},p_{1y},p_{1z},p_{2x},\ldots,x_1,y_1,z_1,x_2,\ldots,t)$ where p_{1x} is the momentum of particle 1 in the x direction and x_1 is its x coordinate). Over time, the instantaneous value of the property A fluctuates as a result of interactions between the particles. The value that we measure experimentally is an average of A over the time of the measurement and is therefore known as a *time average*. As the time over which the measurement is made increases to infinity, so the value of the following integral approaches the 'true' average value of the property:

$$A_{\text{ave}} = \lim_{\tau\to\infty}\frac{1}{\tau}\int_{t=0}^{\tau} A(\mathbf{p}^N(t),\mathbf{r}^N(t))\,dt \tag{6.1}$$

To calculate average values of the properties of the system, it would therefore appear to be necessary to simulate the dynamic behaviour of the system (i.e. to determine values of $A(\mathbf{p}^N(t),\mathbf{r}^N(t))$, based upon a model of the intra- and intermolecular interactions present).

To calculate average values of the properties of the system, it would therefore appear to be necessary to simulate the dynamic behaviour of the system (i.e. to determine values of $A(\mathbf{p}^N(t),\mathbf{r}^N(t))$, based upon a model of the intra- and intermolecular interactions present). In principle, this is relatively straightforward to do. For any arrangement of the atoms in the system, the force acting on each atom due to interactions with other atoms can be calculated by differentiating the energy function. From the force on each atom it is possible to determine its acceleration via Newton's second law. Integration of the equations of motion should then yield a trajectory that describes how the positions, velocities and accelerations of the particles vary with time, and from which the average values of properties can be determined using the numerical equivalent of Equation (6.1). The difficulty is that for 'macroscopic' numbers of atoms or molecules (of the order of 10^{23}) it is not even feasible to determine an initial configuration of the system, let alone integrate the equations of motion and calculate a trajectory. Recognising this problem, Boltzmann and Gibbs developed statistical mechanics, in which a single system evolving in time is replaced by a large number of replications of the system that are considered simultaneously. The time average is then replaced by an *ensemble average*:

$$\langle A\rangle = \iint d\mathbf{p}^N\,d\mathbf{r}^N\,A(\mathbf{p}^N,\mathbf{r}^N)\rho(\mathbf{p}^N,\mathbf{r}^N) \tag{6.2}$$

The angle brackets $\langle\ \rangle$ indicate an ensemble average, or *expectation value*; that is, the average value of the property A over all replications of the ensemble generated by the simulation. Equation (6.2) is written as a double integral for convenience but in fact there should be $6N$ integral signs on the integral for the $6N$ positions and momenta of all the particles. $\rho(\mathbf{p}^N\mathbf{r}^N)$ is the *probability density* of the ensemble; that is, the probability of finding a configuration with momenta $\mathbf{p}^N$ and positions $\mathbf{r}^N$. The ensemble average of the property A is then determined by integrating over all possible configurations of the system. In accordance with the *ergodic hypothesis*, which is one of the fundamental axioms of statistical mechanics, the ensem-

SCHOOL OF BIO AND CHEMICAL ENGINEERING
DEPARTMENT OF BIOINFORMATICS

UNIT – 5- SBI1208 – Cheminformatics

Drug

A **drug** is any substance that causes a change in an organism's physiology or psychology when consumed. Drugs are typically distinguished from food and substances that provide nutritional support. Consumption of drugs can be via inhalation, injection, smoking, ingestion, absorption via a patch on the skin, suppository, or dissolution under the tongue.

In pharmacology, a drug is a chemical substance, typically of known structure, which, when administered to a living organism, produces a biological effect. A pharmaceutical drug, also called a medication or medicine, is a chemical substance used to treat, cure, prevent, or diagnose a disease or to promote well-being.[3] Traditionally drugs were obtained through extraction from medicinal plants, but more recently also by organic synthesis. Pharmaceutical drugs may be used for a limited duration, or on a regular basis for chronic disorders.

Pharmaceutical drugs are often classified into drug classes—groups of related drugs that have similar chemical structures, the same mechanism of action (binding to the same biological target), a related mode of action, and that are used to treat the same disease. The Anatomical Therapeutic Chemical Classification System (ATC), the most widely used drug classification system, assigns drugs a unique ATC code, which is an alphanumeric code that assigns it to specific drug classes within the ATC system. Another major classification system is the Biopharmaceutics Classification System. This classifies drugs according to their solubility and permeability or absorption properties

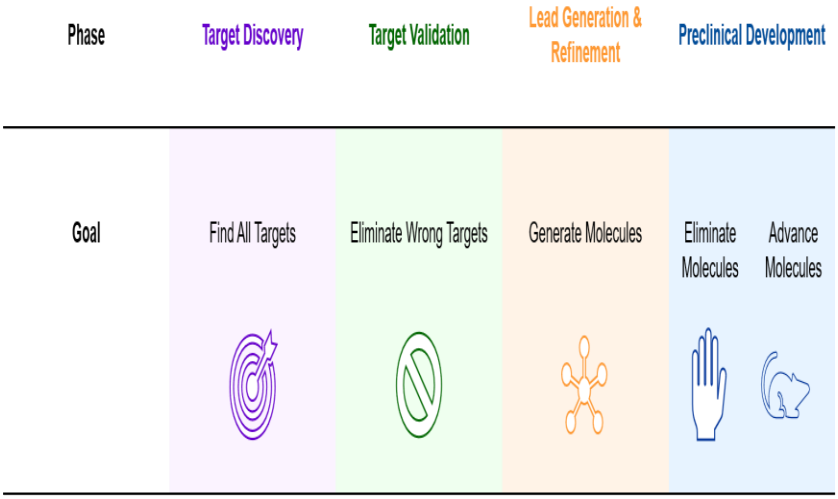
Phases Stages

There are five critical steps in the U.S. drug development process, including many phases and stages within each of them. We will discuss these different phases and stages to develop an in-depth understanding of the entire process. The five steps are –

- Step 1: Discovery and Development
- Step 2: Preclinical Research
- Step 3: Clinical Development
- Step 4: FDA Review

- Step 5: FDA Post-market Safety Monitoring.

Step 1: Discovery & Development



Drug discovery is how new medications are discovered. Historically, drugs were mostly found by identifying active ingredients from traditional medicines or purely by chance. Afterward, classical pharmacology was used to investigate chemical libraries including small molecules, natural products, or plant extracts, and find those with therapeutic effects. Since human DNA was sequenced, reverse pharmacology has found remedies to existing diseases through testing.
Disease processes, molecular compound tests, existing treatments with unanticipated effects, and new technologies spur drug discovery through the cycle below.
Today drug discovery involves screening hits, medicinal chemistry, and optimization of hits to reduce potential drug side effects (increasing affinity and selectivity). Efficacy or potency, metabolic stability (half-life), and oral bioavailability are also improved in this step of the drug development process.

Target Identification & Validation

Target identification finds a gene or protein (therapeutic agent) that plays a significant role in disease. When identified, therapeutic characteristics are recorded. Targets are efficacious, safe, usable as drugs, and capable of meeting clinical and commercial requirements. Researchers use disease association, bioactive molecules, cell-based models, protein interactions, signaling pathways analysis, and functional analysis of genes to validate targets, or in vitro genetic manipulation, antibodies, and chemical genomics. The Sanger Whole Genome CRISPER library and Duolink PLA are excellent sources for drug discovery targets.

Hit Discovery Process

Following target validation, compound screening assays are developed.

Assay Development & Screening

Assays are test systems that evaluate the effects of the new drug candidate at the cellular, molecular, and biochemical levels.

High Throughput Screening

High Throughput Screening (HTS) uses robotics, data processing/control software, liquid handling devices, and sensitive detectors to rapidly conduct millions of pharmacological, chemical, and genetic tests, eliminating hours of painstaking testing by scientists. HTS identifies active compounds, genes, or antibodies that affect human molecules.

Hit to Lead

In the Hit to Lead (H2L) process, small molecule hits from an HTS are evaluated and optimized in a limited way into lead compounds. These compounds then move on to the lead optimization process.

Lead Optimization

In the lead optimization (LO) process, the lead compounds discovered in the H2L process are synthesized and modified to improve potency and reduce side effects. Lead optimization conducts experimental testing using animal efficacy models and ADMET tools, designing the drug candidate.

Active Pharmaceutical Ingredients

Active pharmaceutical ingredients (APIs) are biologically active ingredients in a drug candidate that produce effects. All drugs are made up of the API or APIs and excipients. (Excipients are inactive substances that deliver the drug into the human system.). High Potency Active Pharmaceutical Ingredients (HP APIs) are molecules that are effective at much smaller dosage levels than standard APIs. They are classified based on toxicity, pharmacological potency, and occupational exposure limits (OELs), and used in complex drug development involving more than ten steps.

The drug discovery process ends when one lead compound is found for a drug candidate, and the process of drug development starts.

Step 2: Preclinical Research

Once a lead compound is found, drug development begins with preclinical research to determine the efficacy and safety of the drug. Researchers determine the following about the drug:

- Absorption, distribution, metabolization, and excretion information
- Potential benefits and mechanisms of action
- Best dosage, and administration route
- Side effects/adverse events
- Effects on gender, race, or ethnicity groups
- Interaction with other treatments
- Effectiveness compared to similar drugs

Preclinical trials test the new drug on non-human subjects for efficacy, toxicity, and pharmacokinetic (PK) information. These trials are conducted by scientists in vitro and in vivo with unrestricted dosages.

Absorption, Distribution, Disposition, Metabolism, & Excretion

Absorption, Distribution, Disposition, Metabolism, & Excretion (ADME) is a PK process of measuring the ways the new drug affects the body. ADME involves mathematical descriptions of each effect.

Proof of Principle / Proof of Concept

Proof of Principle (PoP) are studies that are successful in preclinical trials and early safety testing. Proof of Concept (PoC) terminology is used almost interchangeably with PoP in drug discovery and development projects. Successful PoP/PoC studies lead to program advancement to the Phase II studies of dosages.

In Vivo, In Vitro & Ex VivoAssays

These three types of studies are conducted on the whole, living organisms or cells, including animals and humans; or using non-living organisms or tissue extract. In vivo, preclinical research

examples are the development of new drugs using mice, rat, and dog models. In vitro is research conducted in a laboratory. Ex vivo uses animal cells or tissues from a non-living animal. Examples of ex vivo research assays are finding effective cancer treatment agents; measurements of tissue properties (physical, thermal, electrical, and optical); and realistic modeling for new surgical procedures. In an ex vivo assay, a cell is always used as the basis for small explant cultures that provide a dynamic, controlled, and sterile environment.

In SilicoAssays

In silico assays are test systems or biological experiments performed on a computer or via computer simulation. These are expected to become increasingly popular with the ongoing improvements in computational power, and behavioral understanding of molecular dynamics and cell biology.

Drug Delivery

New drug delivery methods include oral, topical, membrane, intravenous, and inhalation. Drug delivery systems are used for targeted delivery or controlled release of new drugs. Physiological barriers in animal or human bodies may prevent drugs from reaching the targeted area or releasing when they should. The goal is to prevent the drug from interacting with healthy tissues while still being effective.

- **Oral:** Oral delivery of medications is reliable, cost-effective, and convenient for patients. Oral drug delivery may not monitor precise dosages to the desired area but is ideal for prophylactic vaccinations and nutritional regimens. Delayed action, stomach enzyme destruction, absorption inconsistencies, or patients with gastrointestinal issues or upset can occur, and patients must be conscious during administration.
- **Topical:** Topical drug delivery involves ointments, creams, lotions, or transdermal patches that deliver a drug by absorption into the body. Topical delivery is more useful for patient skin or muscular conditions — it is preferred by patients due to non-invasive delivery and their ability to self-administer the medicine.
- **Parenteral (IM, SC or LP Membrane):** Parenteral drug delivery utilizes bodily membranes, including intramuscular (IM), intraperitoneal (IP), or subcutaneous or (SC). It is often used for unconscious patients and avoids epithelial barriers that are difficult for drugs to cross.
- **Parenteral (Intravenous):** Intravenous injection is one of the fastest drug delivery absorption methods. IV injection ensures entire doses of drugs enter the bloodstream, and it is more effective than IM, SC, or LP membrane methods.
- **Parenteral (Inhalation):** Inhalation drug delivery gets the drug rapidly absorbed into the mucosal lungs, nasal passages, throat, or mouth. Problems with inhalation delivery include difficulty delivering the optimum dosage due to small mucosal surface areas and patient discomfort. Pulmonary inhalation drug delivery uses fine drug powders or macromolecular drug solutions. Lung fluids resemble blood, so they can absorb small particles easily and deliver them into the bloodstream.

Formulation Optimization & Improving Bioavailability

Formulation optimization is ongoing throughout pre-clinical and clinical stages. It ensures drugs are delivered to the proper place at the right time and in the right concentration. Optimization may include overcoming solub

Step 3: Clinical Development

Once preclinical research is complete, researchers move on to clinical drug development, including clinical trials and volunteer studies to finetune the drug for human use.

Complexity of Study Design, Associated Cost & Implementation Issues

The complexity of clinical trial design and its associated costs and implementation issues may affect trials carried out during this phase. Trials must be safe and efficacious and be completed under the drug development budget, using a methodology to ensure the drug works as well as possible for its intended purpose. This rigorous process must be set up correctly and enroll many volunteers to be effective.

Clinical Trials– Dose Escalation, Single Ascending & Multiple Dose Studies

Proper dosing determines medication effectiveness, and clinical trial examine dose escalation, single ascending, and multiple dose studies to determine the best patient dosage.

Phase I – Healthy Volunteer Study

This phase is the first time the drug is tested on humans; less than 100 volunteers will help researchers assess the safety and pharmacokinetics, absorption, metabolic, and elimination effects on the body, as well as any side effects for safe dosage ranges.

Phase II and Phase III – Studies in Patient Population

Phase II assesses drug safety and efficacy in an additional 100-500 patients, who may receive a placebo or standard drug previously used as treatment. Analysis of optimal dose strength helps create schedules while adverse events and risks are recorded. Phase III enrolls 1,000-5,000 patients, enabling medication labeling and instructions for proper drug use. Phase III trials require extensive collaboration, organization, and Independent Ethics Committee (IEC) or Institutional Review Board (IRB) coordination and regulation in anticipation of full-scale production following drug approval.

Biological Samples Collection, Storage & Shipment

During clinical trials, biological samples are collected, stored, and shipped from testing sites according to global standards and regulations. Transport containers of biological samples may include dry ice packs or other temperature stabilizing methods. Different requirements apply to different types of biological samples.

Pharmacodynamic (PD) Biomarkers

PD biomarkers are molecular indicators of the drug's effects on the target human area, and link drug regimen and biological responses. This data can help select rational combinations of targeted agents and optimize drug regimens and schedules. Rationality and hypothesis-testing power are increased through the use of PD endpoints in human trials.

Pharmacokinetic Analysis

Pharmacokinetic analysis is an experimental trial that determines the theory of how a new drug behaves in the human body. The volume of distribution, clearance, and terminal half-life are defined through compartmental modeling.

Bioanalytical Method Development and Validation

Bioanalytical methods detect analytes and metabolites such as drug or biomarkers in biological or human samples to determine drug efficacy and safety. The complete bioanalytical assay consists of sample collection, clean-up, analysis, and detection.

Drug (Analyte) & Metabolite Stability in Biological Samples

Stability is important in determining human drug efficacy, and biological samples are required. Drug and drug metabolites are susceptible to degradation, which can lower drug concentration over the life of the drug.

Blood, Plasma, Urine & Feces Sample Analysis for Drug and Metabolites

Biological samples used in clinical trials include blood, plasma, urine, and feces to determine and analyze various properties and effects of the drug and its metabolites on humans.

Patient Protection – GCP, HIPAA, & Adverse Event Reporting

Human patients must always be protected during clinical trials, and Good Clinical Practices (GCP), the Health Insurance Portability and Accountability Act (HIPAA), and adverse event reporting to IEC/IRB regulates and ensures their safety.

Step 4: FDA Review

Once the new drug has been formulated for its best efficacy and safety, and the results from clinical trials are available, it's advanced forward for wholistic FDA review. At this time, the FDA reviews and approves, or does not approve, the drug application submitted by the drug development company.

Regulatory Approval Timeline

The new drug regulatory approval timeline may be standard, fast track, breakthrough, accelerated approval, or priority review depending on its applications and necessity for patients. If standard or priority review is required, the approval timeline may be up to an year. Fast track, breakthrough, or accelerated approvals may occur sooner.

IND Application

IND applications are submitted to the FDA before starting clinical trials. If clinical trials are ready to be conducted, and the FDA has not responded negatively about the drug, developers may start the trials.

NDA / ANDA / BLA Applications

An NDA abbreviated new drug application (ANDA), or BLA is submitted to the FDA after clinical trials demonstrate drug safety and efficacy. The FDA reviews study data and decides whether to grant approval or not. Additional research or an expert advisory panel may be required before a final decision is made.

Orphan Drug

An orphan drug is intended to treat disease so rare that financial sponsors are unwilling to develop it under standard marketing conditions. These drugs may not be approved quickly or at all.

Accelerated Approval

New drugs may be granted accelerated approval if there is strong evidence of positive impact on a surrogate endpoint instead of evidence of impact on actual clinical benefits the drug provides. Expedition of approval means the medication can help treat severe or life-threatening conditions.

Reasons for Drug Failure

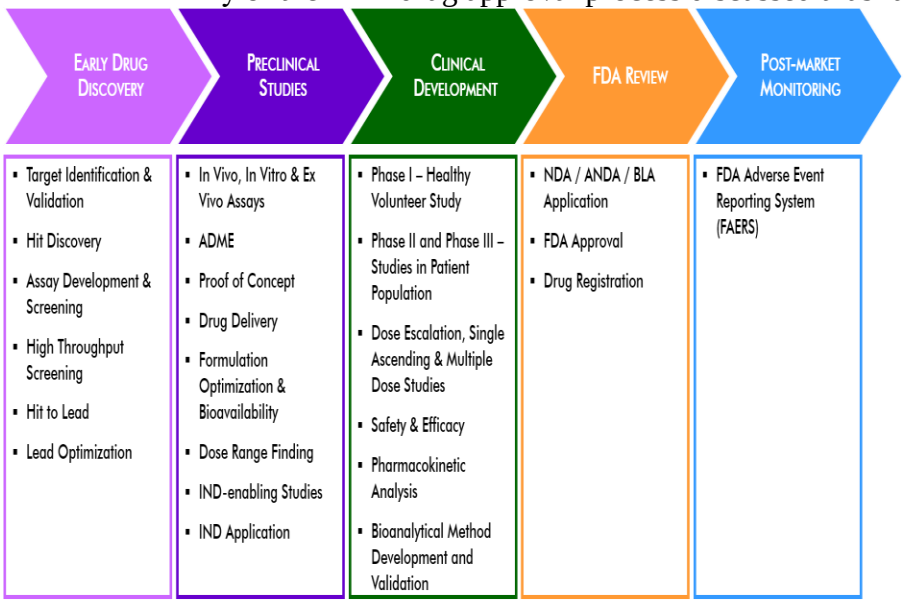
New drug applications may fail for a variety of reasons, including toxicity, efficacy, PH properties, bioavailability, or inadequate drug performance.

- **Toxicity:** If the toxicity of a new drug is too high in human or animal patients, the drug may be rejected due to safety concerns about its use following manufacture.
- **Efficacy:** If a new drug's efficacy is not high enough or evidence is inconclusive, the FDA may reject it.
- **PK Properties or Bioavailability:** PK properties or poor bioavailability due to low aqueous solubility, or high first-pass metabolism, may also cause a drug to fail FDA review. PK causes of drug failure include inadequate action duration and unanticipated human drug interactions.
- **Inadequate Drug Performance:** If the new drug performs the desired function, but only at a shallow level, the FDA may reject the application in favor of a formulation that performs better.

Step 5: Post-market Monitoring

Following drug approval and manufacturing, the FDA requires drug companies to monitor the safety of its drug using the FDA Adverse Event Reporting System (FAERS) database. FAERS helps FDA implement its post-marketing safety surveillance program. Through this program, manufacturers, health professionals, and consumers report problems with approved drugs.

Here's a summary of the FDA drug approval process discussed thus far.



Drug metabolism

Drug metabolism is the metabolic breakdown of drugs by living organisms, usually through specialized enzymatic systems. More generally, **xenobiotic metabolism** (from the Greek xenos "stranger" and biotic "related to living beings") is the set of metabolic pathways that modify the chemical structure of xenobiotics, which are compounds foreign to an organism's normal biochemistry, such as any drug or poison. These pathways are a form of biotransformation present in all major groups of organisms and are considered to be of ancient origin. These reactions often act to detoxify poisonous compounds (although in some cases the intermediates in xenobiotic metabolism can themselves cause toxic effects). The study of drug metabolism is called pharmacokinetics.

The metabolism of pharmaceutical drugs is an important aspect of pharmacology and medicine. For example, the rate of metabolism determines the duration and intensity of a drug's pharmacologic action. Drug metabolism also affects multidrug resistance in infectious diseases and in chemotherapy for cancer, and the actions of some drugs as substrates or inhibitors of enzymes involved in xenobiotic metabolism are a common reason for hazardous drug interactions. These pathways are also important in environmental science, with the xenobiotic metabolism of microorganisms determining whether a pollutant will be broken down during bioremediation, or persist in the environment. The enzymes of xenobiotic metabolism, particularly the glutathione S-transferases are also important in agriculture, since they may produce resistance to pesticides and herbicides.

Drug metabolism is divided into three phases. In phase I, enzymes such as cytochrome P450 oxidases introduce reactive or polar groups into xenobiotics. These modified compounds are then conjugated to polar compounds in phase II reactions. These reactions are catalysed by transferase enzymes such as glutathione S-transferases. Finally, in phase III, the conjugated xenobiotics may be further processed, before being recognised by efflux transporters and pumped out of cells. Drug metabolism often converts lipophilic compounds into hydrophilic products that are more readily excreted.

Permeability barriers and detoxification

The exact compounds an organism is exposed to will be largely unpredictable, and may differ widely over time; these are major characteristics of xenobiotic toxic stress.[1] The major challenge faced by xenobiotic detoxification systems is that they must be able to remove the almost-limitless number of xenobiotic compounds from the complex mixture of chemicals involved in normal metabolism. The solution that has evolved to address this problem is an elegant combination of physical barriers and low-specificity enzymatic systems.

All organisms use cell membranes as hydrophobic permeability barriers to control access to their internal environment. Polar compounds cannot diffuse across these cell membranes, and the uptake of useful molecules is mediated through transport proteins that specifically select substrates from the extracellular mixture. This selective uptake means that most hydrophilic molecules cannot enter cells, since they are not recognised by any specific transporters.[2] In contrast, the diffusion of hydrophobic compounds across these barriers cannot be controlled, and organisms, therefore, cannot exclude lipid-soluble xenobiotics using membrane barriers.

However, the existence of a permeability barrier means that organisms were able to evolve detoxification systems that exploit the hydrophobicity common to membrane-permeable xenobiotics. These systems therefore solve the specificity problem by possessing such broad substrate specificities that they metabolise almost any non-polar compound.[1] Useful metabolites are excluded since they are polar, and in general contain one or more charged groups.

The detoxification of the reactive by-products of normal metabolism cannot be achieved by the systems outlined above, because these species are derived from normal cellular constituents and usually share their polar characteristics. However, since these compounds are few in number, specific enzymes can recognize and remove them. Examples of these specific detoxification systems are the glyoxalase system, which removes the reactive aldehyde methylglyoxal,[3] and the various antioxidant systems that eliminate reactive oxygen species

Phases of detoxification

The metabolism of xenobiotics is often divided into three phases:- modification, conjugation, and excretion. These reactions act in concert to detoxify xenobiotics and remove them from cells.

Phase I – modification

In phase I, a variety of enzymes act to introduce reactive and polar groups into their substrates. One of the most common modifications is hydroxylation catalysed by the cytochrome P-450-dependent mixed-function oxidase system. These enzyme complexes act to incorporate an

atom of oxygen into nonactivated hydrocarbons, which can result in either the introduction of hydroxyl groups or N-, O- and S-dealkylation of substrates.[5] The reaction mechanism of the P-450 oxidases proceeds through the reduction of cytochrome-bound oxygen and the generation of a highly-reactive oxyferryl species, according to the following scheme:[6]

$$O_2 + NADPH + H^+ + RH \rightarrow NADP^+ + H_2O + ROH$$

Phase I reactions (also termed nonsynthetic reactions) may occur by oxidation, reduction, hydrolysis, cyclization, decyclization, and addition of oxygen or removal of hydrogen, carried out by mixed function oxidases, often in the liver. These oxidative reactions typically involve a cytochrome P450 monooxygenase (often abbreviated CYP), NADPH and oxygen. The classes of pharmaceutical drugs that utilize this method for their metabolism include phenothiazines, paracetamol, and steroids. If the metabolites of phase I reactions are sufficiently polar, they may be readily excreted at this point. However, many phase I products are not eliminated rapidly and undergo a subsequent reaction in which an endogenous substrate combines with the newly incorporated functional group to form a highly polar conjugate.

A common Phase I oxidation involves conversion of a C-H bond to a C-OH. This reaction sometimes converts a pharmacologically inactive compound (a prodrug) to a pharmacologically active one. By the same token, Phase I can turn a nontoxic molecule into a poisonous one (toxification). Simple hydrolysis in the stomach is normally an innocuous reaction, however there are exceptions. For example, phase I metabolism converts acetonitrile to $HOCH_2CN$, which rapidly dissociates into formaldehyde and hydrogen cyanide.[7]

Phase I metabolism of drug candidates can be simulated in the laboratory using non-enzyme catalysts.[8] This example of a biomimetic reaction tends to give products that often contains the Phase I metabolites. As an example, the major metabolite of the pharmaceutical trimebutine, desmethyltrimebutine (nor-trimebutine), can be efficiently produced by in vitro oxidation of the commercially available drug. Hydroxylation of an N-methyl group leads to expulsion of a molecule of formaldehyde, while oxidation of the O-methyl groups takes place to a lesser extent.

Oxidation

- Cytochrome P450 monooxygenase system
- Flavin-containing monooxygenase system
- Alcohol dehydrogenase and aldehyde dehydrogenase
- Monoamine oxidase
- Co-oxidation by peroxidases

Reduction

- NADPH-cytochrome P450 reductase

Cytochrome P450 reductase, also known as NADPH:ferrihemoprotein oxidoreductase, NADPH:hemoprotein oxidoreductase, NADPH:P450 oxidoreductase, P450 reductase, POR, CPR, CYPOR, is a membrane-bound enzyme required for electron transfer to cytochrome P450 in the microsome of the eukaryotic cell from a FAD- and FMN-containing enzyme NADPH:cytochrome P450 reductase The general scheme of electron flow in the POR/P450 system is: NADPH → FAD → FMN → P450 → O_2

- Reduced (ferrous) cytochrome P450

During reduction reactions, a chemical can enter *futile cycling*, in which it gains a free-radical electron, then promptly loses it to oxygen (to form a superoxide anion).

Hydrolysis

- Esterases and amidase
- Epoxide hydrolase

Phase II – conjugation

In subsequent phase II reactions, these activated xenobiotic metabolites are conjugated with charged species such as glutathione (GSH), sulfate, glycine, or glucuronic acid. Sites on drugs where conjugation reactions occur include carboxy (-COOH), hydroxy (-OH), amino (NH_2), and thiol (-SH) groups. Products of conjugation reactions have increased molecular weight and tend to be less active than their substrates, unlike Phase I reactions which often produce active metabolites. The addition of large anionic groups (such as GSH) detoxifies reactive electrophiles and produces more polar metabolites that cannot diffuse across membranes, and may, therefore, be actively transported.

These reactions are catalysed by a large group of broad-specificity transferases, which in combination can metabolise almost any hydrophobic compound that contains nucleophilic or electrophilic groups. One of the most important classes of this group is that of the glutathione S-transferases (GSTs).

Mechanism	**Involved enzyme**	**Co-factor**	**Location**
methylation	methyltransferase	S-adenosyl-L-methionine	liver, kidney, lung, CNS
sulphation	sulfotransferases	3'-phosphoadenosine-5'-phosphosulfate	liver, kidney, intestine
acetylation	• N-acetyltransferases • bile acid-CoA:amino acid N-acyltransferases	acetyl coenzyme A	liver, lung, spleen, gastric mucosa, RBCs, lymphocytes
glucuronidation	UDP-glucuronosyltransferases	UDP-glucuronic acid	liver, kidney, intestine, lung, skin, prostate, brain
glutathione conjugation	glutathione S-transferases	glutathione	liver, kidney
glycine conjugation	Two step process: 1. XM-ligase (forms a xenobiotic acyl-CoA) 2. Glycine N-acyltransferase (forms	glycine	liver, kidney

	the glycine conjugate)		

Phase III – further modification and excretion

After phase II reactions, the xenobiotic conjugates may be further metabolized. A common example is the processing of glutathione conjugates to acetylcysteine (mercapturic acid) conjugates.[11] Here, the γ-glutamate and glycine residues in the glutathione molecule are removed by Gamma-glutamyl transpeptidase and dipeptidases. In the final step, the cysteine residue in the conjugate is acetylated.

Conjugates and their metabolites can be excreted from cells in phase III of their metabolism, with the anionic groups acting as affinity tags for a variety of membrane transporters of the multidrug resistance protein (MRP) family.[12] These proteins are members of the family of ATP-binding cassette transporters and can catalyse the ATP-dependent transport of a huge variety of hydrophobic anions,[13] and thus act to remove phase II products to the extracellular medium, where they may be further metabolized or excreted

Endogenous toxins

The detoxification of endogenous reactive metabolites such as peroxides and reactive aldehydes often cannot be achieved by the system described above. This is the result of these species' being derived from normal cellular constituents and usually sharing their polar characteristics. However, since these compounds are few in number, it is possible for enzymatic systems to utilize specific molecular recognition to recognize and remove them. The similarity of these molecules to useful metabolites therefore means that different detoxification enzymes are usually required for the metabolism of each group of endogenous toxins. Examples of these specific detoxification systems are the glyoxalase system, which acts to dispose of the reactive aldehyde methylglyoxal, and the various antioxidant systems that remove reactive oxygen species.

Sites

Quantitatively, the smooth endoplasmic reticulum of the liver cell is the principal organ of drug metabolism, although every biological tissue has some ability to metabolize drugs. Factors responsible for the liver's contribution to drug metabolism include that it is a large organ, that it is the first organ perfused by chemicals absorbed in the gut, and that there are very high concentrations of most drug-metabolizing enzyme systems relative to other organs. If a drug is taken into the GI tract, where it enters hepatic circulation through the portal vein, it becomes well-metabolized and is said to show the *first pass effect*.

Other sites of drug metabolism include epithelial cells of the gastrointestinal tract, lungs, kidneys, and the skin. These sites are usually responsible for localized toxicity reactions.

Factors that affect drug metabolism

The duration and intensity of pharmacological action of most lipophilic drugs are determined by the rate they are metabolized to inactive products. The Cytochrome P450 monooxygenase system is the most important pathway in this regard. In general, anything that *increases* the rate of metabolism (*e.g.*, enzyme induction) of a pharmacologically active metabolite will *decrease* the duration and intensity of the drug action. The opposite is also true (*e.g.*, enzyme inhibition). However, in cases where an enzyme is responsible for metabolizing a pro-drug into a drug, enzyme induction can speed up this conversion and increase drug levels, potentially causing toxicity.

Various *physiological* and *pathological* factors can also affect drug metabolism. Physiological factors that can influence drug metabolism include age, individual variation (*e.g.*, pharmacogenetics), enterohepatic circulation, nutrition, intestinal flora, or sex differences.

In general, drugs are metabolized more slowly in fetal, neonatal and elderly humans and animals than in adults.

Genetic variation (polymorphism) accounts for some of the variability in the effect of drugs. With N-acetyltransferases (involved in *Phase II* reactions), individual variation creates a group of people who acetylate slowly (*slow acetylators*) and those who acetylate quickly, split roughly 50:50 in the population of Canada. This variation may have dramatic consequences, as the slow acetylators are more prone to dose-dependent toxicity.

Cytochrome P450 monooxygenase system enzymes can also vary across individuals, with deficiencies occurring in 1 – 30% of people, depending on their ethnic background.

Dose, frequency, route of administration, tissue distribution and protein binding of the drug affect its metabolism.

Pathological factors can also influence drug metabolism, including liver, kidney, or heart diseases.

In silico modelling and simulation methods allow drug metabolism to be predicted in virtual patient populations prior to performing clinical studies in human subjects.[15] This can be used to identify individuals most at risk from adverse reaction.

Pharmacokinetics and Pharmacodynamics

Pharmacokinetics And Dynamics:

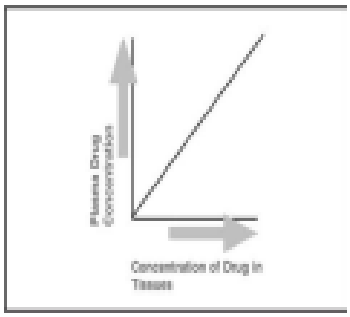
- The purpose of studying pharmacokinetics and pharmacodynamics is to understand the drug action, therapy, design, development and evaluation
- Pharmacokinetics is what the Body Does To The Drug like how the drug is Absorbed, Distributed, Metabolized, and Excreted by the body – Drug disposition.
- Pharmacodynamics is what the Drug Does To The Body which may be the therapeutic effects or the adverse side effects - Drug action.

Effect Of A Drug

- A drug's effect is often related to its concentration at the site of action, so it would be useful to monitor this concentration. Receptor sites of drugs are generally inaccessible to our observations or are widely distributed in the body, and therefore direct measurement of drug concentrations at these sites is not practical.
- So we can measure drug concentration in the blood or plasma, urine, saliva, and other easily sampled fluids because of the Kinetic homogeneity *principle*.

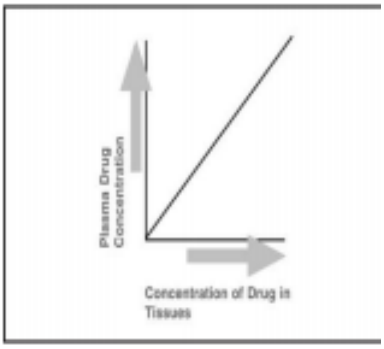
Relationship of plasma to tissue drug concentrations

- Kinetic Homogeneity describes the predictable relationship between plasma drug concentration and concentration at the receptor site where a given drug produces its therapeutic effect
- Changes in the plasma drug concentration reflect changes in drug concentrations at the receptor site, as well as in other tissues.
- As the concentration of drug in plasma increases, the concentration of drug in most tissues will increase proportionally

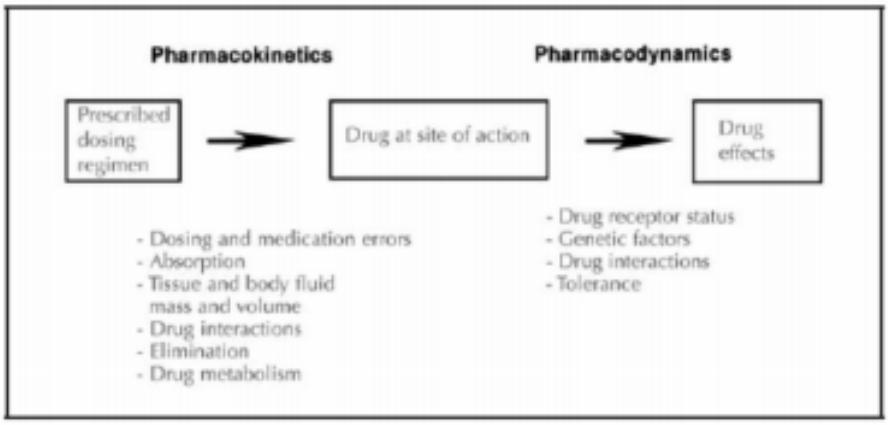


Relationship of plasma to tissue drug concentrations

- Kinetic Homogeneity describes the predictable relationship between plasma drug concentration and concentration at the receptor site where a given drug produces its therapeutic effect
- Changes in the plasma drug concentration reflect changes in drug concentrations at the receptor site, as well as in other tissues.
- As the concentration of drug in plasma increases, the concentration of drug in most tissues will increase proportionally

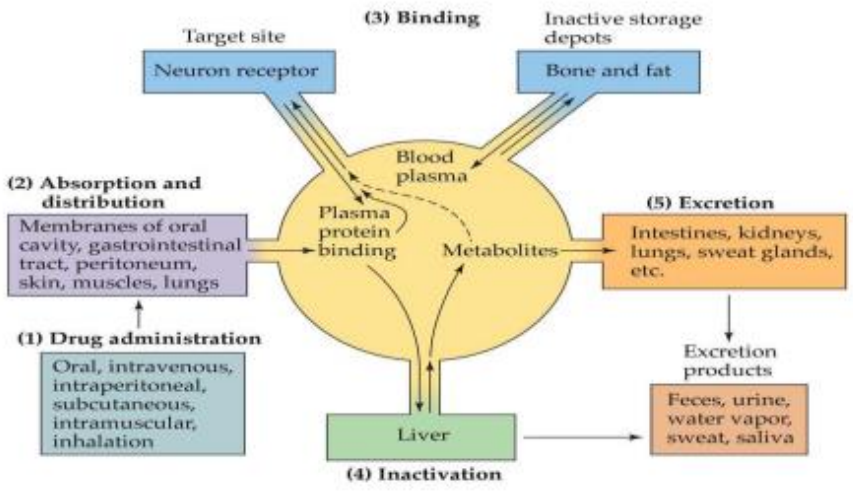


Relationship of pharmacokinetics & pharmacodynamics

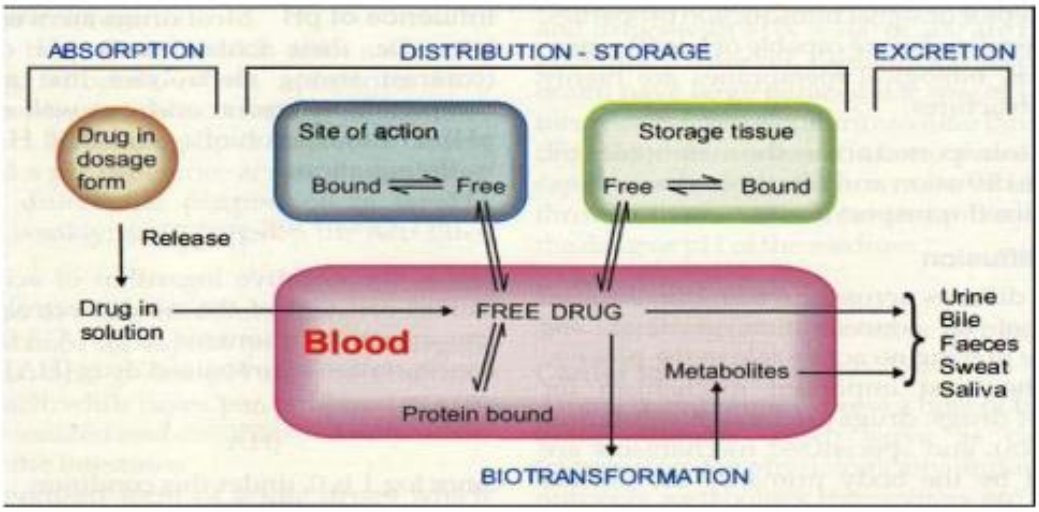


Pharmacokinetics

- Pharmacokinetics is the study of Movement of drugs in the body and it describes the drug absorption, distribution within body, and drug elimination over time.
- It involves Four Processes
  1. Absorption
  2. Drug distribution
  3. Metabolism
  4. Drug elimination



Schematic depiction of pharmacokinetic processes



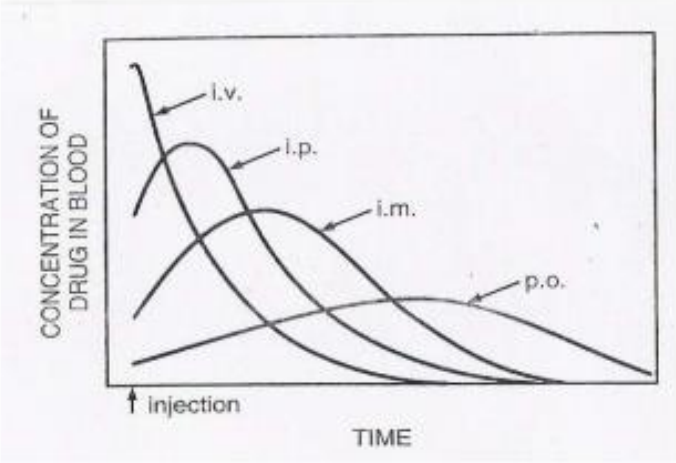
1.Absorption

It is the process of entry of drug from site of administration into systemic circulation.

The bioavailability of the drug depends on the extent of the absorption.

- Bioavailability is the percentage of drug that reaches the systemic circulation in an unchanged form and becomes available for biological effect following administration by any route.
- Bioequivalence occurs when two formulations of the same compound have the same bioavailability and the same rate of absorption.

Route of Administration Determines Bioavailability



Distribution of Drugs

Distribution is the movement of drug from the central compartment (blood) to peripheral compartments. Here the concentration gradient is being the driving force for the movement from plasma to tissues.

It depends on,

Ionization,

Molecular size,

Binding to plasma proteins,

Differences in regional blood flow

Presence of tissue-specific transporters

Volume of distribution

It is defined as the volume of fluid required to contain the total amount of drug Q in the body at the same concentration as that present in the plasma C_p

$$V_d = Q/C_p$$

Importance of V_d:

1.It helps in estimating the total amount of drug in body at any time.

Amount of drug = V_d x plasma concentration of drug at certain time.

2.Vd is important to determine the loading dose

Loading dose = V_d x desired concentration

Metabolism (Biotransformation)

- Biotransformation means chemical alteration of the drug in the body.
- It is needed to render nonpolar (lipid-soluble) compounds polar (lipid insoluble) so that they are not reabsorbed in the renal tubules and are excreted.
- The primary site for drug metabolism is liver; others are-kidney, intestine, lungs and plasma.

Phases of biotransformation:

Phase I (Non-synthetic) reactions - A functional group is generated or exposed-metabolite may be active or inactive.

Phase II (Synthetic) reactions – Mostly a conjugation reaction - Metabolite is mostly inactive (except few drugs).

Phase I (Non-synthetic) Reactions

Introduction or unmasking of functional group by oxidation, reduction hydrolysis, Cyclization, Decyclization

These reactions may result in

1. Drug inactivation (most of drugs)
2. Conversion of inactive drug into active metabolite (cortisone→ cortisol)
3. Conversion of active drug into active metabolite (phenacetin→ paracetamol)
4. Conversion to toxic metabolite (methanol → formaldehyde)

Phase II (Synthetic) reactions

- Functional group or metabolite formed by phase I is masked by conjugation with natural endogenous constituent as glucuronic acid , glutathione, sulphate , acetic acid, glycine or methyl group.
- These reactions usually result in drug inactivation with few exceptions e.g. morphine-6-conjugate is active
- Most of drugs pass through phase I only or phase II only or phase I then phase II.
- Some drugs as isoniazid passes first through phase II then phase I (acetylated then hydrolyzed to isonicotinic acid).

Factors affecting drug metabolism

1. Drugs
2. Genetic variation
3. Nutritional state
4. Dosage
5. Age

Elimination Or Excretion

Elimination-Termination of Drug Action by which a drug or metabolite is eliminated from the body. Drugs and their metabolites are excreted in Urine, Faeces, Exhaled air, Saliva and sweat.

- Two-stage kidney process (filter, absorption)
- Metabolites that are poorly reabsorbed by kidney are excreted in urine.
- Some drugs have active (lipid soluble) metabolites that are reabsorbed into circulation (e.g., pro-drugs)
- Other routes of elimination: lungs, bile, skin

Terminologies In Pharmacokinetics

1. Elimination Half-Life - time required for drug blood levels to be reduced by 50%

2. Volume of Distribution (theoretical volume that would have to be available for drug to disperse) $= \frac{\text{Dose}}{\text{Plasma Concentration}}$

3. Clearance = Volume of blood cleared of drug per unit time

Pharmacodynamics

- Pharmacodynamics refers to the relationship between drug concentration at the site of action and the resulting effect, including the time course and intensity of therapeutic and adverse effects.
- The effect of a drug present at the site of action is determined by that drug's binding with a receptor.
- The concentration at the site of the receptor determines the intensity of a drug's effect

Drug Action

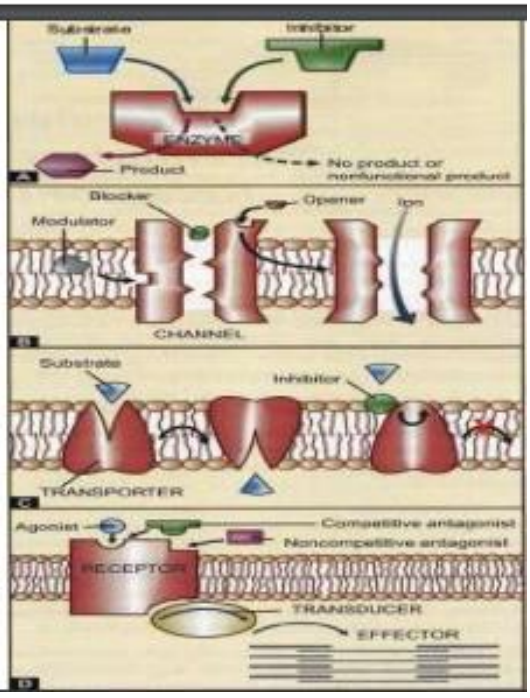
Four major types of bio-macromolecular targets of drug action is there,

(A) Enzyme

(B) Transmembrane ion channel

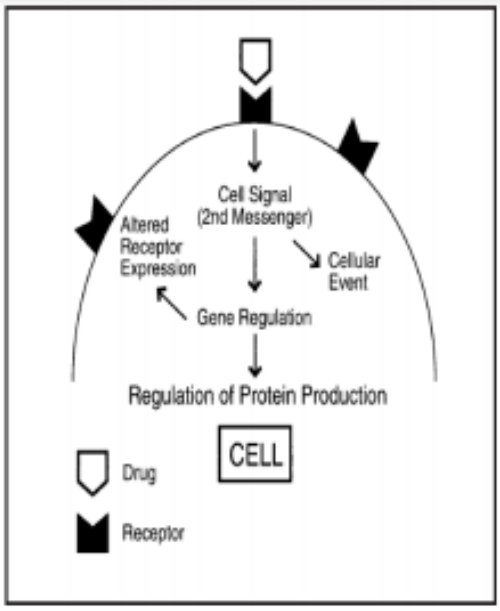
(C) Membrane bound transporter

(D) Receptor



Factors Affect Drug Response

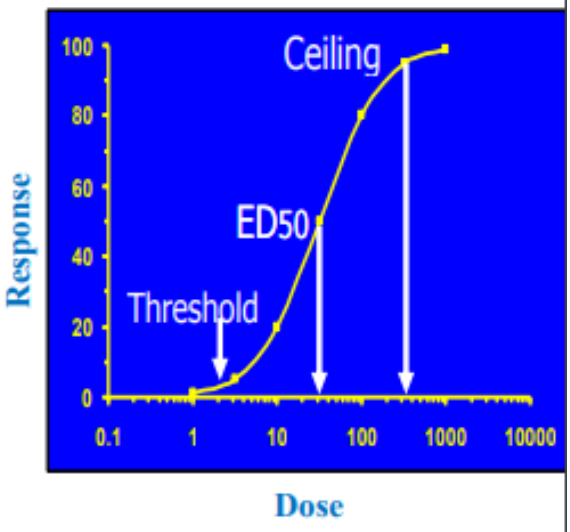
- Density of receptors on the cell surface
- the mechanism by which a signal is transmitted into the cell by second messengers.
- Regulatory factors that control gene translation and protein production may influence drug effect



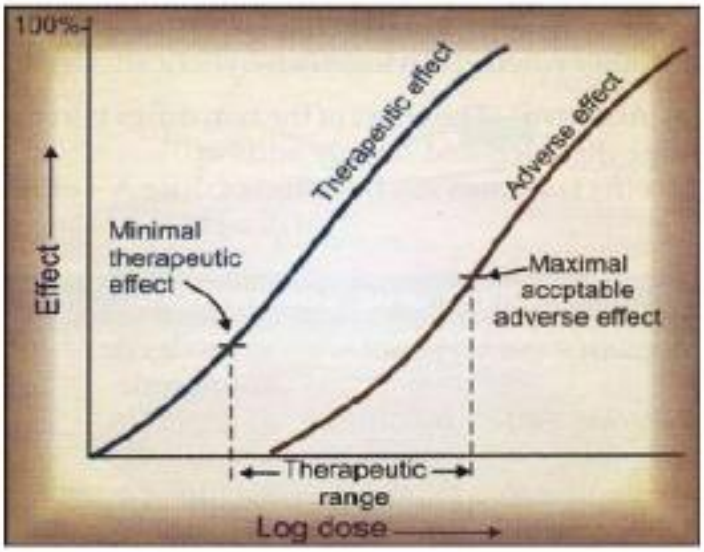
Dose-response Curves

Individual responses to varying doses

- ✓Threshold: Dose that produces a just-noticeable effect.
- ✓ED_{50}: Dose that produces a 50% of maximum response.
- ✓Ceiling: Lowest dose that produces a maximal effect.



Dose-response Curves for Therapeutic Effect And Adverse Effect Of The Same Drug



Dose-Response Functions

Efficacy ED_{50} = median effective dose

Lethality LD_{50} = median lethal dose

Therapeutic Index = LD_{50}/ED_{50}
= toxic dose/effective dose

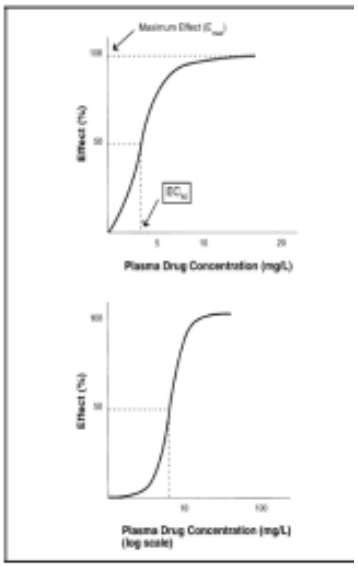
This is a measure of a drug's safety

A large number = a wide margin of safety

A small number = a small margin of safety

Maximum Effect(E_{max}) & EC_{50}

- When the logarithm of concentration is plotted versus effect, one can see that there is a concentration below which no effect is observed and a concentration above which no greater effect is achieved.
- 50% effective concentration Or EC50 the concentration at which 50% of the maximum effect is achieved.
- The EC50 does not, however, indicate other important determinants of drug response, such as the duration of effect.



Duration of effect

Duration of effect is determined by a complex set of factors, including

The time that a drug is engaged on the receptor

Intracellular signaling

Gene regulation.

Time Course Studies important for

Predicting dosages/dosing intervals

Maintaining therapeutic levels

Determining time to elimination

Tolerance

- The effectiveness can decrease with continued use is referred to as tolerance.

Tolerance may be caused by

The pharmacokinetic factors, such as increased drug metabolism, that decrease the concentrations achieved with a given dose.

The pharmacodynamic factors like when the same concentration at the receptor site results in a reduced effect with repeated exposure.