SEMESTER III (PAPER II)

(UNIT 1)

INTRODUCTION TO CHEMOINFORMATICS

1. Explain Chemoinformatics.

1) The intrinsic information in huge amounts of data produced in chemistry and pharmaceutical research is often difficult to grasp, since data contains information about various characteristics of chemical compounds and various methods have to be applied to extract the relevant information.
2) Thus, it has been realized that this huge amount of data and information can only be processed and analyzed by computer methods.
3) Furthermore, many of the problems faced in chemistry are so complex that novel approaches utilizing solutions that are based on informatics methods are needed.
4) Thus, methods were developed for building databases on chemical compounds and reactions, for the prediction of physical, chemical and biological properties of compounds and materials, for drug design, for structure elucidation, for the prediction of chemical reactions and for the design of organic syntheses.
5) Through the application of information technology, chemical informatics helps chemists organize and analyze known scientific data and extract new information from that data to assist in the development of novel compounds, materials, and processes.
6) Therefore, this database on chemical compounds is variously known as chemoinformatics, cheminformatics, or even chemiinformatics.
7) By accelerating the process of chemical synthesis, this method is having a profound effect on all branches of chemistry, but especially on drug discovery.
8) Through the rapidly evolving technology of combinatorial chemistry, it is now possible to produce compound libraries to screen for novel bioactivities.
9) This powerful new technology has begun to help pharmaceutical companies to find new drug candidates quickly, save significant money in preclinical development costs and ultimately change their fundamental approach to drug discovery.
10) Instead of preparing and examining a single compound, families of new substances are synthesized and screened.

- **Definitions:**

1) "The use of information technology and management has become a critical part of the drug discovery process. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of

making better decisions faster in the areaof drug lead identification and organization." - **Frank K.Brown** (1998)

2) "Chemoinformatics - a new name for an old problem."- **M. Hann, R Green3**
3) "Chem(o)informatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization and use of chemical information." - **Greg Paris** (In August 1999)
4) Chemoinformatics is the application of informatics methods to solve chemical problems.

– J.Gasteiger

2. Why is it required to study Chemoinformatics? Describe.

1) Chemical Abstracts Service adds over three-quarters of a million new compounds to its database annually, for which large amounts of physical and chemical property data are available.
2) Some groups generate hundreds of thousands to millions of compounds on a regular basis through combinatorial chemistry that are screened for biological activity.
3) When making combinatorial libraries of chemical compounds, information is needed on the molecular components, their biological effects, and information on how to prepare the compound.
4) In high throughput screening, the test results need to be captured, stored, and then analyzed.
5) Three dimensional structures determined by x -ray crystallography known for about 300,000 organic compounds. Or the largest database of infrared spectra contains about 200,000 spectra. It is only 1% of all the available compounds.
6) This is another reason why there was a need for informatics methods in chemistry.
7) This is true, for the relationships between the structure of a compound and its biological activity, or for the influence of reaction conditions on chemical reactivity. All these problems in chemistry require novel approaches for managing large amounts of chemical structures and data, for knowledge extraction from data, and for modeling complex relationships.
8) Therefore, all these limitations could only be solved through chemoinformatics that can help in this endeavor; it can integrate a comprehensive knowledge of chemistry with an extensive understanding of information technology.
9) There are four key problems a cheminformatics system solves:
   i. Store a molecule;
   ii. Find exact molecule;
   iii. Substructure search;

iv. Similarity search.

10) It helps to know about the relationships between the structure of a compound and its biological activity,or for the influence of reaction conditions on chemical activity.

11) It can help out in producing and manage information metrics to reduce risk, e.g.,

i. Virtual screening

ii. Library design

iii. Docking

iv. Cost/benefit analysis

12) The main purpose that chemoinformatics serves is, that it involves organization of chemical data in a logical form to facilitate the process of: data organization, understanding chemical properties, their relationship to structures, making inferences, assessing the properties of new compounds by comparison with known compounds.

3. Write a note on history of Chemoinformatics.

1) Cheminformatics (sometimes spelled as chemoinformatics or chemiinformatics) is a relatively new discipline.
2) It has emerged from several older disciplines such as computational chemistry, computer chemistry, chemometrics, QSAR, chemical information, etc.
3) The names identifying these older disciplines can be controversial, but they have been studied for many years.
4) Cheminformatics involves the use of computer technologies to process chemical data.
5) Initial activities in the field started with chemical document processing (the Journal of Chemical Documentation was published in 1961 by ACS. It was renamed the Journal of Chemical Information & Computer Science after 1974).
6) The differentiation between chemical data processing from other data processing is that chemical data involves the requirement to work with chemical structures.
7) This requirement necessitated the introduction of special approaches to represent, store and retrieve structures in a computer system.
8) Another challenge faced by this new field was to establish clear relationships between structural patterns and activities or properties.
9) One of the earliest Cheminformatics studies involved chemical structure representations, such as structural descriptors.
10) The first book Computer Handling of Chemical Structure Information was appeared in 1971.
11) The term ‘Chemoinformatics’ was defined by F.K. Brown in 1998.

4. Comment on Chemoinformatics Vs Cheminformatics.

1) The term cheminformatics was defined by F.K.Brown in 1998.
2) Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and optimization.
3) Since then, both spellings have been used, and some have evolved to be established as Cheminformatics, while European Academia settled in 2006 for Chemoinformatics.

5. Describe application of Chemoinformatics in details (Any 6)

1) Drug mining:

➔ It is the nontrivial extraction of implicit, previously unknown and potentially useful information from data, or the search for relationships and global patterns that exist in databases.

2) Drug design:

➔ It includes not only ligand design, but also pharmacokinetics (Pharmocogenomics) toxicity, which are mostly beyond the possibilities of structure- and/ or computer-aided design.

➔ Nevertheless, appropriate chemometric (Chemoinformatics) tools, including experimental design and multivariate statistics, can be of value in the planning and evaluation of pharmacokinetic and toxicological experiments and results.

➔ The molecular designing of drugs for specific purposes are based on knowledge of molecular properties such as activity of functional groups, molecular geometry, and electronic structure, and also on information cataloged on analogous molecules.

3) Storage and retrieval:

➔ The storage, indexing and search of information is the primary application of chemoinformatics that relates to the compounds including: unstructured data, digital libraries, information retrieval, and information extraction.

➔ The data should be arranged properly in the database so that it can be retrieved easily, and therefore, software is needed for drawing and saving up the information to manage the flood of data.

4) Virtual screening:

➔ Virtual screening involves computationally screening *in silico* libraries of compounds.

➔ It is the process of screening of large databases on the computer for molecules having desired properties and biological activity.

➔ A major application of virtual screening techniques is the identification of novel active molecules in large compound databases.

5) Virtual database assembly:

➔ A crucial activity as it enables access to the large number of drug- like molecules that could theoretically be made... can serve several purposes: for example, to generate a maximally diverse virtual library for lead generation, a biased library aimed at a specific target or target family, or a lead optimization library.

6) Quantitative structure-activity relationship (QSAR):

➔ QSAR is the calculation of quantitative structure-activity relationship and quantitative structure property relationship values, which can be used to predict the activity of compounds from their structures.

6. Elaborate on types of learning approach used in Chemoinformatics.

- There are two types of learning approaches, they are as follows:-

1. Inductive learning:

- In inductive learning, knowledge is extracted from chemical information.
- Inductive reasoning is part-to-whole wherein the content of the subject-matter is sequenced from particular concepts to general concept.
- Specifically the concepts related to each other are organized in a way that most prerequisite knowledge to the general concept is presented firstly.
- It uses collected information to derive a general principle.
- It is used in learning from data, for making predictions on chemical phenomena.
- The process of data-driven modeling using computational statistical techniques is an example of inductive learning.
- Other examples include: Artificial intelligence, machine learning, statistics, structure-property relationships and model building from experimental data.

2. Deductive learning:

- In deductive learning, data is extracted from knowledge.
- Deductive reasoning is whole-to-part wherein the content of the subject matter is organized from general concept to the particular concepts.
- It makes use of a fundamental theory which allows us to calculate properties and predict the behavior of molecules.
- It is used to infer specific conclusions.
- It allows us to calculate properties and predict the behavior of molecules.
- Quantum methods, molecular mechanics.

7. Representation of chemical structures/ molecular structures:

1) There was an emphasis on computational representation of molecular structures and creation of structural databases.
2) Special methods had to be devised to uniquely represent a chemical structure, to perceive features such as rings and aromaticity, and to treat stereochemistry, 3D structures, or molecular surfaces.
3) A whole range of methods for the computer representation of chemical compounds and structures has been developed: linear codes, connection tables, matrices.

- **Types of Chemical Representations for 2D molecular structures:**

- Structures are needed to be included in computer readable form.

  - **Linear Notations**
  - **Graph theory**
  - **Connection tables**

1. Linear notations:

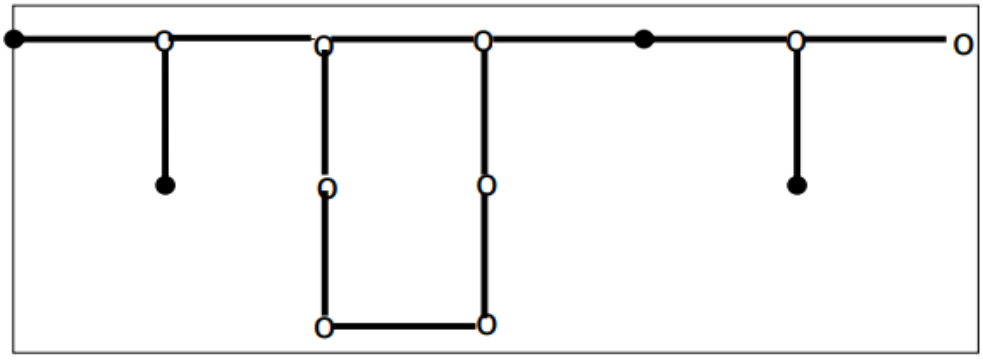
- Linear/line notations represents the structure of chemical compounds as a linear sequence of letters and numbers.
- They are useful for storing and transmitting large number of molecules.
- Various line notations are: WLN12 (Wiswesser Line Notation), ROSDAL (Representation of Organic Structures Description Arranged Linearly), SMILES13 (Simplified Molecular Input Line Entry Specification) and SLN (Sybyl).
- The earliest structure linear notation was the Wiswesser Line Notation (WLN).
- SMILES are widely accepted. In SMILES, atoms are represented by their atomic symbols, Upper case – for aliphatic, Lower case – for aromatic, Hydrogen - not represented, Double bonds – "=" , Triple bonds - "#", Single and aromatic bonds are not explicitly represented by any symbol.
- This system efficiently compressed structural data and, was very useful for storing and searching chemical structures in low performance computer systems.
- However, the WLN is difficult for non-experts to understand. Later, David Weininger suggested a new linear notation designated as SMILESTM.
- To successfully represent a structure, a linear notation should be canonicalized.
- That is, one structure should not correspond to more than one linear notation string, and conversely, one linear notation string should only be interpreted as one structure.
- For example:2- Methyl propane – CC(C)C

- **Canonical representation of molecular structures:**

- ➔ Canonical representation is unique ordering of atoms for a given graph.
- ➔ It is required due to different ways of constructing connection table or the SMILES string for a given molecule.
- ➔ In a connection table one may choose different ways to number the atoms and in SMILES notation the SMILES string may be written starting at a different atom or by following a different sequence through the molecule.
- ➔ A well-known and widely used method for determining a canonical order of the atoms is the Morgan algorithm and SEMA.
- ➔ An algorithm called CANGEN has been developed to generate a unique SMILES string for each molecule.
- ➔ Example: Aspirin is: CC (=O) Oc1ccccc1C (=O) O.

2. Graph theory:

- A molecular graph (2D structure) can also be canonicalized into a real number through a mathematical algorithm.
- The graph theory approach was given by H.L.Morgan in 1965.
- The real number is identified as a molecular topologic index.
- However, two different structures can have the same topologic index.
- Therefore, topologic indices can only be used as screens for accelerating structure database searching.
- Actually, the concept of molecular topological index was originally proposed for QSAR and QSPR studies.
- Wiener reported the first molecular topological index in 1947.
- Example:



Aspirin

Source: J. Zupan *et.al*.,(1999).

The aspirin structure can be represented by Graph theory, where Oxygen atom is represented by filled bullet and carbon atom is represented by vacant bullet and hydrogen atom is not represented here. So, the aspirin structure will be-

2.1. Topological graph theory:

- It is the study of graphs which consist of a set of "nodes" and a set of "edges" joining pairs of nodes.
- It provides details about graphs in terms of charges, stereochemistry, etc.
- Properties of graphs:
  - Graphs are only about connectivity.
  - Spatial position of nodes is irrelevant.
  - Length of edges is irrelevant.
  - Crossing edges are irrelevant.
- If a molecule and its specific topologic index had a one-to-one relationship, then structure search could be done by number comparison.
- However, substructure search still had to use an atom-by-atom matching algorithm, which, as mentioned earlier, could be very time-consuming.

2.2 Matrix representations:

- It was an extended method of graph theory.
- A molecular structure with n atoms may be represented by an n × n matrix (H-atoms are often omitted).
- The matrix to represent a graph in this theory way is called **Adjacent matrix** that indicates which atoms are bonded.
- 1 and 0 representation shows bonded and non-bonded atoms.
- **Distance matrix**: encodes the distances between atoms.

- The distance is defined as the number of bonds between atoms on the shortest possible path.
- Distance may also be defined as the 3D distance between atoms.
- **Bond matrix:** indicates which atoms are bonded, and the corresponding bond orders.

3. Connection tables:

- A disadvantage of matrix representations is that the matrix size increases with the square of the number of atoms.
- Therefore, connection tables were a way to communicate the molecular graph to and from the computer.
- **The simplest type of connection tables consists of two sections:**

i) List of the bonds, specified as pairs of bonded atoms.

ii) More detailed form of connection table includes: List of atomic numbers, List of the bonds, Hybridization state, Bond order, Information about xy or xyz coordinates of atoms

4. Fingerprints:

➔ A chemical structure may be indexed on the basis of specific chemical characteristics, usually fragments.

➔ Fragments may be: a small group of atoms, a functional group, rings. These are defined beforehand.

➔ It is an ambiguous representation: different structures may have common fragments.

➔ Therefore, Fingerprints characterizes the 2D structure of a molecule, usually through a string of '1's and '0's.

➔ It encodes the presence and absence of certain features in a compound, eg., fragments.

➔ There are 2 basic types of fingerprint:

1) Structural keys:

➔ Structural keys contain a string of bits ('1's and '0's) where each bit is set to 1 or 0 depending on the presence or absence of a particular fragment. They usually employ a pre-defined dictionary of fragments.

2) Hashed fingerprints:

➔ In hashed fingerprints, there is no set dictionary or 1:1 relationship between bits and features.

➔ All possible fragments in a compound are generated.

➔ The number of fragments represented can be huge.

➔ Thus rather than assigning one bit position for each fragment, the bits are "hashed" down onto a fixed number of bits.

➔ Thus hashed fingerprints are a less precise form, but they carry more information.

➔ Once fingerprint representations are available, similarity coefficients can be used to give a measure of similarity between two fingerprints.

- **Types of Chemical Representations for 3D molecular structures:**

➔ Two Dimensional (2D) representations of molecules only tell about atoms, which are bonded together. It doesn't tell about steric & electronic parameters and atom positions in 3D space.

➔ Three Dimensional representations of molecules have following challenges:

➔ Molecules can adopt more than one low energy conformation and

➔ The number of accessible structures is very large.

➔ So there is need to represent molecular structures in 3D. The data stored in a 3D database either comes from Experimental methods or Computational methods.

1) Experimental 3D databases:

➔ It includes structures solved using X- ray Crystallography.

➔ The Cambridge Structural Database contains the X-ray structures of more than 250,000 organic and organometallic compounds.

➔ It stores crystal structures of small molecules and provides a fertile resource for geometrical data on molecular fragments for calibration of force fields and validation of results from computational chemistry.

➔ As protein crystallography gained momentum, the need for a common repository of macromolecular structural data led to the Protein Data Base (PDB).

➔ The PDB (Protein Data Bank) contains more than 20,000 X-ray & NMR structure of proteins and protein-ligand complexes and some nucleic acid and carbohydrate structures.

➔ Both these databases are widely used and continue to grow rapidly.

2) Theoretical 3D databases:

➔ The PDB & CSD are extremely useful but for most compounds no crystal structure is available. T here is also an increasing need to evaluate virtual compounds –
structures that could be synthesized but which have not yet been made.

➔ Even when experimental data is available, this usually provides just a single conformation of the molecule which will not necessarily correspond to the active conformation; most molecules have a number of conformations accessible to them.

➔ It is thus desirable to include mechanisms for taking conformational space of the molecules into ac count during 3D database searching.

➔ Structure-generation programs such as CONCORD20, CORINA21, and COBRA take 2D representation of molecule and generate a low energy conformation.

➔ These programs generate one conformation.

8. How are these representations saved into the computer?

1) MDL Molfile format:

➔ An MDL Molfile is a file format for holding information about the atoms, bonds, connectivity and coordinates of a molecule.

➔ The molfile consists of some header information, the Connection Table (CT) containing atom info, then bond connections and types, followed by sections for more complex information.

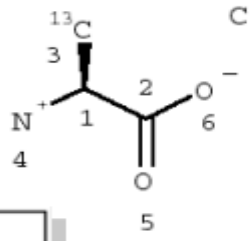
➔ The molfile is sufficiently common that most, if not all, cheminformatics software systems/applications are able to read the format, though not always to the same degree.

The bond block

111222tttsssxxxrrrccc

```
1  2  1  0  0  0
1  3  1  1  0  0
1  4  1  0  0  0
2  5  2  0  0  0
2  6  1  0  0  0
```

L-Alanine



Chiral

Field	Meaning	Values	Notes
111	first atom number	1 - number of atoms	[Generic]
222	second atom number	1 - number of atoms	[Generic]
ttt	bond type	1 = Single, 2 = Double, 3 = Triple, 4 = Aromatic, 5 = Single or Double, 6 = Single or Aromatic, 7 = Double or Aromatic, 8 = Any	[Query] Values 4 through 8 are for SSS queries only.
sss	bond stereo	Single bonds: 0 = not stereo, 1 = Up, 4 = Either, 6 = Down, Double bonds: 0 = Use x-, y-, z-coords from atom block to determine cis or trans, 3 = Cis or trans (either) double bond	[Generic] The wedge (pointed) end of the stereo bond is at the first atom (Field 111 above)
xxx	not used		
rrr	bond topology	0 = Either, 1 = Ring, 2 = Chain	[Query] SSS queries only.
ccc	reacting center status	0 = unmarked, 1 = a center, -1 = not a center, Additional: 2 = no change, 4 = bond made/broken, 8 = bond order changes 12 = 4+8 (both made/broken and changes); 5 = (4 + 1), 9 = (8 + 1), and 13 = (12 + 1) are also possible	[Reaction, Query]

2) SDF (.sdf) format:

➔ Includes structural information in the Molfile format but gives associated data items for one or more compounds.

➔ The associated data comes into different categories that include charge information, exact mass, and

Marwin form of structure representation.

3) SMILES:

➔ Short and readable descriptions of molecular graphs are linear notations.

➔ A clear example is the broadly used Simplified Molecular Input Line System (SMILES), which captures a molecules' structure in the form of an unambiguous text string using alphanumeric characters.

➔ They allow the efficient storage and fast processing of large numbers of molecules.

The SMILES notation uses the following basic rules for encoding molecules:

a) Atoms are represented by their atomic symbols.

b) Hydrogen atoms saturating free valences are not represented explicitly.

c) Neighboring atoms stand next to each other, and bonds are characterized as being single (-), double (=), triple (#), or aromatic (:).

d) Single and aromatic bonds are usually omitted.

e) Enclosures in parentheses specify branches in the molecular structure.

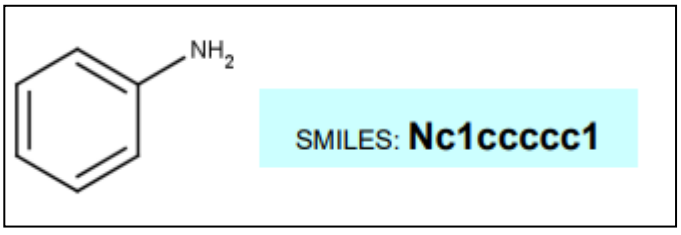
f) For the linear representation of cyclic structures, a bond is broken in each ring and the connecting ring atoms are followed by the same digit in the textual representation.

g) Atoms in aromatic rings are indicated by lower case letters. In some cases, there may be problems with aromaticity perception.

➔ Although SMILES strings are unambiguous in describing chemical structures, they are not unique because multiple valid SMILES representations exist for the same molecular graph.

➔ Canonical SMILES strings are often used to ensure the uniqueness of molecules in a database.

➔ Example:



➔ Representation of cis-trans stereochemistry in double bonds Stereochemistry around a double bond (cis/trans) is specified with characters '\' and '/'.

➔ To represent rings, you need to break a ring bond and replace it by a ring opening symbol and a corresponding ring closure.

4) SMARTS:

➔ SMILES Arbitrary Target Specification (SMARTS) is a language developed to specify sub-structural patterns used to match molecules and reactions.

➔ Substructure specification is achieved using rules that are extensions of SMILES.

➔ In particular, the atom and bond labels are extended to also include logical operators and other special symbols, which allow SMARTS atoms and bonds to be more inclusive.

➔ This notation is especially useful for finding molecules with a particular substructure in a database.

➔ SMARTS can also be used to filter out molecules with substructures that are associated with toxicological problems or that appear as frequent hitters in many biochemical high-throughput screens.

➔ Other applications are the separation of active from inactive compounds and the evaluation of ligand selectivity.

➔ The SMARTS language provides several primitive symbols describing atomic and bond properties beyond those used in SMILES (atomic symbol, charge, and isotopic specifications).

5) InChI and InChI Keys:

1) InChI is the International Chemical Identifier developed under IUPAC's auspices, the International Union of Pure and Applied Chemistry, with principal contributions from NIST (the U.S. National Institute of Standards and Technology) and the InChI Trust.
2) The InChI objective is to establish a unique label for each compound and allow an easier linking of diverse data compilations.
3) This notation resolves many of the chemical ambiguities not addressed by SMILES, particularly concerning stereocenters, tautomers, and other valence model problems.
4) However, InChIs are difficult to read and interpret by humans in most cases.
5) InChIs comprise different layers and sub- layers of information separated by slashes (/).
6) Each InChI string starts with the InChI version number, followed by the main layer.
7) This main layer contains sub-layers for empirical formula, atom connections, and hydrogen atoms positions.
8) The identity of each atom and its covalently bonded partners provide all of the information necessary for the main layer.
9) The main layer may be followed by additional layers, for example, for the charge, isotopic composition, tautomerism, and stereochemistry.

10) The InChIKey is a fixed-length (27-character) condensed digital representation of an InChI, developed to make it easy to perform web searches for chemical structures.
11) The first block of 14 characters for an InChIKey encodes core molecular constitution, as described by a formula, connectivity, hydrogen positions, and charge sub-layers of the InChI main layer.
12) The other structural features complementing the core data—namely exact positions of mobile hydrogens, stereochemical, isotopic, and metal ligands, whichever are applicable—are encoded by the second block of InChIKey.
13) The possible protonation or deprotonation of the core molecular entity is encoded in the very last InChIKey flag character.

9. Searching Chemical structures:

1) In order to retrieve data and information from databases, access has to be provided to chemical structure information.
2) Structure Searching involves determination of features like bond orders, rings and aromaticity.
3) It includes searching the whole structure, substructure, structure similarity and diversity.
4) A connection table is essentially a representation of the molecular graph.
5) Therefore, for storing a unique representation of a molecule and for allowing its retrieval, the graph isomorphism problem had to be solved to define from a set of potential representations of a molecule a single one as the unique one.
6) The first solution was the Morgan algorithm for numbering the atoms of a molecule in a unique and unambiguous manner.
7) This provided the basis for full structure searching.
8) Then, methods were developed for substructure searching, for similarity searching, and for 3D structure searching.
   - **Full structure searches**
   - **Substructure search**
   - **Three dimensional structure search:**

1) Full structure searches:

➔ It includes the searching as well as retrieval of information from databases.
➔ Structure searching can be done by use of molecular formula, molecular weight, Trade and/or trivial name, registry number, and hash codes.
➔ In it, structure is first converted to canonical representation and then Hash code is generated.
➔ Hash code corresponds to physical location on the computer disk.

- ➔ A hash code is a fixed length representation of a data structure used as an index or key to a direct access file.
- ➔ The input structure cannot be restored from a hash code, and due to the limited range of values, two different data structures may be represented by the same hash code.
- ➔ Ihlenfeldt and Gasteiger proposed to represent chemical structures with hash codes by using a hierarchical algorithm:
- ➔ atom hash codes are computed first, merged into molecule hash codes and the molecule hash codes are combined to give a molecular ensemble hash code.

2) Substructure search:

- ➔ A substructure search involves finding all the structures in a database that contain one or more particular structural fragments.
- ➔ For example, to find all of the structures in a database which contain the nitro group: Substructure searching requires some method of specifying a query (i.e., we want to find this and that, but not this, etc.).
- ➔ One popular example is SMARTS, an extension to SMILES.
- ➔ Mathematically, substructure searching is performed, as with structure searching, using a graph representation, but this time a sub-graph isomorphism algorithm finds occurrences of sub-graphs (i.e. substructures) in a structure.
- ➔ Two methods are used for substructure search, they are ,
  1) Sub-graph isomorphism:
     - ➢ To determine whether one graph is entirely contained within another. It is a slow process and has factorial degree of complexity.
  2) Bitstrings:
     - ➢ It consist of a sequence of "0"s and "1"s. A "1" in a bitstring usually indicates the presence of a particular structural feature and a "0" its absence.

3) **Three dimensional structure search:**

- ➔ Its objective is to identify conformations that match the query.
- ➔ It is a two-stage process:

  a) Rapid screen to encode information about the distances between relevant groups in molecular conformation.

  b) A sub-graph isomorphism such as Ullmann algorithm.

  Potential matches are then confirmed by fitting the relevant conformation to the query in

Cartesian space.

CONFORMATIONAL SEARCH AND ANALYSIS

- **Conformational analysis:**

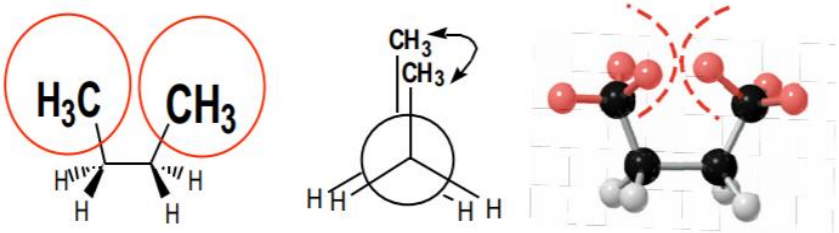
1. Conformational analysis is the study of the conformations of a molecule and their influence on its properties: structural, energetic, temporal, etc.
2. Conformations are the different 3-dimensional arrangements that the molecule can acquire by freely rotating around σ-bonds.
3. Conformations play an important role in prediction of not just physico-chemical properties but also the biological activity of the drug.
4. A key component of a conformational analysis is the conformational search, the objective of which is to identify the 'preferred' conformations of a molecule: those conformations that determine its behavior.
5. It is done by exploring the energy surface of a molecule and determining the conformation with minimum energy.
6. Energy minimization methods play a crucial role in conformational analysis.
7. An important feature of methods for performing energy minimization is that they move to the minimum point that is closest to the starting structure.
8. For this, it is necessary to have a separate algorithm which generates the initial starting structures for subsequent minimization.
9. Each of these in turn is then subjected to energy minimization in order to derive the associated minimum energy structure.
10. Global minimum-energy conformation is the conformation with the lowest energy.
11. Conformational analysis is an important step in molecular modeling as it is necessary to reduce time spent in screening of compounds for activity.

12. Most drugs are flexible molecules with the ability to adopt different conformations by means of rotation about single bonds.
13. The usual strategy in conformational analysis is to use a search algorithm to generate a series of initial conformations.

❖ **Conformational Energy:**

1. The relationship between molecular structure and potential energy is a major area of study in organic chemistry.
2. It has applications in conformational analysis because the main interest is in relating the structure of conformers to their energy, and therefore to their relative stabilities.
3. The higher the potential energy of a system the lower its stability.
4. These are always relative concepts there is no absolute energy or stability.
5. They are always measured relative to a previously agreed upon standard.
6. There are 3 factors that increase the potential energy of conformers and therefore, decrease their stability. They are as follows:
    a) **Steric interactions:** Crowding of alkyl groups or other substituent's as they come close together.
    b) **Torsional strain** - Tendency of s-bonds to rotate in order to acquire a more stable conformation.
    c) **Angle strain** - Increase in potential energy due to bond angles being forced to depart from ideal values in cycloalkanes and other rings.

❖ **Steric interactions:**

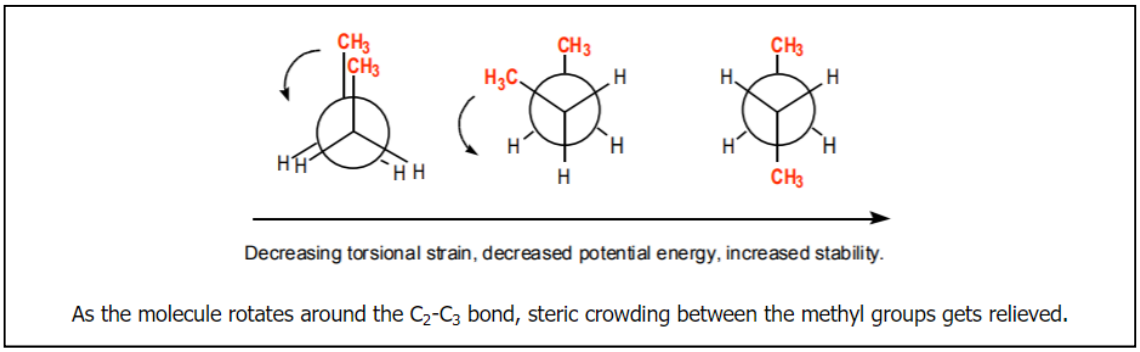


Three representations of steric interactions, or "crowding," between the two methyl groups in the eclipsed conformation of *n*-butane.

1. The term steric interactions refer to interference that occurs between atoms or groups of atoms by virtue of their size, or volume.
2. When bulky groups or large molecules get too close to each other, steric interactions can become severe, raising the potential energy of the system.
3. When these groups are allowed to drift far apart, steric interactions are relieved and the system gains stability.

❖ **Torsional strain:**

1. This term is closely related to dihedral angle (also called torsional angle).
2. Free rotation around sigma bonds changes the dihedral angle and therefore can bring bulky groups together or apart.
3. When the dihedral angle is small and bulky groups interact (such as in the illustration above), there is a driving force for the molecule to rotate around the C_2-C_3 axis to relieve the steric interactions between the methyl groups.
4. If this drive is impeded and the molecule is forced to acquire the eclipsed conformation, or a small dihedral angles between bulky groups, then torsional strain results.
5. This might be the case for example in cyclic systems, where free rotation around sigma bonds is limited or impeded.
6. Obviously, torsional strain increases the energy of a system and therefore decreases its stability.

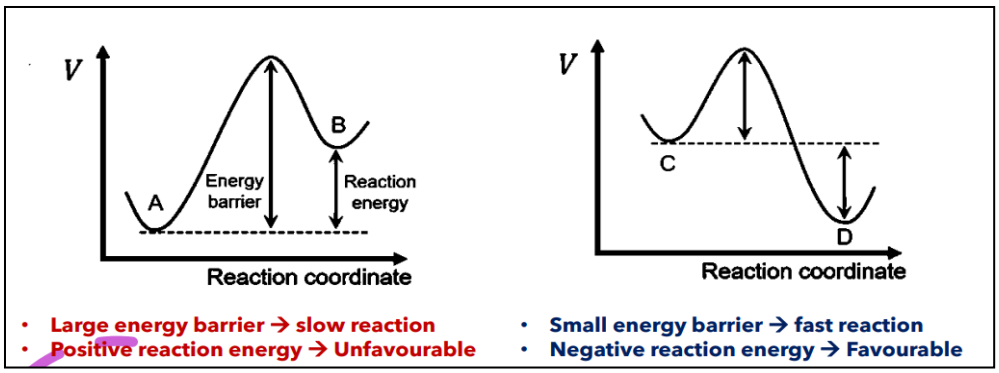


❖ **Potential Energy:**

1. The potential energy dictates the behaviour of the system.
2. It is the interaction energy among all the particles of the system.

$$\mathbf{V = V_{NN} + V_{ee} + V_{eN}}$$

3. V_{NN} , V_{ee} = Repulsive (positive)
4. V_{En} = Attractive (negative)



❖ **Force Fields:**

1. A force field is a simple equation that relates the PE of the system with its internal coordinates (bond distances, bond angle, ….).
2. In most used force fields, the PE is split into bonded and non-bonded interactions.

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

$$E_{bonded} = E_{bond} + E_{angle} + E_{dihedral}$$

$$E_{nonbonded} = E_{electrostatic} + E_{van\ der\ waals}$$

3. The atoms of the molecules are classified in different atom types to distinguish interactions between the same chemical classes of atoms.

ca: sp^2C in aromatic system

ha: H bonded to aromatic C

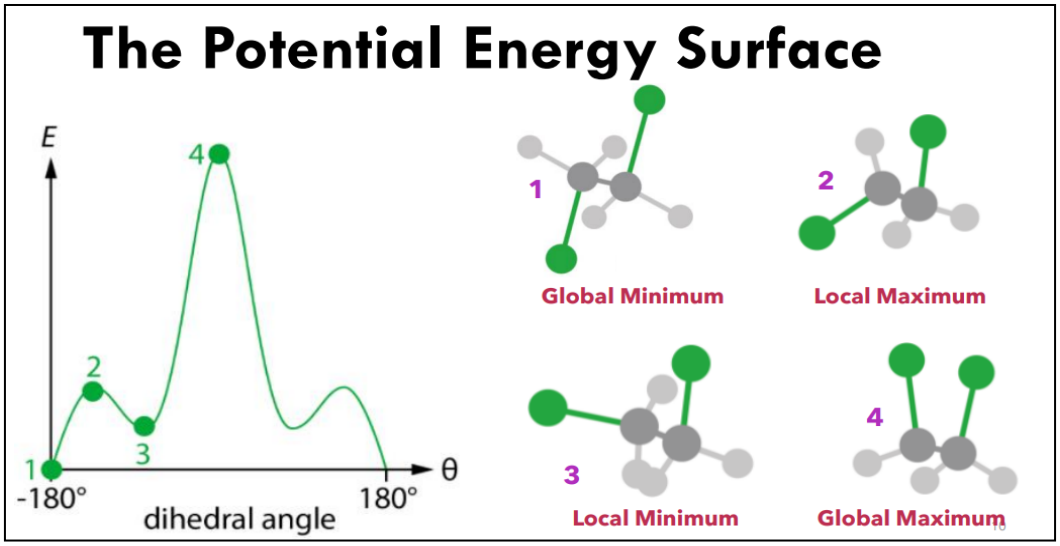
c: sp^2C of carbonyl group

o: O with one connected atom

oh: O in hydroxyl group

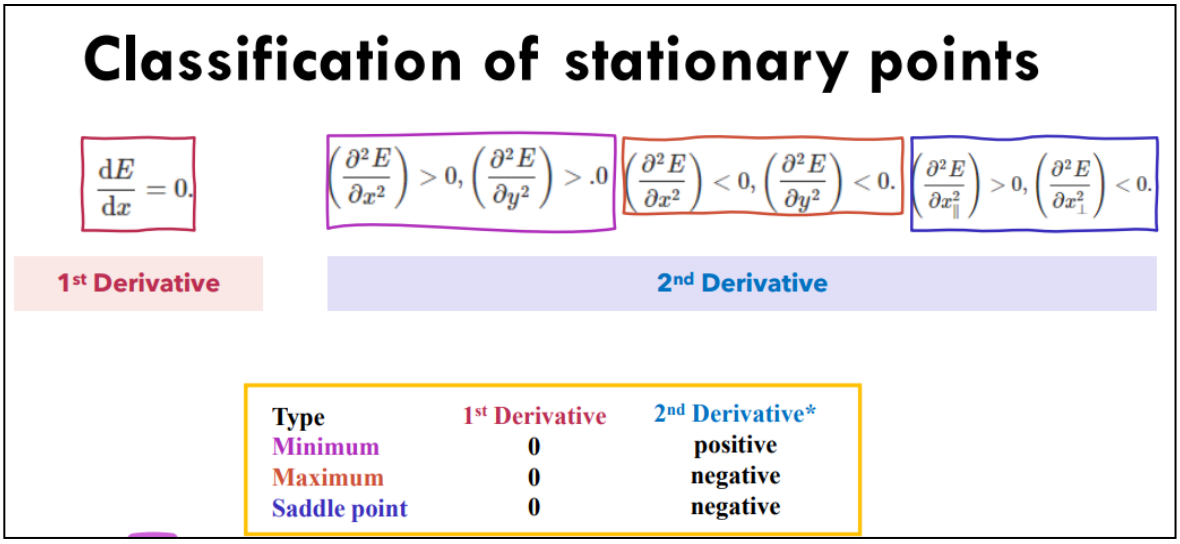
ho: H in hydroxyl group

❖ **Potential Energy Surface:**



1. **Global minimum:** Maximum stable
2. **Local maximum:** slightly unstable
3. **Local minimum:** slightly unstable
4. **Global maximum:** unstable

❖ **Classification of stationary points:**



❖ **Conformational search methods:**

- Conformational search methods can be divided into following categories: systematic search algorithms, model-building methods, random approaches, distance geometry and molecular dynamics.
- Most common methods are: Molecular dynamics and Monte-Carlo.

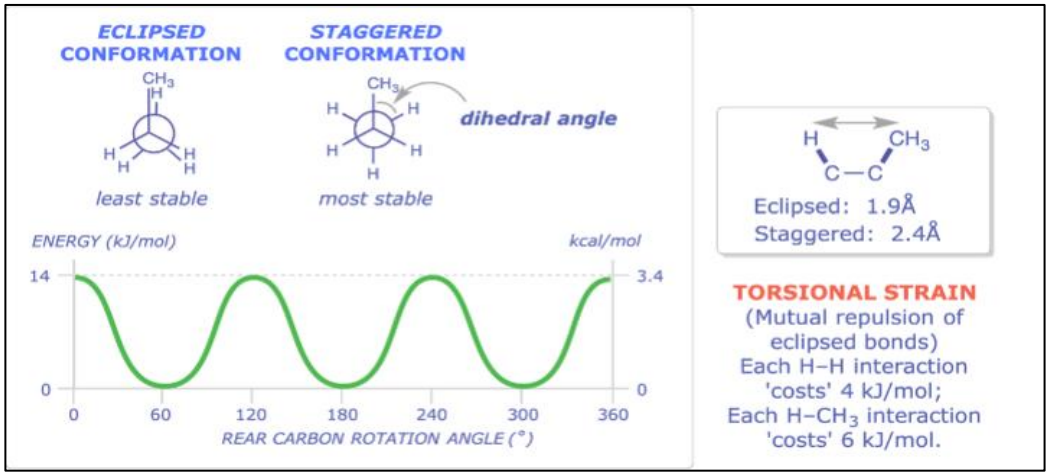
1. Systematic search:

1) In this type of method, every configuration is taken into consideration.
2) In systematic search, exploration of the energy surface of the molecule is carried out in a predictable pattern.
3) **Procedure:**

- All rotatable bonds are identified.
- Bond length and angles are fixed.
- Each bind is rotated one by one with fixed increment.
- A minimization follows each move.
- Finished when all possible combinations are done.

4) It explores the conformational space by making regular and predictable changes to the conformation.
5) It explores the energy surface of the molecule in a predictable fashion.
6) Example: Conformations of a propane:
   - ➔ Experiments show that there is a 14 kJ/mol (3.4 kcal/mol) barrier to rotation in propane.
   - ➔ The most stable (low energy) conformation is the one in which all of the bonds as far away from each other as possible i.e., staggered conformation in a Newmann projection.

- ➔ The least stable (high energy) conformation is the one in which, for any two adjacent carbon atoms, the six bonds (five C–H and one C–C) are as close as possible (eclipsed in a Newman projection).
- ➔ All other conformations lie between these two limits.
- ➔ The barrier to rotation is the result of one C–C/C–H eclipsing interaction and two equal C–H/C–H eclipsing interactions.
- ➔ We know from ethane that each C–H/C–H eclipsing interaction 'costs' about 4 kJ/mol (1 kcal/mol), so we can assign a value of about 6 kJ/mol (1.4 kcal/mol) to the C–C/C–H eclipsing interaction in propane.
- ➔ The 14 kJ/mol of extra energy in the eclipsed conformation of propane is called torsional strain.
- ➔ Energy minima occur at staggered conformations, and energy maxima occur at eclipsed conformations.
- ➔ The torsional strain is thought to be due to the slight repulsion between electron clouds in the eclipsed bonds.



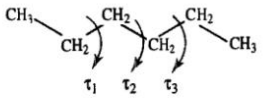
- ➔ Eclipse: High energy (unstable).
- ➔ Staggered: Low energy (stable).

7) Example: Conformations of a propane:

- **Combinatorial explosion**

$$\text{Number of conformations} = \prod_{i=1}^{N} \frac{360}{\theta_i}$$

Dihedral over bond i



- For **5 bonds** and 30 increments, the total resultant structures is **248832** conformers→ (if 1s per structure = 69 hours)
- For **7 bonds** and 30 increments, the total resultant structures is 36 million (**36, 000,000**) conformers → (if 1s per structure = 415 days)

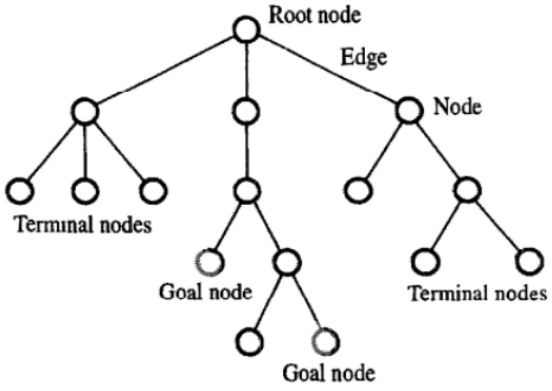
➔ It explores large regions with high energy.

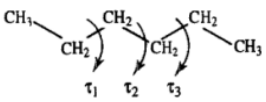
➔ It is not good for highly flexible molecules.

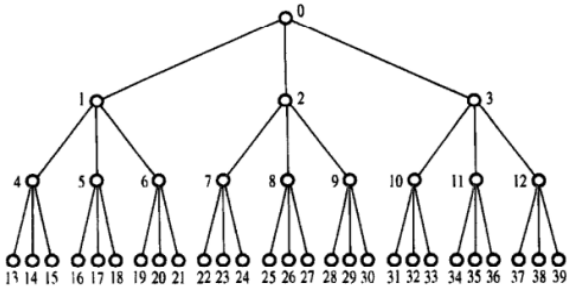
➔ Therefore, it is time consuming and cannot be used for large system.

8) Example: Search tree:

- **Search Trees**







➔ Search trees are widely used to represent the different states that a problem can adopt.

➔ A tree contains nodes that are connected by edges.

➔ The presence of an edge indicates that the two nodes it connects are related in some way.

➔ Each node represents a state that the system may adopt.

➔ The root node represents the initial state of the system.

➔ Terminal nodes have no child nodes.

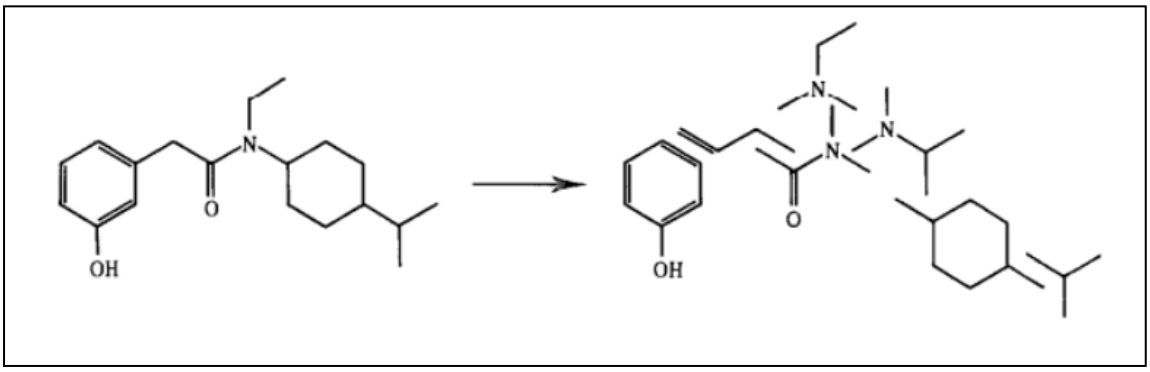
➔ A goal node is a special kind of terminal node that corresponds to an acceptable solution to a problem. It has the least energy possible state.

➔ 1 root node, multiple terminal nodes and the sum of them (1/more) is the goal node.

➔ A search tree eliminates the time consuming energy minimization stage for structures that have a very high energy or some other problem.

9) Systematic searches are therefore, subject to the effects of combinatorial explosion, and they are not naturally suited to molecules with rings.
10) However, they do have a definite endpoint, when the search has finished, one can be guaranteed to have found all conformations for a given dihedral increment.

2. Model-building approach:



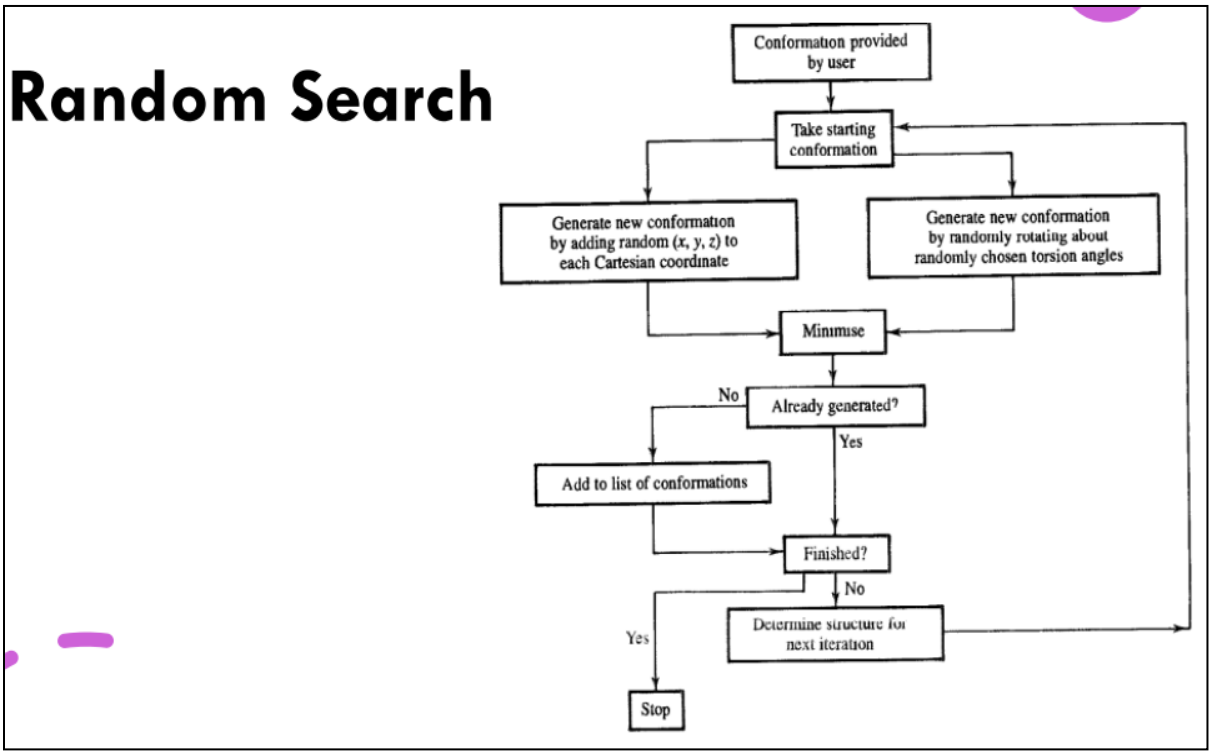
1) Fragment- or model-building approaches to conformational analysis construct conformations of molecules by joining together three dimensional structures of molecular fragments.
2) This approach would be expected to be more efficient than the normal systematic search because there are usually many fewer combinations of fragment conformations than combinations of torsion angle values.
3) Many molecular modeling systems offer a facility for constructing structures from molecular fragments, though the user usually has to specify manually which fragments are to be joined and how this is to be achieved.
4) Clearly, if each fragment can adopt a number of conformations, then it is impractical to tackle the problem manually and some means of automating the method is required.
5) A program to explore conformational space automatically using the fragment-building approach must first decide which fragments are needed to construct the molecule.
6) This is done using a substructure search algorithm, which determines whether each of the fragments that the program 'knows about' is present in the molecule.
7) Having identified the fragments that are required, conformations can be generated.
8) The conformations available to each fragment should span the range of conformations the fragment can adopt.

9) A conformation of the molecule is constructed by assigning a template to each fragment and then attempting to join the templates together.
10) The search problem can be represented as a tree, as for a systematic search, and so all of the usual tree-searching algorithms are applicable.
11) The search can be significantly enhanced by tree pruning.
12) The fragment-based approach to conformational analysis relies upon 2 assumptions: The first assumption is that each fragment must be conformationally independent of the other fragments in the molecule.
13) The second assumption is that the conformations stored for each fragment must cover the range of structures that are observed in fully constructed molecules.
14) The fragment conformations can be obtained from a variety of sources: 2 common approaches are by analyzing a structural database or from some other conformational search method.
15) A third limitation is that one can obviously only analyze molecules for which there are fragments available.

3. Random search:

1) A random search is, in many ways, the antithesis of a systematic search.
2) It is not possible to predict the order in which conformations will be generated by a random method.
3) A random search can move from one region of the energy surface to a completely unconnected region in a single step.
4) A random search can explore conformational space by changing either the atomic Cartesian coordinates or the torsion angles of rotatable bonds.
5) At each iteration, a random change is made to the 'current' conformation.
6) The new structure is then refined using energy minimization.
7) If the minimized conformation has not been found previously, it is stored.
8) The conformation to be used as the starting point for the next iteration is then chosen and the cycle starts again.
9) The procedure continues until a given number of iterations have been performed or until it is decided that no new conformations can be found.
10) Random search methods can require long runs to ensure that the conformational space has been covered, and they can generate the same structure many times.

11) Therefore, in a random search, there is no natural end point; one can never be absolutely sure that all of the minimum energy conformations have been found.
12) The usual strategy is to generate conformations until no new structures can be obtained.
13) This usually requires each structure to be generated many times and so the random methods inevitably explore each region of the conformational space a large number of times.



4. Distance Geometry:

1) Distance geometry uses the interatomic distances and various mathematical procedures to generate structures of conformations for energy minimization.
    - Matrix containing the maximum and minimum values permitted to each interatomic distance in the molecule is calculated.
    - Each interatomic distance is arbitrarily assigned values between the upper and lower bounds.
    - Distance matrix is transformed into trial set of Cartesian coordinates.
    - Refinement of structure and generation of conformation is the last stage.
2) Refinement of structure is carried out in accordance to simple trigonometric restrictions.
3) The distance between A and C can be no greater than the sum of maximum values of distances between AB and BC.

$$U_{AC} \leq U_{AB} + U_{BC}$$

4) Minimum value of AC distance can be no less than the difference between the lower bound on AB and the upper bound on BC.

$$L_{AC} \geq L_{AB} - U_{BC}$$

5) Distance geometry is particularly useful when experimental information can be incorporated, as it restrained molecular dynamics.

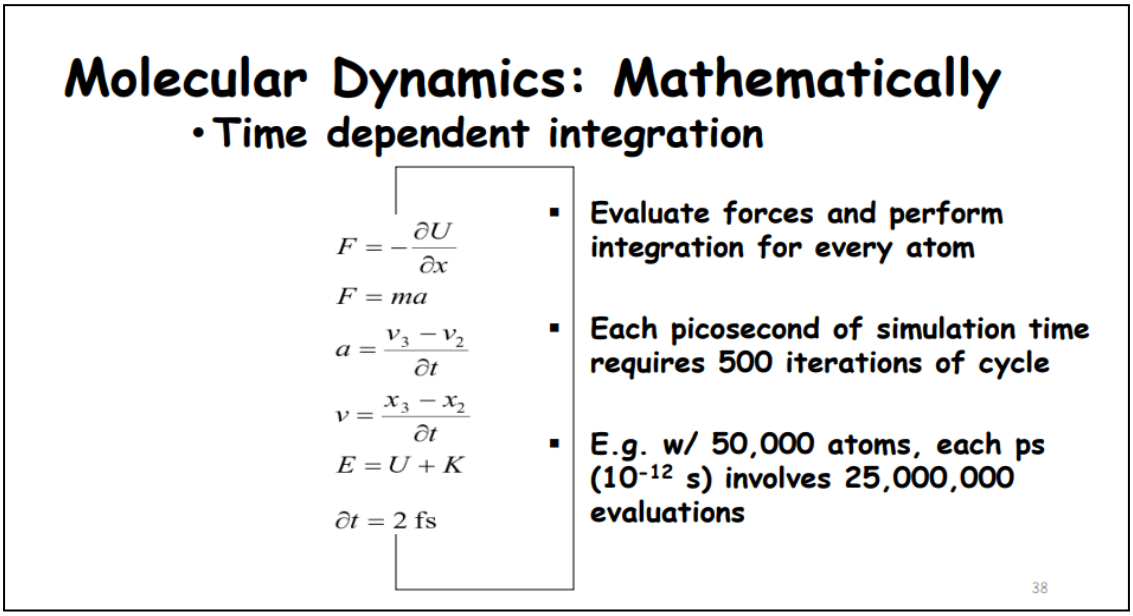
5. Molecular dynamics (Prediction concept):

1) Molecular dynamics (MD) is a conformational space search procedure in which the atoms of molecule are given an initial velocity and are then allowed to evolve in time according to the laws of Newtonian mechanics.
2) In this method, generation of successive configurations is carried out by incorporation of Newton's law of motion for the atoms in the system, to provide a trajectory that defines how the positions and velocities of the particles of the system vary with time.
3) Molecular dynamics is a computer program that treats the atoms within the molecules as moving spheres.
4) After 10^{-15} sec of movement, determination of the position and velocity of each atom in structure is used for the estimation of the forces by utilizing the values of bond lengths, bond angles, torsional terms and non-bonded interactions.
5) The calculation of potential energy of each atom and Newton's Law of Motion helps in the computation of acceleration and direction of each atom.
6) Generation of different conformations is carried out by program by "Heating" of molecule which implies that the molecule undergoes bond stretching and bond rotation as if it was being heated.
7) The process can be repeated automatically to give any number of practical structures.
8) Molecular Dynamics thus provides not only information about the conformation system but also the way in which the conformation changes with time.
9) The additional kinetic energy enhances the ability of the system to explore the energy surface and can prevent the molecule getting stuck in a localized region of conformational space.
10) Force fields:

- 1 and 2: bonds
- 1 and 3: angles.
- 1 and 4: either eclipsed or staggered conformation.

11) Molecular dynamics simulation are performed to:

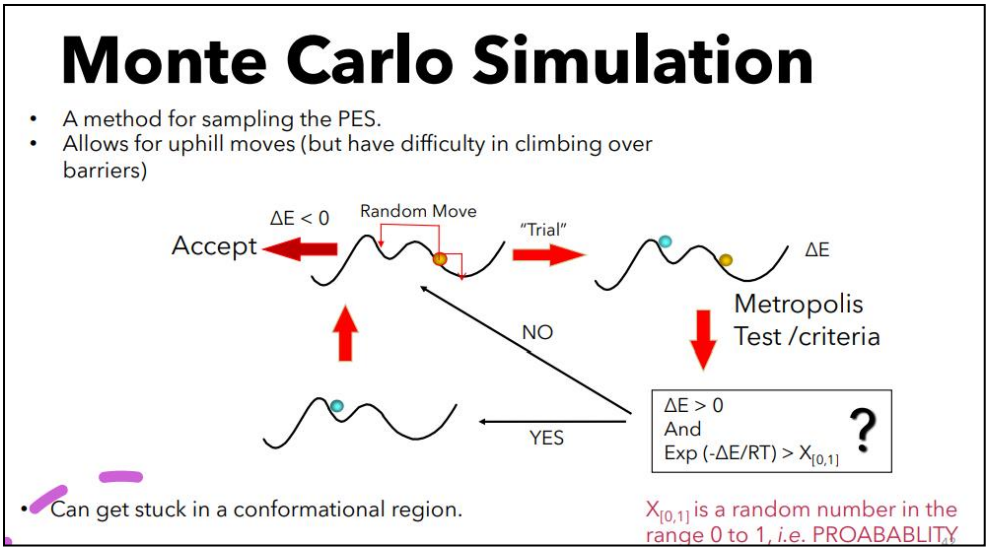
- Link physics, chemistry and biology
- Model phenomena that cannot be observed experimentally
- Understand protein folding
- Access to thermodynamics quantities (free energies, binding energies, etc)



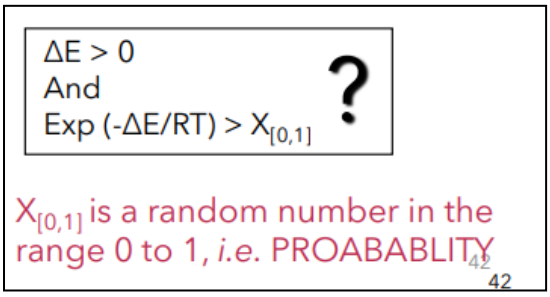
6. Monte-Carlo (Probability concept):

1) Monte Carlo simulation is a simulation that makes use of internally generated (pseudo) random numbers.
2) An important feature of this method is that is that the Monto Carlo scheme can study, sample and calculate energy of all the states.
3) It is a method for sampling the potential energy surface (PES).
4) It allows for uphill moves (but have difficulty in climbing over barriers).
5) It gives exact solutions to statistical mechanical problems.
6) It relies on transition probabilities between different states of the simulated system.
7) These transitions are traced according to the following scheme:

- Generation of an initial configuration.
- Trial of a randomly generated system configuration.
- Evaluation of an acceptance criterion for the trial configuration.
- The acceptance criterion is usually formulated in terms of the potential energy change between trial (new) and existing (old) states and some other properties of the new and old configuration.



- **Acceptance criteria with respect to random number:**



8) The Boltzman factor of energy difference can also be calculated and compared with a random number between 0 and 1.
9) If the random number is lower than the Boltzman factor, the new configuration is accepted.
10) If the random number is higher, the new configuration is rejected and original configuration is used for the next cycle.

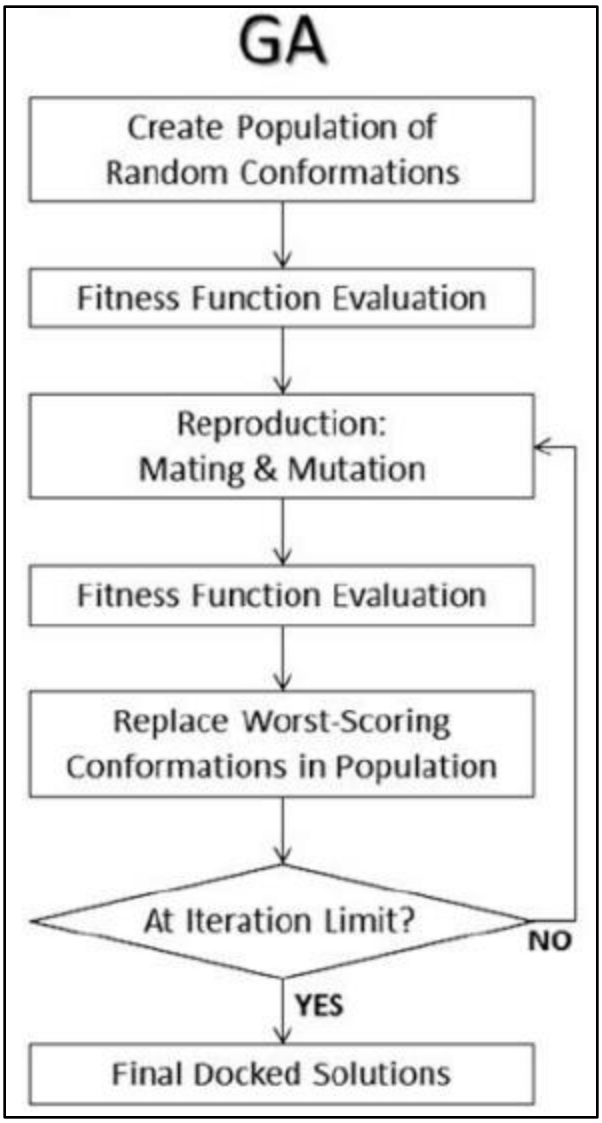
7. Simulated annealing:

1) It helps in climbing configuration.
2) A key feature of annealing is the use of very careful temperature control at the liquid-solid phase transition.
3) The perfect crystal that is eventually obtained corresponds to the global minimum of the free energy.
4) Simulated annealing is a computational method that mimics the process in order to find the 'optimal' or 'best' solutions to problems which have a large number of possible solutions.
5) It solves the hill climbing problem by heating (providing energy) and cooling (i.e. lowest energy).

6) It is a probabilistic technique for approximating the global optimum (least energy) the best solution of a given function.
7) Simulated annealing is a special case where temperature is gradually reduced during the simulation.
8) So, the system is heated and then cooled.
9) If a slow cooling is applied to a liquid, the liquid freezes naturally to a state of minimum energy. This concept is applied here in simulated annealing.
10) At higher temperature, the system is allowed to equilibrate using Molecular dynamics or Monte Carlo simulations.
11) As the temperature falls, it reaches the global minimum energy configuration.
12) It guarantees to find global minima if the equilibrium is infinitely long and cooling is very slow.
13) To confirm if it is a global minimum, repeat the steps.

8. Genetic algorithms:

1) Genetic algorithms (GAs) are a class of optimization methods that are based on various computational models of Darwinian evolution.
2) A genetic algorithm is a large-scale optimization algorithm mimicking a biological evolution in a randomly generated population.
3) A number of conformations form this population.
4) The adaptation is calculated, and a new population is created in accordance to operators (crossover, and mutation).
5) The process is repeated until it converges to a minimum energy structure.
6) Selection of the initial population of conformers analogous to parent is carried out with a statistical bias such that only stable conformations are selected.
7) The chromosome represented by torsional angles or any other parameter may alter due to crossover or mutation resulting in new and diversified conformers.
8) The process may be repeated for as long as it is practically possible.
9) Stable configurations may be formed early and can be lost due to crossover or mutation.
10) To prevent this most programs have an elitist strategy to carry forward the most stable conformations.


GA
Create Population of Random Conformations
Fitness Function Evaluation
Reproduction: Mating & Mutation
Fitness Function Evaluation
Replace Worst-Scoring Conformations in Population
At Iteration Limit?
NO
YES
Final Docked Solutions

(UNIT 1.4)

3D PHARMACOPHORE GENERATION

❖ **Pharmacophore:**

1. Pharmacophore is defined as a specific three dimensional arrangement of the functional molecule, which is crucial to attach or bind to an active site of an enzyme or a molecule.
2. Building a pharmacophore model for a set of active ligands or for a target protein will effectively is used to perceive the interactions between ligand and protein and to facilitate the hit identification process.
3. The concept of pharmacophore was first introduced in 1909 by Ehrlich, who defined the pharmacophore as 'a molecular framework that carries (phoros) the essential features responsible for a drug's (pharmacon) biological activity'.
4. A pharmacophore model is 'an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response'.
5. A Pharmacophore does not represent a concrete molecule, but an abstract concept which describes the common molecular properties of interaction with the receptor.
6. Thus a pharmacophore is a common descriptor shared by a set of active ligands or from the protein binding site.
7. Therefore, the pharmacophore model is able to accurately explain the nature and location of the functional groups of a ligand with the binding site of a target protein.
8. It also provides information of various types of non-covalent interactions and their characteristics.
9. A pharmacophore model is comprised of features that are moieties or regions with specific atoms or bond types in a molecule.
10. These features include hydrophobic, aromatic, hydrogen bond acceptor, hydrogen bond donor, positive and negative ionizable moieties, etc.
11. A pharmacophore model can be established either in a ligand-based manner or in a structure-based manner.
12. Pharmacophore approaches have been used extensively in virtual screening, de novo design and other applications such as lead optimization and multitarget drug design.

Characterization:

1. Location of the functional groups (e.g. proton donor/acceptor, hydrophobic parts).
2. Stabilization of the most effective conformation.

3. Lipinski's rule of five:

   The following properties are essential for good permeation. They are as follows:

- The molecule has less than five proton-donators.
- The molecular weight is smaller than 500 Dalton.
- Lop P smaller than 5.
- The molecule has less acceptors than 10.
- The molecule should use biological transporters otherwise the ligand is attached too strong or it cannot be transported.
- Minimum of pharmacophoric points: 3

❖ **Three-dimensional pharmacophore:**

1. This pattern is derived, manually or computationally, from a three-dimensional molecular model.
2. The pattern is based upon a physical model and binding mechanism.
3. It is sensitive to conformation changes.
4. Better results are obtained when supported by crystal or NMR structural data.
5. It is suitable for lead optimization.

❖ **Types of pharmacophore:**

1. Ligand-based pharmacophore modeling.
2. Structure-based pharmacophore modeling.

❖ **Ligand-based pharmacophore modeling:**

1. Ligand-based pharmacophore modeling has become a key computational strategy for facilitating drug discovery in the absence of a macromolecular target structure.
2. It is usually carried out by extracting common chemical features from 3D structures of a set of known ligands representative of essential interactions between the ligands and a specific macromolecular target.
3. In general, pharmacophore generation from multiple ligands (usually called training set compounds) involves two main steps:

➔ Creating the conformational space for each ligand in the training set to represent conformational flexibility of ligands, and aligning the multiple ligands in the training set and determining the essential common chemical features to construct pharmacophore models.

➔ Handling conformational flexibility of ligands and conducting molecular alignment represent the key techniques and also the main difficulties in ligand-based pharmacophore modeling.

4. Despite the great advances, several key challenges in ligand-based pharmacophore modeling still exist.
5. The first challenging problem is the modeling of ligand flexibility.
6. Currently, two strategies have been used to deal with this problem:

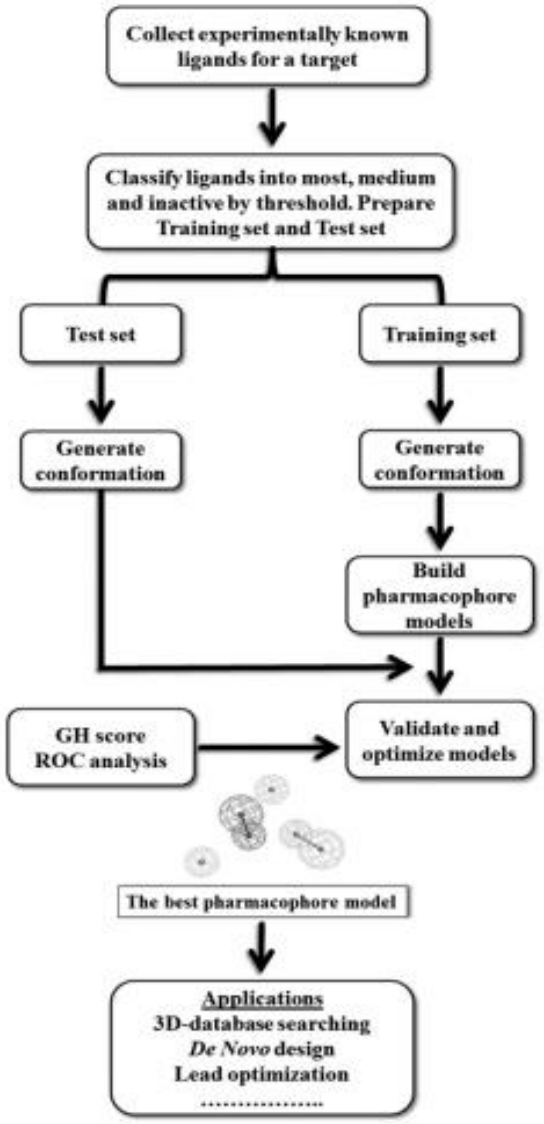
➔ The first is the pre-enumerating method, in which multiple conformations for each molecule are pre-computed and saved in a database.

➔ The second is the on-the-fly method, in which the conformation analysis is carried out in the pharmacophore modeling process.

➔ The first approach has the advantage of lower computing cost for conducting molecular alignment at the expense of a possible need for a mass storage capacity.

➔ The second approach does not need mass storage but might need higher CPU time for conducting rigorous optimization.

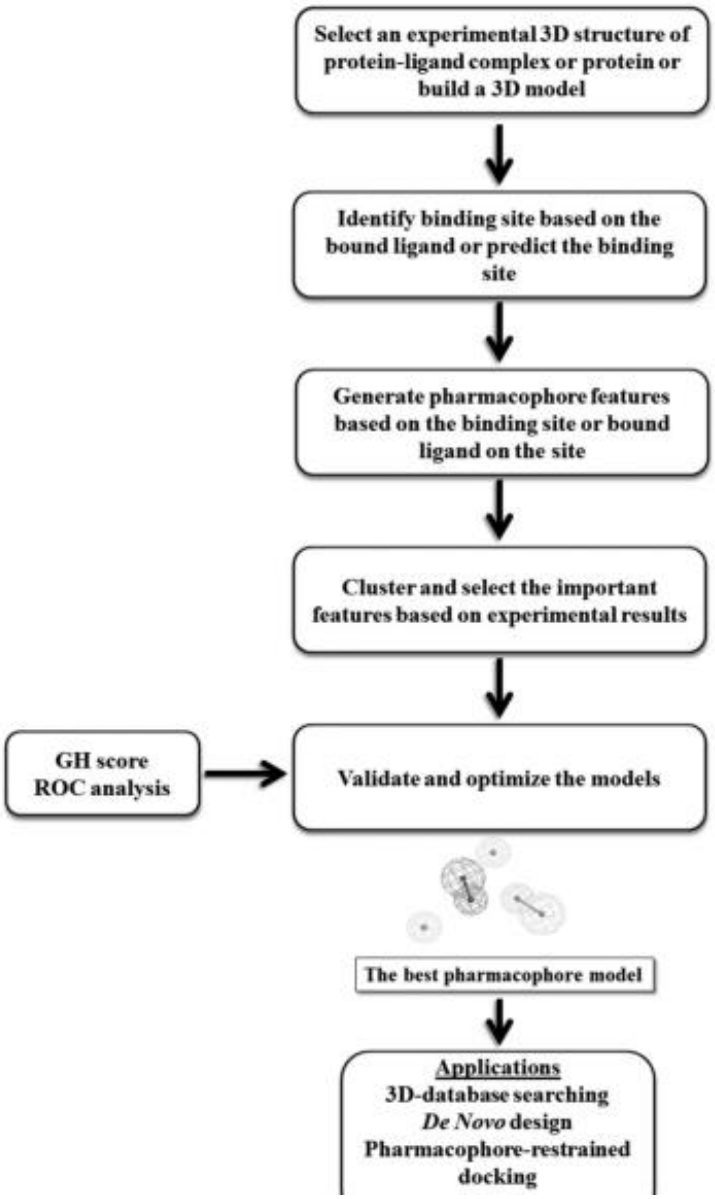
➔ It has been demonstrated that the pre-enumerating method outperforms the on-the-fly calculation approach.



❖ **Structure-based pharmacophore modeling:**

1. Structure-based pharmacophore modeling works directly with the 3D structure of a macromolecular target or a macromolecule– ligand complex.
2. The protocol of structure-based pharmacophore modeling involves an analysis of the complementary chemical features of the active site and their spatial relationships, and a subsequent pharmacophore model assembly with selected features.
3. The structure-based pharmacophore modeling methods can be further classified into two subcategories: Macromolecule (ligand-complex based) and macromolecule (without ligand-based).
4. The macromolecule–ligand-complex-based approach is convenient in locating the ligand-binding site of the macromolecular target and determining the key interaction points between ligands and macromolecule.

5. The limitation of this approach is the need for the 3D structure of macromolecule–ligand complex, implying that it cannot be applied to cases when no compounds targeting the binding site of interest are known.
6. This can be overcome by the macromolecule-based approach.



❖ **Application of pharmacophore models**

1. Pharmacophore model–based
2. Pharmacophore-based *de novo* design

❖ **Pharmacophore-model-based virtual screening:**

1. The computational technique of virtual screening or 3D database screening is considered one of the main techniques and complementary approaches to HTS in identifying hits molecules.

2. Once a pharmacophore model is generated by either the ligand based or the structure-based approach, it can be used for querying the 3D chemical database to search for potential ligands, which is so-called 'pharmacophore-based virtual screening' (VS).
3. Pharmacophore-based VS reduces the problems arising from inadequate consideration of protein flexibility or the use of insufficiently designed or optimized scoring functions by introducing a tolerance radius for each pharmacophoric feature.
4. In the pharmacophore-based VS approach, a pharmacophore hypothesis is taken as a template.
5. The purpose of screening is actually to find such molecules (hits) that have chemical features similar to those of the template.
6. Some of these hits might be similar to known active compounds, but some others might be entirely novel in scaffold.
7. The searching for compounds with different scaffolds, while sharing a biological activity is usually called 'scaffold hopping'.
8. The screening process involves two key techniques and difficulties: handling the conformational flexibility of small molecules and pharmacophore pattern identification.
9. The strategies for handling the flexibility of small molecules in pharmacophore-based VS are very similar to those used in pharmacophore modeling.
10. Again, the flexibility of small molecules is handled by either pre-enumerating multiple conformations for each molecule in the database or conformational sampling at search time.
11. Pharmacophore pattern identification, usually called 'substructure searching', is actually to check whether a query pharmacophore is present in a given conformer of a molecule.

❖ **Pharmacophore-based *de novo* design:**

1. Another application of pharmacophore is de novo design of ligands.
2. The compounds obtained from pharmacophore-based VS are usually existing chemicals, which might be patent protected.
3. In contrast to pharmacophore-based VS, the de novo design approach can be used to create completely novel candidate structures that conform to the requirements of a given pharmacophore.
4. The pharmacophore-based de novo design can guide medicinal chemists by suggesting chemical modification to the synthesized compounds during lead optimization.
5. This is particularly employed in the absence of 3D structure of protein target.
6. In contrast to pharmacophore-based virtual screening, the de novo approach may result in completely new and/or novel chemotypes.

7. In this approach, diverse and synthetically accessible fragments are first collected based on the pharmacophore.
8. Then, new molecules are constructed by using fragments under the consideration of the 3D pharmacophore
9. Therefore, pharmacophore-based de novo design shows a unique advantage in building completely novel hit compounds.

❖ **Advantages and limitations of pharmacophore:**

1. Like any other approach, pharmacophore has both advantages and disadvantages.
2. The major advantages and limitations are as follows:

Advantages:

1. Pharmacophore models can be used for VS on a large database.
2. There is no need to know the binding site of the ligands in the macromolecular
3. Target protein, although this is true only for LBP modeling.
4. It can be used for the design, optimization of drugs, and scaffolds hopping.
5. It can conceptually be obtained even for 2D structural representation.
6. This approach is comprehensive and editable.
7. By adding or omitting chemical feature constraints, information can be easily traced to its source.

Limitations:

1. 2D pharmacophore is faster but less accurate than 3D pharmacophore.
2. A pharmacophore is based only on the ligand structure and conformation.
3. No interactions with the proteins are integrated. It is interesting to point out that in this case, SBP modeling can be used to solve the problem.
4. It is sensitive to physicochemical features.

(UNIT 2)

INTRODUCTION TO MOLECULAR DESCRIPTORS

❖ **Molecular descriptors:**

1. The molecular descriptor is the final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment.
2. They are the numerical values designed to capture the structure and properties of a molecule.
3. To make a reasonable prediction for any set of molecules, the physical or biological data must be related to the molecule through a series of descriptors.
4. These descriptors can be structural, relating data about the relative position of atoms and types, or calculated data such as electron density using quantum chemical methods.
5. Descriptors should be able to give possible characteristics of molecule.
6. It should be able to quantify chemical changes.

❖ **Types of molecular descriptors:**

Descriptors can be classified by the following representations:

Molecular representation	**Examples**
0D	Atom types, molecular weight, bond types (i.e., constitutional descriptors, count descriptors)
1D	Count of atom types, counts of hydrogen bond donors or acceptors, number of rings, number of functional groups by type (i.e., list of structural fragments, fingerprints)
2D	Mathematical representation by graph theory or calculated values such as lipophilicity or topological polar surface area. (i.e., graph invariants)
3D	Geometrical descriptors or polar surface area. (i.e., quantum-chemical descriptors, size, steric, surface and volume)
4D	GRID or CoMFA methods, Volsurf

0D molecular descriptors:

1. 0D depend on the atom types that conform the molecule and their bonds.
2. The total number of carbon, nitrogen, oxygen or halogen atoms can potentially adequately describe a molecule.

1D molecular descriptors:

1. In addition to the types of atoms present, molecules can be further represented by bonding or bonding fragments.
2. 1D only depend on the functional group types.
3. Other functional groups can also be used to adequately describe a molecule by similarity.

2D molecular descriptors:

1. 2D descriptors consider the molecular representation through a chemical graph.
2. A graph is an abstract structure that contains nodes connected by edges.
3. In representing molecules nodes are the atoms, and edges are the bonds.
4. Hydrogen atoms are usually omitted and thus called "hydrogen depleted molecular graphs."

3D molecular descriptors:

1. 3D descriptors contain the 3D geometric information of a molecule that is its molecular surface area, polar surface area, volume, excluded volume and chiral centers.
2. Three-dimensional representation of molecules provides a more accurate description of molecular dimensionality.

❖ **What should a descriptor be like?**

1. It should have structural interpretation.
2. It should have a good correlation with at least one property.
3. It should preferably discriminate among isomers.
4. It should be possible to apply to local structure.
5. It should be possible to generalize to "higher" descriptors.
6. It should be simple.
7. It should not be based on experimental properties.
8. It should not be trivially related to other descriptors.
9. It should be possible to construct efficiently.
10. It should use familiar structural concepts.
11. It should change gradually with gradual change in structures.
12. It should have the correct size dependence, if related to the molecular size.

❖ **Physicochemical Properties of Descriptors:**

1. Hydrophobicity
   - ➔ Partition coefficient (P)
   - ➔ Substituent hydrophobicity constant (π)
2. Electronic effects

3. Steric factors
   - ➔ Taft's steric factor(Es)
   - ➔ Molar Refractivity (MR)

❖ **Lipophilicity/Hydrophobicity:**

1. An important physicochemical property for a drug compound is its Lipophilicity.
2. Lipophilicity plays a crucial role in determining ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties and the overall suitability of drug candidates.
3. There is increasing evidence to suggest that control of physicochemical properties such as lipophilicity, within a defined optimal range, can improve compound quality and the likelihood of therapeutic success.
4. Lipophilicity is calculated using a collection of training data-set available for molecules and fragments that contribute to sub-structures and functional groups of molecules.
5. The more lipophilic a molecule is, the more soluble it is in lipophilic organic phase.
6. For the same reason drug penetration into a biological membrane is also influenced by the lipophilicity of the drug.
7. Most drugs diffuse through the cell membrane.
8. To diffuse through a cell membrane, a drug must be lipophilic, that is uncharged and non-polar.
9. It dissolves in non-polar liquids such as oil according to the 'like dissolves like' principle.
10. Such drugs are termed 'fat-loving'.
11. Since, bio-membranes are composed of fatty acids they represent a lipophilic barrier through which hydrophilic molecules are unable to diffuse.
12. Whether a drug is lipophilic or hydrophobic, it has a great effect on its pharmacokinetic properties, especially regarding its distribution, metabolism and excretion.
13. Therefore, drugs in their active form are usually lipophilic.
14. They penetrate the lipid bilayer of most cellular membrane, and thus have a good absorption.
15. Therefore, lipophilicity can be defined as the affinity of a drug for a lipid environment.

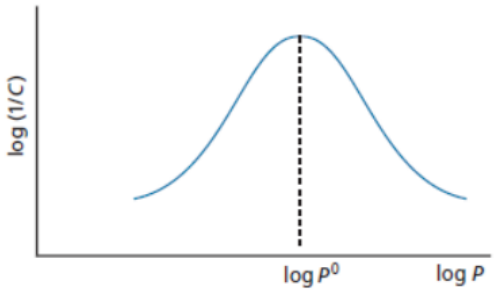
❖ **Partition coefficient (P):**

1. A partition coefficient is a measure of drugs lipophilicity and an indication of its ability to cross the cell membrane.

The partition coefficient (P)

$$P = \frac{\text{Concentration of drug in octanol}}{\text{Concentration of drug in aqueous solution}}$$

High P value : Hydrophobic compounds
Low P value : Hydrophilic compounds



- Parabolic curve
- Optimum value for Log P exists
- Beyond Log P^o-
  increasing hydrophobicity
  decreases the activity

2. The hydrophobic character of a drug can be measured experimentally by testing the drug's relative distribution in an n -octanol/water mixture.
3. Hydrophobic molecules will prefer to dissolve in the n-octanol layer of this two-phase system, whereas hydrophilic molecules will prefer the aqueous layer.
4. The relative distribution is known as the partition coefficient (P).
5. Hydrophobic compounds have a high P value, whereas hydrophilic compounds have a low P value.

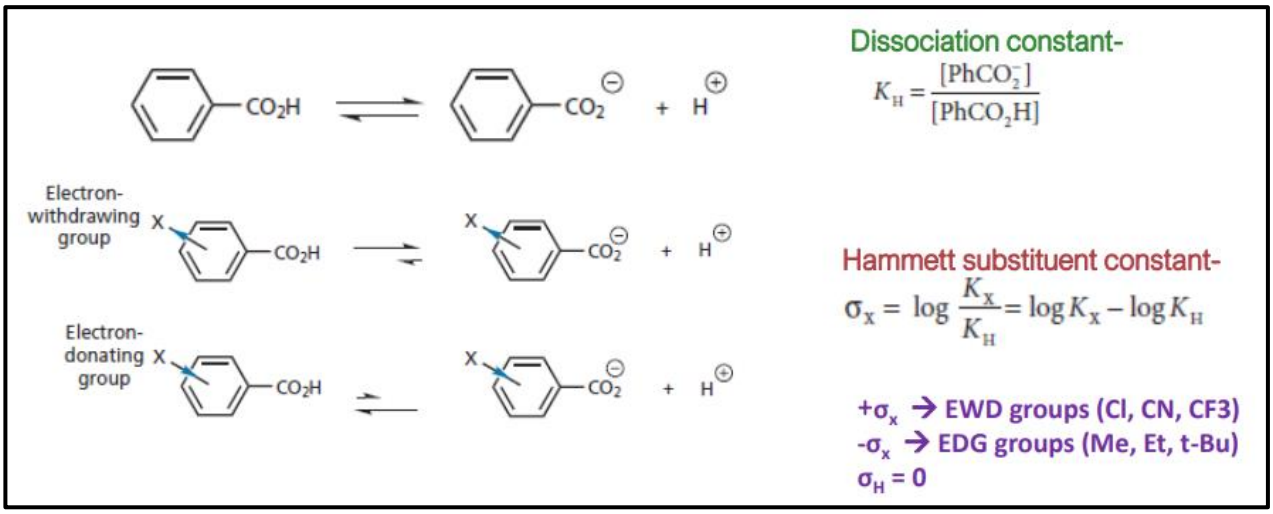
❖ **Linear Free Energy Relationships:**

1. Linear Free Energy Relationships allow a correlation of substituents with a reaction rate, biological activity, pKa, etc.

❖ **Hammett substitution constant (σ):**

1. In 1935, Prof. Louis Hammett developed an electric constant, now known as 'Hammett constant' for a substituent based on the Pka values.
2. Hammett substitution constant is a measure of the electron-withdrawing or electron-donating ability of a substituent.
3. The Hammett substituent constant takes into account both resonance and inductive effects.
4. It has been determined by measuring the dissociation of a series of substituted benzoic acids compared with the dissociation of benzoic acid itself.
5. The Hammett equation describes a free-energy relationship relating reaction rates and equilibrium constants for many reactions, involving benzoic acid derivatives with meta- and para-substituents to each other with just two parameters: a substituent constant and a reaction constant.

6. Louis Hammett correlated electronic properties of organic acids and bases with their equilibrium constants and reactivity.
7. This electronic effect of various substituents will clearly have an effect on drug ionization and polarity.
8. Example: σ for aromatic substituents is measured by comparing the dissociation constants of substituted benzoic acids with benzoic acid.



- Benzoic acid is a weak acid and only partially ionizes in water
- Equilibrium is set up between the ionized and non-ionized forms, where the relative proportion of these species is known as the equilibrium or dissociation constant.
- When a substituent is present on the aromatic ring, this equilibrium is affected.
- Electron-withdrawing groups, such as a nitro group, result in the aromatic ring having a stronger electron-withdrawing and stabilizing influence on the carboxylate anion, and so the equilibrium will shift more to the ionized form.
- Therefore more on the ionised form of the drug will be formed.
- Hence the equilibrium constant (K)
- K = [ionised]/[unionised] will be high
- If the substituent X is an electron-donating group such as an alkyl group, then the aromatic ring is less able to stabilize the carboxylate ion since the H+ will prefer to be attached to the electron rich COO- and therefore the COOH will be dominate and the value of K will be less.
- The equilibrium shifts to the left indicating a weaker acid with a smaller K value.

9. Dissociation constant-

$$K_H = \frac{[PhCO_2^-]}{[PhCO_2H]}$$

where,

K_H – Dissociation constant for Hydrogen

10. Hammett substitution constant-

$$\sigma_x = \log \frac{K_X}{K_H} = \log K_X - \log K_H$$

where,

σ X is the Hammett constant of the substituent X.

K_H – Dissociation constant for Hydrogen

K_X – Dissociation constant for Substituent

11. Benzoic acids containing electron-withdrawing substituents will have larger K_X values than benzoic acid itself (K_H) and, therefore, the value of σX for an electron-withdrawing substituent will be positive.
12. Substituents such as Cl, CN, or CF3 have positive σ values.
13. Benzoic acids containing electron-donating substituents will have smaller K_X values than benzoic acid itself and, hence, the value of σ X for an electron-donating substituent will be negative.
14. Substituents such as Me, Et, and t -Bu have negative values of σ.
15. The Hammett substituent constant for H is zero.
16. The value of σ for a particular substituent will depend on whether the substituent is meta or para.
17. σ can only be used for aromatic substituents.
18. Traditionally, the Hammett Linear Free Energy Relationship is written as:

$$\log\left(\frac{k_X}{k_H}\right) = \rho\sigma$$

k_X – rate of reaction with substituent X

k_H – rate of reaction with no substituent (hydrogen)

σ – substituent constant, relative pK_a of X-substituted benzoic acid vs. benzoic acid:

19. The reaction constant, or sensitivity constant, ρ describes the susceptibility of the reaction to substituents, compared to the ionization of benzoic acid.
20. It is equivalent to the slope of the Hammett plot.
21. If the value of:

➔ ρ>1: the reaction is more sensitive to substituents than benzoic acid and negative charge is built during the reaction.

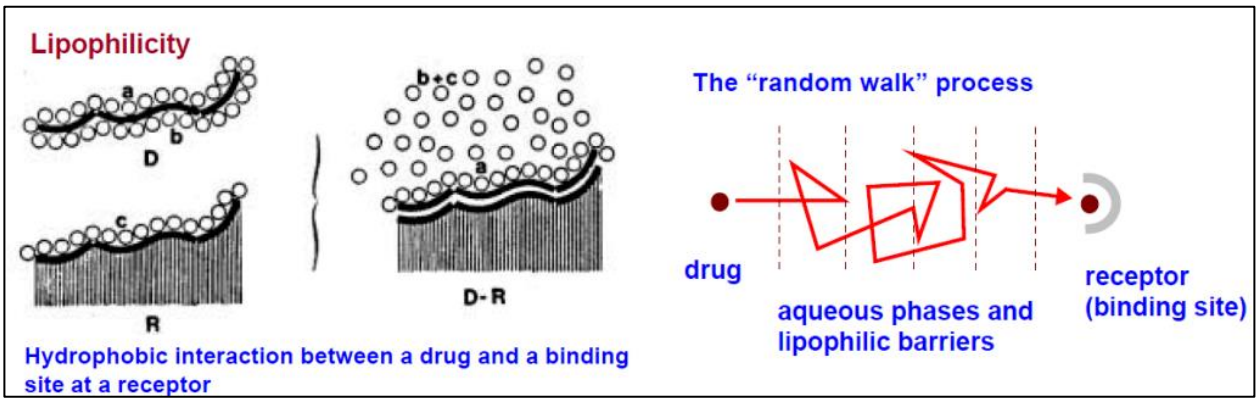
➔ 0<ρ<1: the reaction is less sensitive to substituents than benzoic acid and negative charge is built.

➔ ρ=0: no sensitivity to substituents and no charge are built.

➔ ρ<0: the reaction builds positive charge.

❖ **Lipophilicity effects:**

1. Just as the Hammett equation relates the electronic effects of substituents to reactions rates, Hansch believed that a linear free-energy relationship should exist for lipophilicity and biological activity.
2. Hansch suggested that the drug in the aqueous phase surrounding the cell made a random walk through the cell membrane, which is lipophilic, to interact with a particular site in the cell, the rate of which is dependent on the structure of the drug.
3. In 1969, Hansch developed an equation that related biological activity to certain electronic characteristics and the hydrophobicity of a set of structures.



4. As a measure of lipophilicity, Hansch proposed the partition coefficient, P as

$$P = [\text{compound}]_{oct}/[\text{compound}]_{aq}\ (1 - \alpha)$$

where, α is the degree of dissociation of the compound in water calculated from the ionization constant.

If P < 1, it means that the compound is more soluble in water.

If P >1, it means that the compound is more soluble in octanol.

❖ **Why n-Octanol is used?**

1. The partition coefficient of a molecule observed in a water– *n* -octanol system has been adopted as the standard measure of lipophilicity.
2. n-Octanol is used because of similarity with biological cell membrane system having long alkyl chain and polar hydroxyl groups.

3. Thus, it has a membrane analogous structure.
4. It has a hydrogen bond donor and acceptor.
5. It is practically insoluble in water.
6. There is no desolvation on transfer into organic phase.
7. It has a very low vapor pressure.
8. It is transparent in the UV region.

❖ **LogP vs LogD:**

1. The partition coefficient (P) is a measure of lipophilicity and can be defined by the distribution of a drug between the organic phase, which is generally n-octanol pre-saturated with water, and the aqueous phase, which is generally water pre-saturated with n-octanol.
2. According to Lipinski's Rule of 5, an oral drug should have a LogP value <5, ideally between 1.35-1.8 for good oral and intestinal absorption.
3. LogP is a critical measure that not only determines how well a drug will be absorbed, transported, and distributed in the body but also dictates how a drug should be formulated and dosed.
4. For example, a drug with low aqueous solubility and high lipophilicity (high positive LogP) will be compromised in bioavailability.
5. The distribution coefficient (D) is the ratio of the sum of the concentrations of all species of the compound in octanol to the sum of the concentrations of all species of the compound in water.

$$\text{P = Partition Coefficient} = \frac{\text{Concentration of neutral species dissolved in partition solvent}}{\text{Concentration of neutral species dissolved in water}}$$

$$\text{D = Distribution Coefficient} = \frac{\text{Concentration of all species dissolved in partition solvent}}{\text{Concentration of all species dissolved in water}}$$

6. Based on acidic/basic dissociation reactions, the concept of a partition coefficient for cationic and anionic species and for neutral species can be understood.
7. Log P will not give any information about charged or ionic species.
8. It will only give information about neutral species.
9. Whereas, log D will give information about all the species.

$$\log D \cong \log P$$ >>> For unionised molecules

$$\log D \cong \log P + \log (f^0)$$ >>> at a give pH. f^0 is mole fraction of the un-ionized form

$$\log D_{acids} \cong log\ P + pK_a - pH$$ >>> When $pH - pK_a > 1$ for acids

$$\log D_{acids} \cong log\ P + pK_a + pH$$ >>> When $pK_a - pH > 1$ for bases

❖ **Calculation of Log P:**

1. Log P calculation is atom-based as it takes into consideration every atom summation.
2. Log P for a molecule can be calculated from a sum of fragmental or atom based terms plus various corrections.
3. Increase in log P indicates an increase in binding of drugs.
4. Binding is greater for hydrophobic drugs.

- Atom-based

$$\log P = \sum_i a_i f_i + \sum_j b_j f_j$$

❖ **Substitution constant (π):**

1. Partition coefficients can be calculated by knowing the contribution that various substituents make to hydrophobicity.
2. This contribution is known as the substituent hydrophobicity constant (π) and is a measure of how hydrophobic a substituent is relative to hydrogen.
3. Hansch derived substitution constants for the contribution of individual atoms and groups to the partition coefficient.
4. The partition coefficient of a molecule can be calculated by knowing the sum of contribution from various parts of the molecule.
5. A measure of a substituent's hydrophobicity relative to hydrogen.
6. The hydrophobicity constant π for substituent X is given by:

$$\pi X = \log Px - \log PH = \log Px / PH$$

where, πX is the hydrophobic constant of the substituent X.

Px is the partition coefficient of the compound with substituent X and PH is for the parent compound (X = H).

This means πH = 0

7. A positive value of π indicates that the substituent is more hydrophobic than hydrogen.
8. A negative value indicates that the substituent is less hydrophobic.
9. These π values are characteristic for the substituent and can be used to calculate how the partition coefficient of a drug would be affected if these substituents were present.
10. π identify specific regions of the molecule which might interact with hydrophobic regions in the binding sites.
11. πX is the hydrophobic constant of the substituent X.

12. Px is the partition coefficient of the compound with substituent X and PH is for the parent compound (X = H).
13. This means πH = 0
14. π is both:

➔ Additive (multiple substituents exert an influence equal to the sum of the individual substituents).

➔ Constitutive (effect of a substituent may differ depending on the molecule to which it is attached).

❖ **P versus π:**

1. Lipophilic parameters define partitioning of compound between the aqueous and non-aqueous phase.
2. 2 parameters are commonly used to represent lipophilicity which are:

➔ Partition coefficient (P): refers to whole molecules

➔ Substitution constant (π): refers to substituted groups

3. The partition coefficient P is a measure of the drug's overall hydrophobicity and is, therefore, an important measure of how efficiently a drug is transported to its target and bound to its binding site.
4. The π factor measures the hydrophobicity of a specific region on the drug's skeleton.
5. π is the effect of a given substituent on log p of basic skeleton.

$$\Pi = \log Px\text{-}PH$$

6. The π value for the compound is sum of π values of each of separate substituents.
7. P and π are not exactly equivalent.
8. Different equations with different constants would be obtained.
9. Partition coefficient (P) aids in drug development whereas substitution constant aids in drug design.

❖ **Steric effects:**

1. Steric effects are nonbonding interactions that influence the shape (conformation) and reactivity of ions and molecules.
2. Steric effects relate to the bulk, size and shape of a drug.
3. Steric effects complement electronic effects, which usually dictate shape and reactivity.
4. Steric effects are more difficult to quantify compared to electronic and lipophilic factors.
5. The steric factors will influence the ability of the drug to both, approach and interact with a binding site.

Steric factors disadvantages of a bulky group:

- A bulky substituent may act like a shield and hinder the ideal interaction between a drug and its binding site.

Steric factors advantages of a bulky group:

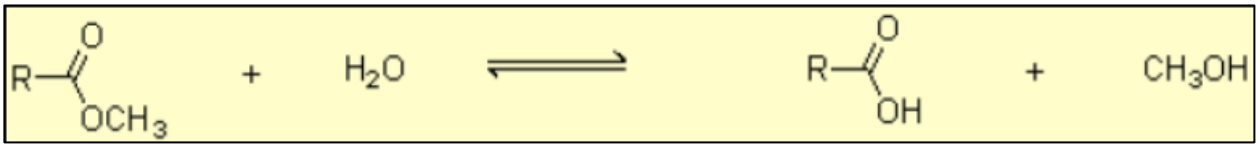
- Alternatively, a bulky substituent may help to orientate a drug properly for maximum binding and increase activity.

❖ **Steric factor descriptors:**

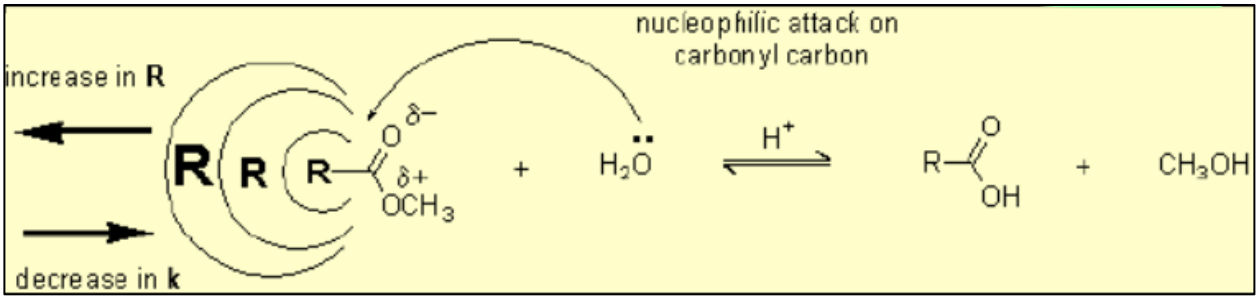
- Taft's steric factor (Es)
- Molar refractivity
- Verloop steric parameter

❖ **Steric effects: Taft's Equation**

1. Since interaction of a drug with its receptor brings together two molecules, steric effects come into play.
2. Steric effects relate to the bulk, size and shape of a drug.
3. The steric factors will influence the ability of the drug to both, approach and interact with a binding site.
4. As the drug is more lipophilic, it enters CNS more rapidly, rapid onset and duration of action, because of bulky substituent's it needs time to orient in favorable conformation and bulky substituent's also delays detachment of drug from receptor, which leads to late onset and duration of action.
5. Size, shape and bulk of drug influences the ease with which it can approach, bind and interact with the target.
6. Bulky substituent may help to orient a drug properly for maximum binding and increase activity.
7. Taft equation was developed by Robert W. Taft in 1952 as a modification to the Hammett equation.
8. While the Hammett equation accounts for how field, inductive, and resonance effects influence reaction rates, the Taft equation also describes the steric effects of a substituent.
9. The value for Es can be obtained by comparing the rates of hydrolysis of substituted aliphatic esters against a standard ester under acidic conditions.



10. Here, the size of R affects the rate of reaction by blocking nucleophilic attack by water.



11. Taft derived the steric parameter Es as:

12. $$E_s = \log k_X - \log k_o = \log k_x/k_o$$

where, kx represents the rate of hydrolysis of an aliphatic ester bearing the substituent X and ko represents the rate of hydrolysis of the reference ester.

13. The Taft steric constant (Es) indicates the size contribution of substituents on a parent molecule.

❖ **Steric effects: Molar Refractivity (MR)**

1. Molar refractivity is a term which represents size and polarizability of a fragment or a molecule as it is closely related with molar properties and refractive index of particular substance which is being tested.
2. It characterizes bulk of molecule or substituent but not shape.
3. Molar volume is corrected by Refractive Index.
4. Molar Refractive (MR) is a measure of molar volume corrected by refractive index independent of physical factors and is useful in differentiating the structurally different compounds.

$$MR = \left[\frac{(n^2 - 1)}{(n^2 + 2)}\right]\left(\frac{MW}{d}\right)$$

n → refractive index
MW → Molecular weight
d → density

5. The term MW/d defines a volume.
6. The $(n^2\text{-}n^1)/(n^2+n^1)$ term provides a correction factor by defining how easily the substituent can be polarized.
7. The greater the positive MR value of substituent the larger its steric or bulk effect and substituent binds to polar surface while a negative value indicates steric hindrance at binding site.
8. The basic idea behind the use of such descriptor is that similar changes in the structure are likely to produce similar changes in reactivity ionization, and binding.

❖ **Graph theory:**

1. Graphs are nodes (vertices) and Edges.

Nodes –> atoms

Edges – > bond

2. Nodes (Atoms) + Edges (bonds) = Topological Graph (2D).
3. In graph theory, a graph carries no geometric information.
4. It only tells about the connectivity.

❖ **Matrix Representation:**

1. Graphs can be easily represented in a "Matrix" form.
2. The idea behind the molecular graph representation lies in mapping the atoms and bonds that make up a molecule into sets of nodes and edges.
3. The atoms in a molecule are represented as nodes and the bonds as edges.

Advantages:

1. The molecular graph is completely coded (each atom and bond is represented).
2. Matrix algebra can be used.

Disadvantages:

1. The number of entries in the matrix grows with the square of the number of atoms.
2. No stereochemistry is included.

Types of matrix representations:

1. **Adjacency matrix:**
   - Adjacency matrix describes connections of atoms.
   - It contains only 0 and 1 (bits).
   - Bond types and bond orders are not present.
   - It does not contain number of free electrons.
2. **Distance matrix:**
   - Distance matrix describes geometric distances.
   - The shortest distance between atoms is considered.
   - Bond types and bond orders are not present.
   - It does not contain number of free electrons.
   - It cannot be represented by bits.
3. **Incidence matrix:**
   - Incidence matrix describes connections and bonds.
   - It contains only 0 and 1(bits).

- Bond types and bond orders are not present.
- It does not contain number of electrons.

4. **Bond matrix:**
   - Bond matrix describes connections and bond orders of atoms.
   - It does not contain number of free electrons.
   - It cannot be represented by bits.
5. **Bond-electron matrix:**
   - Bond-electron matrix describes connections.
   - Bond orders and valence electrons of the atoms are taken into consideration.
   - It cannot be represented by bits.

❖ **Topological Descriptors:**

1. 2D molecular descriptors are based on topological indices.
2. These are based on graph theory and relate to overall topology, dictated by the way in which atoms are connected to each other
3. Simple calculations are done without considering physicochemical properties.
4. Topological descriptors are not rigorous like the Quantum Chemical descriptors.
5. Topological indices are computed by applying a specific algorithm using information obtained from the hydrogen-suppressed graph, that is number of elements defining it and their connectivity information.
6. These are single-valued descriptors that can be calculated from the 2D graph representation of molecules.

❖ **2D Molecular Descriptors:**

1. Mathematical notations provide a method for describing chemical structures, and allow for computational processing of molecules in a data set.
2. These are essentially 2D descriptors.
3. A graph is an abstract structure that contains nodes connected by edges.
4. In representing molecules nodes are the atoms, and edges are the bonds.
5. Hydrogen atoms are usually omitted and thus called “hydrogen depleted molecular graphs.”
6. Descriptors based on the molecular graph representation are widely used because they incorporate precious chemical information:
   - size,
   - degree of branching,

- neighborhood of atoms → electronic & steric effects,
- flexibility
- overall shape

❖ **Weiner Index:**

1. It is one of the first mathematical representations of chemical structure used for prediction of properties was developed in 1947 by Harold Weiner.
2. It is defined at the sum of distances between any two carbon atoms (pairs of nodes) in the molecule.
3. Mathematically it is represented as:

$$W = \sum_{b=1}^{B} N_{i,b} \cdot N_{j,b}$$ Wiener Index

$$W = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} D_{ij}$$ Hosoya Modification

where, Ni,b and Nj,b are the number of vertices on each side of the bond b, including vertices i and j, respectively, and B is the total number of graph edges

4. It involves counting the number of bonds between each pair of atoms and summing the distances between all such pairs –> bond additive index.
5. It is obtained from a distance matrix, also known as Hosoya Modification.

❖ **Zagreb Index:**

1. For each non-hydrogen atom, add up the squares of the number of corrections to other non-hydrogen atoms regardless of bond order.

$$M_1 = \sum_{\text{vertices}} d_i^2$$ ***first Zagreb index***

- d_i = the degree of a vertex *i*

$$M_2 = \sum_{\text{edges}} d_i \cdot d_j$$ ***second Zagreb index***

- $d_i\, d_j$ = the degree of a edge *ij*

2. In these indices one counts the connections from each vertex (node, carbon).
3. The first Zagreb index M (G) is equal to the sum of squares of the degrees of the vertices, and the second Zagreb index M (G) is equal to the sum of the products of the degrees of pairs of adjacent vertices of the underlying molecular graph G.

❖ **Randic's Connectivity Index:**

1. Randic's connectivity index is the best-known branching index.

2. It is calculated from the H-suppressed graph representation of a molecule and is based on the degree δ of each atoms.

$$\textbf{branching index} = \sum_{\textbf{bonds}} {}^{1}\!/\!_{\sqrt{\delta_i \delta_j}}$$

❖ **Keir and Hall's Connectivity Index:**

1. It is a modification of the Randic' connectivity index.
2. It is also, known as chi molecular connectivity indices.
3. Simple δ value, is the number of sigma electrons and the number of H-atoms.

$$\delta_i = \sigma_i - h_i$$

4. Valence δ v value, is the number of sigma, pi and lone pair electrons.

$$\delta_i^v = Z_i^v - h_i$$

5. χ molecular connectivity indices are the sequential indices that sum the atomic δ values over bond paths of different lengths.
6. Zeroth order χ index summation over all atoms in a molecule with path length zero.

$${}^0\chi = \sum_{\text{atoms}} \frac{1}{\sqrt{\delta_i}}; \quad {}^0\chi^v = \sum_{\text{atoms}} \frac{1}{\sqrt{\delta_i^v}}$$

7. First order χ index involves summation of bonds (same as Randic' branching index)

$${}^1\chi = \sum_{\text{bonds}} \frac{1}{\sqrt{\delta_i \delta_j}}; \quad {}^1\chi^v = \sum_{\text{bonds}} \frac{1}{\sqrt{\delta_i^v \delta_j^v}}$$

❖ **Electrotopological State (E-state) Index:**

1. E-state indices are computed for each atom (sometimes including H-atoms).
2. It depends on the intrinsic state (electronic and topological properties) of an atoms (its position in periodic table
3. E-state indices tell only about the topology with respect to electrons.

❖ **Descriptors based on 3D representation:**

1. 3D descriptors contain the 3D geometric information of a molecule.

2. Three-dimensional representation of molecules provides a more accurate description of molecular dimensionality.

- **STERIMOL Parameters:**

➔ It is a multi-parametric method for characterizing the steric features of the substituents in more complex biological systems by Verloop Steric parameter.

➔ It describes molecular reactivity.

➔ STERIMOL parameters are a set of five descriptors (L, B1, B2, B3, and B4).

- L is the length of the substituent along the axis of the bond between the first atom of the substituent and the parent molecule.
- Width parameters B1-B4 are all orthogonal to L and form angles of 90 to each other.

- **Quantum Chemical Descriptors:**

➔ Energies of the HOMO (Highest Occupied Molecular Orbitals) and LUMO (Lowest Unoccupied Molecular Orbitals) are very popular quantum chemical descriptors.

➔ It describes Molecular Reactivity.

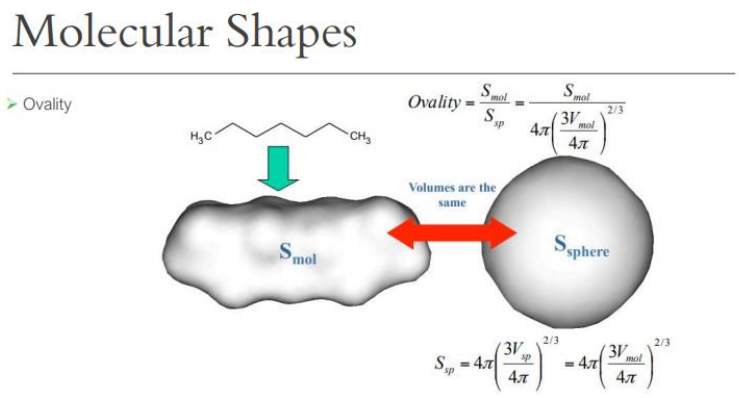
➔ It shows the electron distribution (Molecular electrostatic potential).

➔ The energy of the HOMO is directly related to the ionization potential and characterizes the susceptibility of the molecule toward attack by electrophiles.

➔ The energy of the LUMO is directly related to the electron affinity and characterizes the susceptibility of the molecule toward attack by nucleophiles.

➔ Both the HOMO and the LUMO energies are important in radical reactions

- **Molecular Shapes:**



- **Surface Area:**

➔ It signifies the molecular surface, solvent accessible surface and Van der Waals surface.

- **Surface Polarity (Polar Surface Area):**

➔ Total Surface area occupied by Polar Atoms

- **Field intensity descriptors:**

➔ It is probe atom (pseudo-receptor) based.

QSAR PARAMETERS

Various parameters used in QSAR studies are:

- **Lipophilic Parameters:** partition coefficient, molar refractivity
- **Electronic Parameters:** Hammet constant
- **Steric Parameters:** Taft's constant, verloop steric parameter

- Π : octanol/water partition coefficient, a measure of the hydrophobic bonding power of the drug.
- σ : Hammett substituent constant, which is a measure of the electronic effect on the rate of reaction.
- Es : Taft steric parameter

(UNIT 3)

METHODS TO UNDERSTAND MOLECULAR SIMILARITIES

❖ **Why use similarity measures?**

1. To find a molecule in a database.
2. To screen a database for a particular substructure.
3. To assess whether a new structure is unique.
4. To rank a set of molecules.
5. To cluster a database of molecules.
6. To build a diverse set of molecules for synthesis/testing.

❖ **Pharmacophore vs. Similarity searching**

Substructure/Pharmacophore searching	Similarity searching
• Specify a query which is used to search a database that identifies compounds that meet the query	• The search is to find molecules in the database similar to the query.
• This requires a set of active molecules	• A single compound is sufficient
• The search partitions the database into two types of molecules – those that satisfy the query and those that do not.	• The molecules in the database are ranked in decreasing similarity
• User has no control on the output, as a result the output may be extensive or sparse	• The ranking helps the user to control the size of the output, by setting a threshold.

❖ **How do we access the similarity between two molecules?**

1. A similarity searching method has two components:

➔ A set of numerical descriptors that can be used to compare two molecules.

➔ A similarity coefficient that can quantify the degree of similarity.

❖ **Similarity Measure:**

1. **Similarity** is estimated by the number of matches or overlap between two objects.
2. **Dissimilarity** is estimated by the number of mismatches or difference between two objects.

❖ **Similarity based on 2D fingerprints:**

1. 2D Similarity is commonly based on “fingerprints”.
2. Fingerprints are binary vectors where each bit indicates the presence (“1”) or absence (“0”) of a particular substructure within a molecule.
3. There are 2 types of fingerprints:

➔ Structural keys (fragment-based fingerprints)

➔ Pharmacophore-based fingerprints

4. Similarity can be based on continuous whole molecule properties, e.g., log P, molar refractivity, and topological indexes.
5. Usual approach is to use a distance coefficient, such as Euclidean coefficients.

❖ **Hashed fingerprints:**

1. In hashed fingerprints, there is no set dictionary or 1:1 relationship between bits and features.
2. All possible fragments in a compound are generated.
3. The number of fragments represented can be huge.
4. Up to a given a bond number, all linear paths (linear patterns) consisting bonds and atoms of a structure are detected.
5. All cycles (cyclic patterns) are also detected.
6. Using a proprietary hashing method, a given number of bits in the bit string are set for each pattern. It is possible, that the same bit is set by multiple patterns. This phenomenon is called bit collision.
7. Few bit collisions in the fingerprint is tolerable, but too many may result in losing information in the fingerprint.
8. Once fingerprint representations are available, similarity coefficients can be used to give a measure of similarity between two fingerprints.

❖ **Similarity Coefficients:**

Popular Similarity/Distance Coefficients

- Similarity metrics:
  - Tanimoto coefficient
  - Dice coefficient
  - Cosine coefficient
- Distance metrics:
  - Euclidean distance
  - Hamming distance
  - Soergel distance

1. Tanimoto Coefficient:

➔ The similarity between two molecules represented as binary fingerprints is most frequently quantified by the Tanimoto coefficient.

➔ Tanimoto coefficient is most widely used for binary fingerprints.

➔ **For binary variables**, the Tanimoto similarity between two molecules A and B represented, the coefficient is given by:

$$S_{AB} = \frac{c}{a+b-c}$$

➔ Absent features are not taken into account.

➔ Range: 0 to 1

➔ Eg: For 'a' bits set to 1 in molecule A; 'b' bits set to 1 in molecule B and 'c' bits of 1 common to A and B.

A	1	0	1	1	1	0	1	1	0	0	1	1	a = 8
													c = 5
B	0	0	1	1	0	0	1	0	1	0	1	1	b = 6

$$S_{AB} = \frac{5}{8+6-5} = 0.56$$

➔ **For continuous data**, Tanimoto coefficient is also known as Jaccard coefficient.

➔ Range: -0.33 to + 1.

2. **Hodgkin index:**

➔ Hodgkin index is also known as dice coefficient.

➔ It is monotonic with the Tanimoto coefficient.

➔ Dice coefficient (Hodgkin index) **for continuous variables**,

Range: -1 to +1

➔ **For binary variables**, the coefficient is:

$$S_{AB} = \frac{2c}{a+b}$$

Range: 0 to 1

3. **Carbo Index:**

➔ Carbo index is also known as cosine coefficient.

➔ It is correlated with the Tanimoto coefficient but not strictly monotonic with it.

➔ Cosine similarity or Carbo **index for continuous variabes**,

Range: -1 to +1

➔ **For binary variables**, the coefficient is,

$$S_{AB} = \frac{c}{\sqrt{ab}}$$

Range: 0 to 1

4. **Euclidean distance:**

➔ Euclidean is a distance coefficient measure.

➔ It is used for calculating distance between descriptors.

➔ Euclidean distance, **for continuous variables**,

Range: 0 to ∞

➔ **For binary variables**,

$$D_{AB} = \sqrt{a+b-2c}$$

Range: 0 to N

5. **Hamming distance:**

➔ Hamming distance is also known as Manhattan/City-block distance.

➔ Hamming distance aka Manhattan or City-block, **for continuous variables**,

Range: 0 to ∞

➔ **For binary variables**, the coefficient is

$$D_{AB} = a + b - 2c$$

Range: 0 to N

6. Seorgel index:

➔ Seorgel index is equivalent to (1-Tc) for binary fingerprints.

➔ Seorgel distance **for continuous variables**,

Range: 0 to 1

➔ **For binary variables**, the coefficient is,

$$D_{AB} = \frac{a + b - 2c}{a + b - c}$$

Range: 0 to 1

❖ **Characteristics of the similarity coefficients:**

1. Tanimoto, Dice and Carbo indices measure similarity directly.
2. Hamming, Euclidean and Seorgel formulae provide the distance or dissimilarity between pairs of molecules.
3. Some of the coefficients have values between 0 to 1, while others, this range can be obtained through normalization.
4. Relationship between similarity and distance coefficients is D=1-S
5. The Seorgel distance is the complement of the Tanimoto coefficient ($1-S_{Tan}$) for binary variables.

Metric name	Formula for binary variables	Minimum	Maximum
Tanimoto (Jaccard) coefficient	$S_{AB} = \frac{C}{A+B-C}$	0	1
Dice coefficient (Hodgkin index)	$S_{AB} = \frac{2C}{A+B}$	0	1
Cosine coefficient (Carbo index)	$S_{AB} = \frac{C}{\sqrt{ab}}$	0	1
Soergel distance	$D_{AB} = \frac{a+b-2c}{a+b-c}$	0	1
Euclidean distance	$D_{AB} = \sqrt{a+b-2c}$	0	N[a]
Hamming (Manhattan or city-block) distance	$D_{AB} = a+b-2c$	0	N[a]

❖ **Properties of similarity and distance coefficients:**

1. The distance values must be zero or positive, and the distance from an object to itself must be zero.

$$D_{AB} > 0;\ D_{AA} = D_{BB} = 0$$

2. The distance must be symmetric i.e., $D_{AB} = D_{BA}$
3. The distance values must obey the triangle inequality i.e., $D_{AC} \leq D_{AB} + D_{BC}$
4. The distance between non-identical objects must be greater than zero.

➔ The Hamming, Euclidean and Soergel distance coefficients obey all four properties cited above.

➔ The complements of the Tanimoto, Dice and Cosine coefficients do not obey the triangle inequality.

➔ Coefficients are monotonic when they produce the same similarity rankings.

➔ Hamming and Euclidean distances are monotonic like the Tanimoto and Dice coefficients.

➔ When small molecules are being compared by the Tanimoto coefficient, the coefficient tends to have a small value Smaller molecules also appear closer together (smaller values) when using the Hamming distance

❖ **Effect of size on similarity coefficients:**

❖ **Maximum common sub-graph similarity:**

1. Some similarity measures do not identify local regions of equivalence between two molecules.
2. One way of identifying this is by maximum common sub-graph (MCS).
3. The approach of maximum common sub-graph is to generate alignment between the molecules.
4. MCS is the largest set of atoms and bonds in common between two structures

5. The no. of atoms and bonds in MCS could be used to measure a Tanimoto like coefficient, and quantify the degree of similarity.
6. Several algorithms have been developed to identify the MCS in two molecules which must be represented first as molecular graphs.

❖ **Why consider 3D similarity:**

1. Similarity methods based on 2D descriptors will identify molecules with common substructures.
2. Molecular recognition depends on 3D structure and properties associated with this 3D shape, therefore there is much interest in similarity measures based on 3D properties.
3. The aim is to identify structurally different molecules.
4. To consider 3D properties the conformation of the molecule must be considered, as a consequence such methods are computationally intensive.
5. The complexity increases if the molecule is conformationally flexible.

❖ **Classes of 3D similarity:**

3D similarity methods can be divided into those that are

- Independent of the relative orientation of the molecule.
- Requires the molecules to be aligned in 3D space.

Classes of 3D similarity are as follows:

1. Alignment Independent methods:

➔ Prior to calculating 3D similarity, these methods require consideration of

- ➢ Conformational flexibility.
- ➢ Relative orientation.

➔ If we consider two molecules in a pre-defined conformation, then all that needs to be done is to move one molecule relative to the other, until the similarity measures reaches a maximum (optimal alignment between the two molecules).

➔ Conformational flexibility can be addressed by generating a number of conformations of each molecule and finding the optimal alignment for each of these conformations.

➔ Consideration of conformational flexibility increases greatly the compute time.

➔ This method gives relatively fewer pharmacophoric fingerprints than 2D fingerprints.

➔ It is possible to vary both the conformation and the relative orientation simultaneously to find the best overlap between two molecules.

2. Field-base Alignment methods:

➔ It is now more common to use fields based on Molecular Electrostatic potential (MEP) or various shape derived properties.

- ➔ Field-base Alignment methods considers the electron density of the molecules.
- ➔ These properties are usually mapped to the vertices of 3D grids surrounding the molecules (as in CoMFA).
- ➔ This simplifies the integration over all space.
- ➔ Besides MEP, steric and hydrophobic fields can also be computed and used for comparison.
- ➔ Grid based similarity calculations can be very time consuming, since to maximize the similarity, would involve rotating and translating one grid relative to the other in every possible way and calculating the similarity coefficient.

3. Genomic Projection methods:

- ➔ In this method, the molecule is positioned at the center of a sphere and its properties projected onto the surface of the sphere.
- ➔ The similarities between two molecules are then determined by comparing the spheres.
- ➔ In practice, the sphere is approximated by a tesselated icosahedron or dodecahedron.
- ➔ Each triangular face is divided into a series of smaller triangles.

(UNIT 4)

Introduction to Combinatorial Chemistry and Library Design

❖ **Combinatorial chemistry:**

1) Combinatorial chemistry is a collection of techniques which allow for the synthesis of multiple compounds at the same time.
2) It may be defined as the systematic and repetitive, covalent connection of a set of different "building blocks" of varying structures to each other to yield a large array of diverse molecular entities.
3) It aims to mimic the natural sources to produce pool of chemicals out of which one of them may be proved as lead compound.

Principle:

1) The basic principle of combinatorial chemistry is, to prepare a large number of different compounds at the same time instead of synthesizing compounds in a conventional one at a time manner and then to identify the most promising compound for further development by high throughput screening.
2) The characteristics of combinatorial synthesis is that, different compounds are generated simultaneously under identical reaction conditions (i.e. using the same reaction conditions and the same reaction vessels.) in a systematic manner, so that ideally the products of all possible combinations of a given set of starting materials (termed building blocks) will be obtained at once.
3) The collection of these finally synthesized compounds is referred to as a combinatorial library.
4) The library is then screened for useful properties and the active compounds are identified.

The combinatorial chemistry approach has two phases:

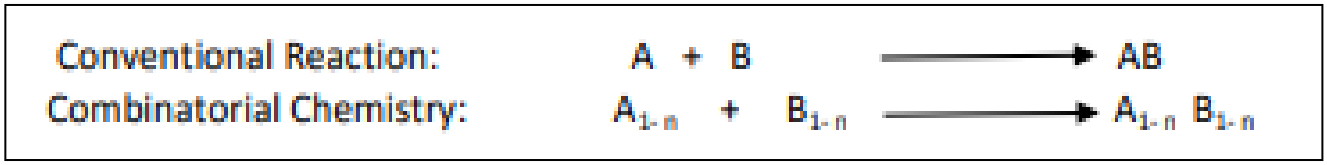
i. Making a library.
ii. Finding the active compound. Screening mixtures for biological activity has been compared to finding a needle in a haystack

1) The combinatorial libraries can be structurally related by a central core structure, termed scaffold (i.e. all compounds of library have a common core structure), or by a common backbone. In both the cases, the accessible dissimilarities of compounds within the library depend on the building blocks which are used for the construction.
2) Combinatorial chemistry is one of the important new methodologies developed by researchers in the pharmaceutical industry to reduce the time and costs associated with producing effective and competitive new drugs.

3) By accelerating the process of chemical synthesis, this method is having a profound effect on all branches of chemistry, but especially on drug discovery.

Example:

1) In a conventional synthesis, one starting material A reacts with one reagent B resulting in one product AB.
2) In a combinatorial synthesis, building blocks of type A (A_1-A_n) are treated simultaneously with different building blocks of type B (B_1-B_n) according to combinatorial principles, each starting material A reacts separately with all reagents B resulting in a combinatorial library A_{1-n}-$B1_{-n}$.
3) Therefore, in combinatorial approach one can cover many combinations An x Bn in one reaction Instead of doing multiple A x B type reactions.



❖ **Need for Combinatorial Chemistry:**

- Earlier, there were problems with Traditional/Conventional Synthesis:

1) The chemist would make only one molecule at a time.
2) Each synthesis was very time consuming.
3) Multistep synthesis has loss at each step.
4) Purification of products very time-consuming between steps.
5) Yields can be low
6) This lead to the production of very few molecules at a time for testing which results in slower lead generation.
7) Hundreds of molecules are generated in a month.
8) There is a high risk of failure.
9) Therefore, to reduce the time and cost, combinatorial chemistry approach was developed by the researchers in the pharmaceutical industry to aid in producing effective and competitive new drugs.
10) One chemist would make multiple molecules at a time.
11) The time and cost associated with the generation and analysis of each individual molecule is significantly less when compared to the time and cost of an individual synthesis.
12) Yields can be high and produces many molecules at a time for testing.
13) Thus leads to faster lead generation and thousands of molecules are generated in a month.
14) Also, there was a low risk of failure.

❖ **Advantages:**

1) The creation of large libraries of molecules in a short time is the main advantage of combinatorial chemistry over traditional.
2) Compounds that cannot be synthesized using traditional methods of medicinal chemistry can be synthesized using combinatorial techniques.
3) The cost of combinatorial chemistry library generation and analysis is very high, but when considered on a per compound basis, the price is significantly lower when compared to the cost of individual synthesis.
4) Yields can be high and produces many molecules at a time for testing. This leads to faster lead generation.
5) Thousands of molecules in a month are generated.
6) There is a low risk of failure.
7) Multiple molecules synthesized at a time.
8) Combinatorial chemistry speeds up the drug discovery process.

❖ **Disadvantages:**

1) Though combinatorial chemistry would solve all the problems associated with drug discovery, one still needs to synthesize the right compound.
2) While a large number of compounds are created, the libraries created are often not focused enough to generate a sufficient number of hits during an assay for biological activity.

❖ **History of Combinatorial chemistry:**

1) The origins of combinatorial chemistry can be traced back at least as far as 1963, when biochemistry professor R. Bruce Merrifield of Rockefeller University, New York City, developed a way to make peptides by solid-phase synthesis.
2) For his work on solid-phase synthesis, Bruce Merrifield won the Nobel Prize in chemistry in 1984 for his work on solid-phase synthesis.
3) During this time, automated peptide synthesizer technology was in its infancy, and the preparation of individual peptides was a challenge.
4) The field in its modern dimensions only began to take shape in the 1980s, when in 1984 research scientist H. Mario Geysen, now at Glaxo Wellcome, Research Triangle Park, N.C., developed a technique to synthesize arrays of peptides on pin-shaped solid supports and in 1985, Richard Houghten developed a technique for creating peptide libraries in tiny mesh "tea bags" by solid-phase parallel synthesis.

5) Another early pioneer was Dr. Árpád Furka who introduced the commonly used split-and-pool method in 1988, which is used to prepare millions of new peptides in only a couple of days and also for synthesizing organic libraries.
6) Through the 80's and into the early 1990's, combinatorial chemistry was focused on peptide synthesis and later oligonucleotide synthesis.
7) In the 1990s, the focus of the field changed predominantly to the synthesis of small, drug like organic compounds and many pharmaceutical companies and biotechnology firms now use it in their drug discovery efforts.

❖ **Pros and Cons of Combinatorial Chemistry:**

Creation of large libraries of molecules in a short time. Compounds that cannot be synthesized using traditional methods of medicinal chemistry done by Combi Chemistry. Cost of combinatorial chemistry library generation and analysis of said library is very high, but when considered on a per compound basis the price is significantly lower when compared to the cost of individual synthesis. More opportunities to generate lead compounds. Combinatorial chemistry speeds up drug discovery	Needs to synthesize the right compound. There is a limit to the chemistry you can do when using solid phase synthesis. The resin you use is often affected by the reaction types available and care must be taken so that the attachment of the reagent to the substrate and bead are unaffected. Each reaction step has to be carefully planned, and often a reaction isn't available because the chemistry affects the resin. There is a great deal of diversity created, but not often a central synthetic idea in the libraries.

❖ **Types of Combinatorial chemistry:**

1) The range of combinatorial techniques is highly diverse, and these products could be made individually in parallel or in mixtures, using either solution or solid phase techniques.
2) Combinatorial chemistry is of two types:
   i. Solid phase combinatorial chemistry (The compound library has been synthesized on solid phase such as resin bead).
   ii. Solution phase combinatorial chemistry (The compound library has been synthesized in solvent in the reaction flask).

1) SOLID PHASE COMBINATORIAL CHEMISTRY:

(Compound library synthesized on solid phase such as resin bead)

Steps:

1) In solid phase combinatorial chemistry, the starting compound is attached to an inert solid/resin bead.
2) Reagents are added to the solution in excess.
3) Separation of products (attached to resin beads) by simple filtration.
4) Cleavage and isolated of products from the beads.

Requirements:

1) A cross-linked insoluble polymeric support which is inert to the synthetic conditions (e.g. a resin bead);
2) An anchor or linker covalently linked to the resin—the anchor has a reactive functional group that can be used to attach a substrate;
3) A bond linking the substrate to the linker, which will be stable to the reaction conditions used in the synthesis;
4) A means of cleaving the product or the intermediates from the linker;
5) Protecting groups for functional groups not involved in the synthetic route.

Example of Solid supports:

1) Partially cross-linked polystyrene beads: Polystyrene is cross linked with divinyl benzene, hydrophobic in nature, causes problems in peptide synthesis due to peptide folding.
2) Sheppard's polyamide resin – more polar
3) Tentagel resin- similar environment to ether
4) Beads, pins and functionalized glass surfaces.

Characteristics of Solid supports:

1) Beads must be able to swell in the solvent used, and remain stable.
2) Most reactions occur in the bead interior.

Advantages:

1) Since, the reaction is carried out on a solid support such as resin beads, a range of different starting materials are available that can be bound to separate resin beads, which are mixed together, such that all the starting material can be treated with another reagent in a single experiment. Therefore, it is possible to do multi-step synthesis and mix-and split synthesis (a technique used to make large number of libraries).

2) Since, the products are bound to solid support, excess reagents or by-products can be easily removed by washing with appropriate solvent. Hence, large excesses of reagents can be used to drive reactions to completion.
3) Intermediates in a reaction sequences are bound to the bead and need not be purified.
4) Individual beads can be separated at the end of reaction to get individual products.
5) Polymeric support can be regenerated and reutilized if appropriate cleavage conditions and suitable anchor/linker group are chosen.
6) Automation is possible.

Disadvantages:

- Not all synthesis can be done on solid phase.

1) Some molecules don't attach well to beads.
2) Some chemistry just doesn't work in this fashion.
3) Removal of product from bead, cab be damaging to product if not careful.

- Typically, kinetics is not the same.

1) Reaction rates can be slower.

- It is difficult to monitor the progress of reaction when the substrate and product are attached to the solid phase.
- Assessment of the purity of the resin attached intermediates is also difficult.
- Purifying the final product after cleavage from the resin also proves to be a challenge.

2) SOLUTION PHASE COMBINATORIAL CHEMISTRY:

(Compound library synthesized in solvent in the reaction flask)

1) All chemical reactions are conducted simultaneously, preferably in well-ordered sets (arrays) of reaction vessels in solution.
2) Soluble polymer are used as support for the product.
3) Most ordinary synthetic chemistry takes place in solution phase.
4) The use of solution phase techniques has been explored as an alternative to solid-phase chemistry approaches for the preparation of arrays of compounds in the drug discovery process.

Advantages:

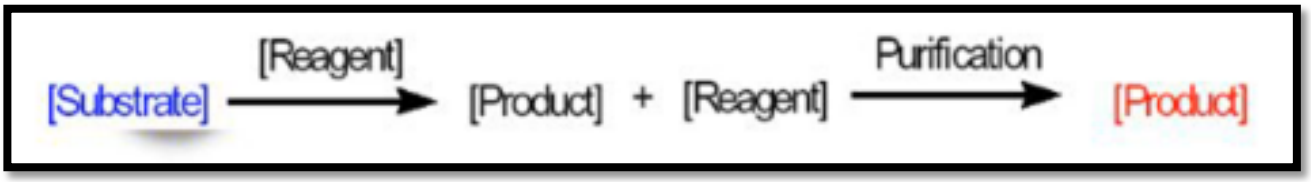
1) Handling of material is easy and can be automated.

Disadvantages:

1) Solution phase work is free from some of the constraints of solid-phase approaches but has disadvantages with respect to purification.
2) In solution phase synthesis we use soluble polymer as support for the product.
3) PEG is a common vehicle which is used in solution phase synthesis it can be liquid or solid at room temperature and show varying degrees of solubility in aqueous and organic solvent.
4) By converting one OH group of PEG to methyl ether (MeO-PEG-OH) it is possible to attached a carboxylic acid to the free OH and use in solution phase combinatorial synthesis.
5) Another common support which is used in solution phase synthesis is liquid Teflon consisting mainly of long chain of (-CF2 -) groups attached to a silicon atom. When these phases are used as a soluble support for synthesis the resulting product can be easily separated from any organic solvent.
6) The main disadvantage of this method is when number of reagents are taken together in a solution, it can result in several side reactions and may lead to polymerization giving a tarry mass.

Limitations:

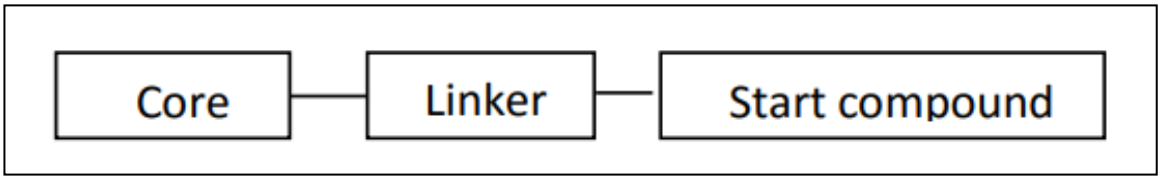
1) When numbers of reagents are taken together in a solution.
    - It can result in several side reactions
    - It leads to polymerization giving a tarry mass.



❖ **Resin (Solid support) used in Solid phase synthesis:**

1) Most solid state combinatorial chemistry is conducted by using polymer beads ranging from 10 to 750 µm in diameter.
2) The solid support must have the following characteristics for an efficient solid phase synthesis:
    a. Physical stability and of the right dimensions to allow for liquid handling and filtration;
    b. Chemical inertness to all reagents involved in the synthesis;
    c. An ability to swell under reaction conditions to allow permeation of solvents and reagents to the reactive sites within the resin;
    d. Derivatization with functional groups to allow for the covalent attachment of an appropriate linker or first monomeric unit.

3) The solid supports are usually composed of two parts: the core and the linker.
4) The starting Compound of the synthesis is attached to the support via the linker.



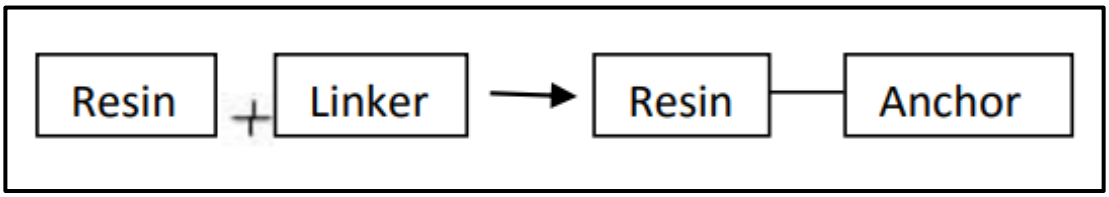
5) The compounds to be synthesized are not attached directly to the polymer molecules but attached by using a **linker moiety** that enables attachment in a way that can be easily reversed without destroying the molecule that is being synthesized and allow some room for rotational freedom of the molecules attach to the polymer.
6) The core ensures the insolubility of the support and determines the swelling properties, while the linker provides the functional group for attachment of the start compound and determines the reaction conditions for the cleavage of the product.
7) The linker itself and the covalent bond formed with the start compound must be stable under the reaction conditions of the synthesis.
8) The bead should be capable of swelling in solvent, yet remain stable.
9) Swelling is important because most of the reactions involved in Solid Phase Synthesis takes place in the interior of the bead rather on the surface.
10) Although beads are the common shape for the solid support, a range of other shapes such as pins have been designed to maximize the surface area available for reaction and hence maximize the amount of compound linked to the solid support.
11) Functionalized glass surfaces have also been used and are suitable for oligonucleotide synthesis

❖ **Types of Solid supports used:**

1) **Polystyrene resins:** Polystyrene is cross linked with divinyl benzene (about 1% cross linking). Polystyrene resin is suitable for non-polar solvents.
2) **Tenta Gel resins:** Polystyrene in which some of the phenyl groups have polyethylene glycol (PEG) groups attached in the para position. The free OH containing resins are suitable for use in polar solvents.**Poly acrylamide resins:** This resin swell better in polar solvent, since they contain amide bonds that more closely resemble biological materials.
3) **Glass and ceramic beads:** This type of solid supports is used when high temperature and high pressure reaction are carried out.

❖ **Linkers / anchors used in solid phase synthesis:**

1) The initial building block of the compound to be prepared by solid phase synthesis is covalently attached to the solid support via the linker.
2) A molecular moiety which is covalently attached to the solid support and which contains a reactive functional group.
3) Allows attachment of the first reactant.
4) The bond formed between the linker and substrate must be stable to the reaction conditions used throughout the synthesis and it should be easily cleaved to release the final compound after the synthesis is completed.
5) It is a bi-functional molecule, one functional group for irreversible attachment to the core resin and a second functional group for forming a reversible covalent bond with the initial building block of the product and the linker remains after cleavage at the resin.
6) Different linkers are available depending on the functional group to be attached and the desired functional group on the product.



7) **A series of selected examples are found below. Resins are named to define the linker e.g.,**

➔ **Merrifield resin:** The Merrifield resin can be used to attach carboxylic acids to the resin. The product can be cleaved from the resin in carboxylic acid form using HF.

➔ **Wang resin:** The resin is used to bind carboxylic acids. The ester linkage formed has a good stability during the solid phase reactions but its cleavage conditions are milder than that of the Merrifield resin. Usually 95% TFA is applied. It is frequently used in peptide synthesis.

➔ **Rink resin** The Rink resin is designed to bind carboxylic acids and cleave the product in carboxamide form under mild conditions. The amino group in the resin is usually present in protected form.

➔ **Hydroxymethyl resin:** The resin can be applied for attachment of activated carboxylic acids and the cleavage conditions resemble that of the Merrifield resin.

➔ **Photolabile anchors:** Photolabile anchors have been developed that allow cleavage of the product from the support by irradiation without using any chemical reagents. Such anchors, like the 2-nitrobenzhydrylamine resin below, usually contain nitro group that absorbs UV light.

➔ **Traceless anchors:** The initial building block of a multi-step solid phase synthesis needs to have one functional group (in addition to others) for its attachment to the solid support. It may happen that in the end product, this group is unnecessary and needs to be removed. For this reason anchors have been developed that can be cleaved without leaving any functionality in the end product at the cleavage site. These traceless anchors usually contain silicon based linkers.

❖ **Protecting groups used in Solid phase synthesis:**

1) Primary function of protecting group is to protect the portion of the molecule that is not covalently bound to the resin and must be protected to avoid subsequent polymerization of excess monomers in solution.
2) A protecting group is reversibly attached to the functional group to convert it to a less reactive form.
3) When the protection is no longer needed, the protecting group is cleaved and the original functionality is restored.
4) A large number of protecting groups were developed for use in peptide synthesis since the amino acids are multifunctional compounds.
5) It is an important requirement for a protecting group to be stable under the expected reaction conditions of each coupling.
6) After coupling is performed, the protecting group is removed toexpose a new reactive site and synthesis continues in a repetitive fashion.
7) Cleavage conditions are dictated by the linker used.
8) Two protecting groups are said to be orthogonal if either of them can be removed without affecting the stability of the other one.
9) Some of the protecting groups most widely used in peptide synthesis are described below. Protection of amino groups:
    a) Benzyl carbonyl (Z) group.
    b) t-butoxy carbonyl (Boc) group.
    c) 9-fluorenyl methoxy carbonyl (9-Fmoc) group.

❖ **Characteristics of solid phase and solution phase combinatorial chemistry:**

SOLID PHASE	SOLUTION PHASE
Make a mixture of products	Makes only one product
Small amounts of products formed	Large amounts of products formed
Simple isolation of product by filtration	Work-up and purification more difficult
Requires two extra reaction steps: linkage & cleavage	No extra steps for attachment & cleavage needed
Limits to chemistry which can be performed	Wide range of reactions can be utilized
Automation possible	Automation difficult
Large excesses of reagent can be used to drive the reaction to completion	Large excesses of reagent cannot be used as it causes subsequent separation problem.
Longer reaction time than in solution phase	Less reaction time
Monitoring of reaction very difficult	Monitoring of reaction easy
Split and mix technique as well as parallel synthesis can be applied	Split and mix strategy not possible Parallel synthesis can be applied

❖ **DIFFERENCE BETWEEN SOLID PHASE AND SOLUTION PHASE SYNTHESIS:**

On a solid support	In solution
Reagents can be used in excess in order to drive the reaction to completion	Reagents cannot be used in excess, unless addition purification is carried out
Purification is easy: simply wash the support	Purification can be difficult
Automation is easy	Automation is difficult
Fewer suitable reactions	In theory any organic reaction can be used
Scale-up relatively expensive	Scale-up is easy and relatively inexpensive
Not well documented and time will be required to find a suitable support and linker for a specific synthesis	Only requires time for the development of the chemistry

❖ **TYPES OF COMBINATORIAL LIBRARIES:**

1) Scaffold-based Libraries:

➔ Core-structure, which is common to all compounds of the library.

➔ Several single building blocks can consist of Scaffold.

➔ Example- Amino acid and Amino Benzophenone.

2) Backbone-based Libraries

➔ Example- Nucleic acid and Carbohydrate.

- **2 approaches to generate libraries are as follows:**

1) Random/diverse libraries:

➔ Synthesis of diverse compounds – large number of molecules – more hits (biological assay).

➔ Little is known about the target – more diverse library (primary screening library).

2) Focused libraries:

➔ Synthesis of focused compounds – small number of molecules.

➔ Incorporate as much information about the therapeutic target as possible.

❖ **TECHNIQUES FOR LIBRARY PREPARATION:**

1) There are two methods, which used for synthesis of compounds in combinatorial chemistry.
2) They include:
    a) Split and mix synthesis or Split and pool synthesis or Portioning – Mixing (PM) synthesis (one bead-one compound library).
    b) Parallel synthesis (one vessel-one compound library)

A. <u>SPLIT AND MIX SYNTHESIS 'OR' PORTIONING– MIXING SYNTHESIS</u>: (One bead-one compound library)

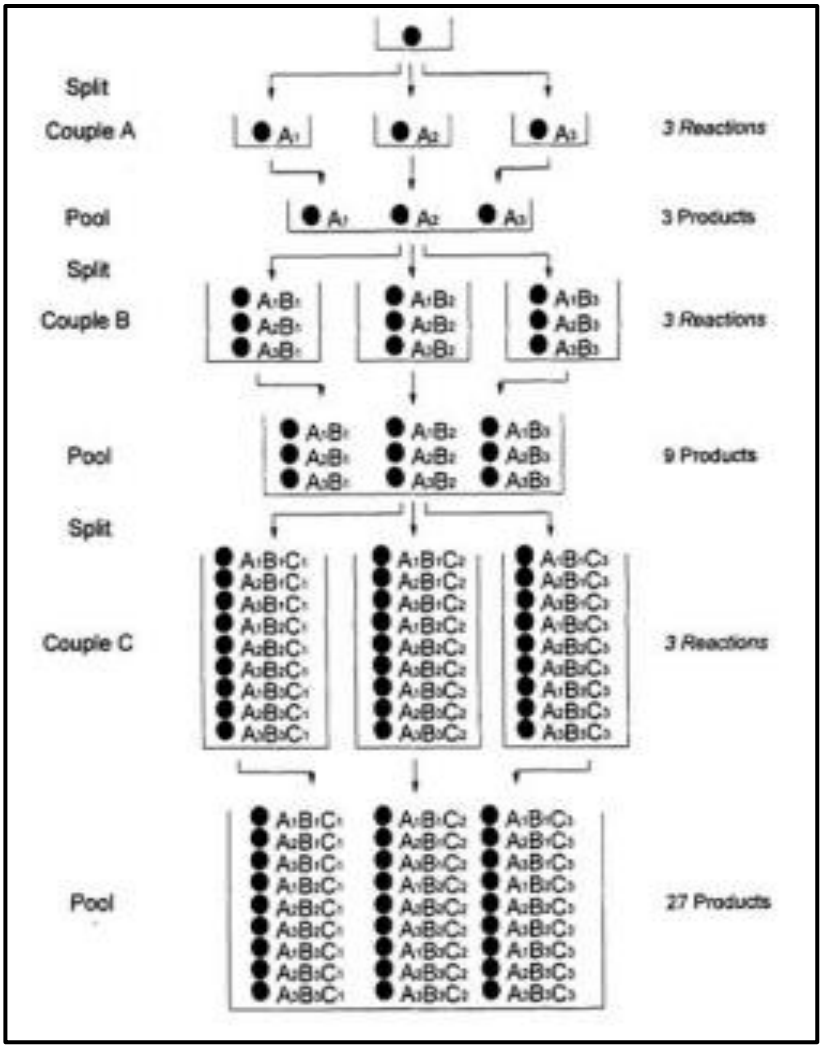
1) This technique was pioneered by Dr.Árpád Furka and co-workers in 1988 for the synthesis of large peptide libraries.

2) This approach was termed divide couple and recombines synthesis by other workers.

Steps:

1) In this method, ingredients are assembled on the surface of the beads or micro particles.
2) In each step, beads from last steps are partitioned into new building block and several groups are added.
3) This leads to the formation of new groups, the different groups of beads are recombined and separated once again.
4) Process is continuous with next building block is added until the desired library has been assembled.
5) After a Split-Pool synthesis, just one single compound is bound to each resin bead.
6) Split-Pool procedure requires a solid support.
7) Therefore, this method is particularly employed for solid phase synthesis.

Example: In following figure spheres represents resin beads, A, B & C represent the sets of building block and borders represents the reaction vessels. In the case, when three building blocks are used, in each coupling step after three stages (ie. divide, couple & recombine), a total number of 27 different compounds, one on each resin bead, are formed using 9 individual reactions (ignoring deprotection). On the resulting products from split and pool synthesis, bioassays is performed and active mixture is discovered. Once an active mixture has been discovered, the next task is discovering which individual compound(s) in that mixture are active. The process of determining these active compounds is known as deconvolution.



Advantages:

a) Only few reaction vessels required.

b) Large libraries can be quickly generated (up to 10^5 compounds).

Disadvantages:

a) Threefold amount of resin beads necessary.

b) The amount of synthesized product is very small.

c) Complex mixtures are formed.

B. <u>PARALLEL SYNTHESIS</u>:

(One vessel-one compound library)

1) It involves multiple reactions, at once instead of in series, each in a separate vessel.

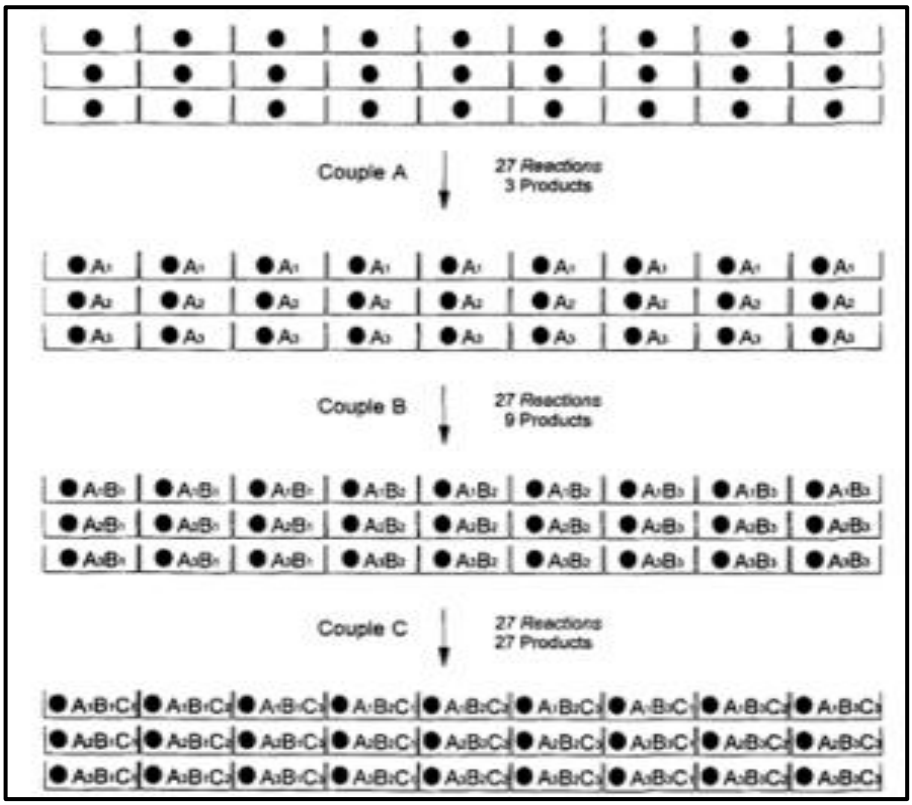
2) A single product is obtained in each different reaction vessel.

Steps:

1) Each compound is synthesized in specific reaction vessel.

2) Each starting material is reacted with each building block separately.
3) Then, product is spilt into portions, reacted with different building block separately again.
3) Methods of parallel synthesis include Houghton's tea bag procedure and Automated parallel synthesis.

Example: In following figure spheres represents resin beads, A, B & C represent the sets of building block and borders represents the reaction vessels. In the case, when three building blocks are used, in each coupling step after three stages, a total number of 27 different compounds, one on each resin bead, are formed using 9 individual reactions (ignoring deprotection).



Advantages:

a) It creates the compounds individually and in their own vessel. Thus the identity of the product is already known.
b) No deconvolution is required.
c) Each compound is substantially pure in its location.
d) Defined location provides the structure of a certain compound.
e) Biological evaluation is easy.

Disadvantage:

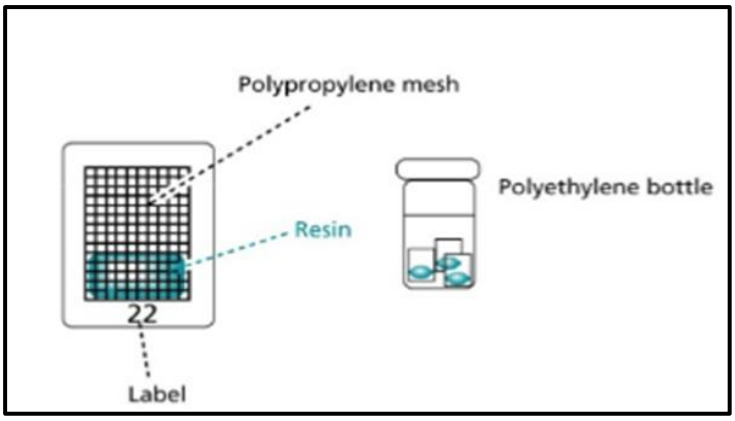
a) Applicable only for medium libraries (several thousand compounds).

b) Large amount of vessels are required.

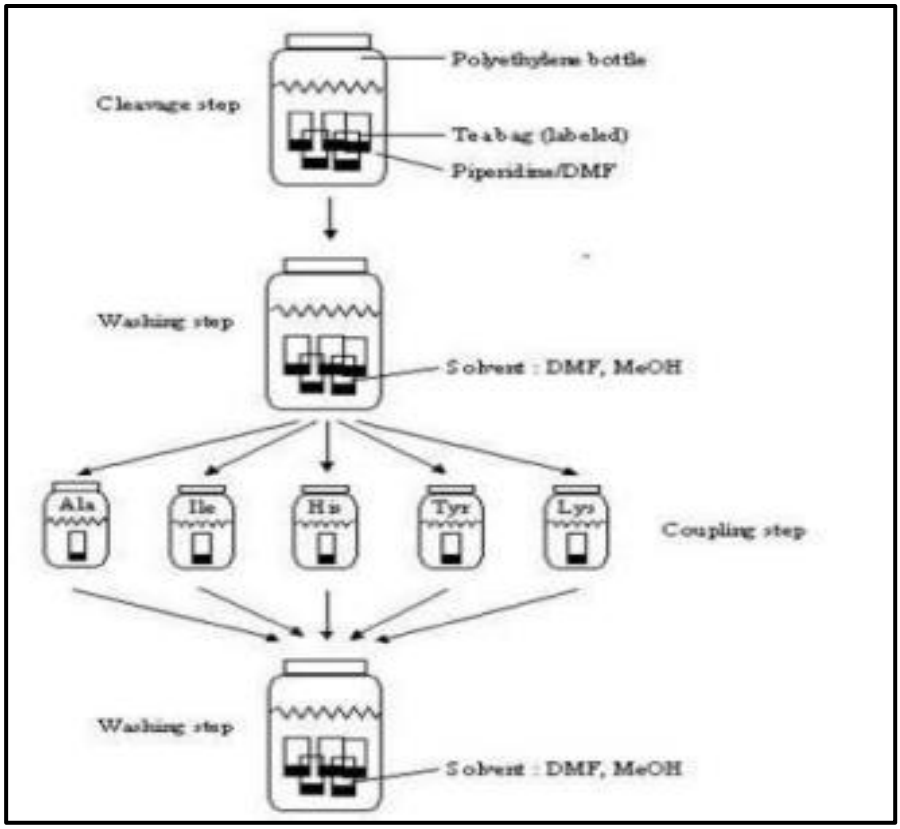
c) Large number of reactions is to be performed.

❖ **METHOD FOR PARALLEL SYNTHESIS:**

- **Houghton Teabag method:**



1) **Definition:** A polypropylene mesh bag, with dimensions of approximately 15 x 20 mm, filled with resin beads, sealed and labeled for a later identification, is known as a tea-bag, designed by Houghten in 1985.
2) The "tea-bag" mesh size is too small to allow resin beads to escape, but solvents and soluble reagents could readily enter.
3) The **principles** of its use are to make multimilligram (up to 500 µmoles) quantities of a single peptide sequence in each packet, which is sufficient for full characterization and screening.
4) It is a manual approach to parallel synthesis.
5) To save time and work while making many peptides simultaneously, bags could be combined into the same reactors for common chemical steps.
6) **Example:**
7) In the synthesis of 40 different peptides, all the bags are initially charged with resin beads bearing a Boc-protected amino acid, and the packets are combined for resin deprotection, washing, and neutralization steps.
8) Then the bags are sorted into groups for the addition of the next amino acid.
9) Then the bags could be combined again for deprotection, washing, and neutralization.
10) After an appropriate number coupling steps, all the bags can then be treated with HF/anisole to cleave the peptides from the beads.
11) As the first intention was to speed up peptide synthesis, nowadays the tea-bag method is a classic example for combinatorial synthesis, its speed, and effectiveness.
12) **Schematic overview of a typical group of steps carried out using the tea-bag procedure:**



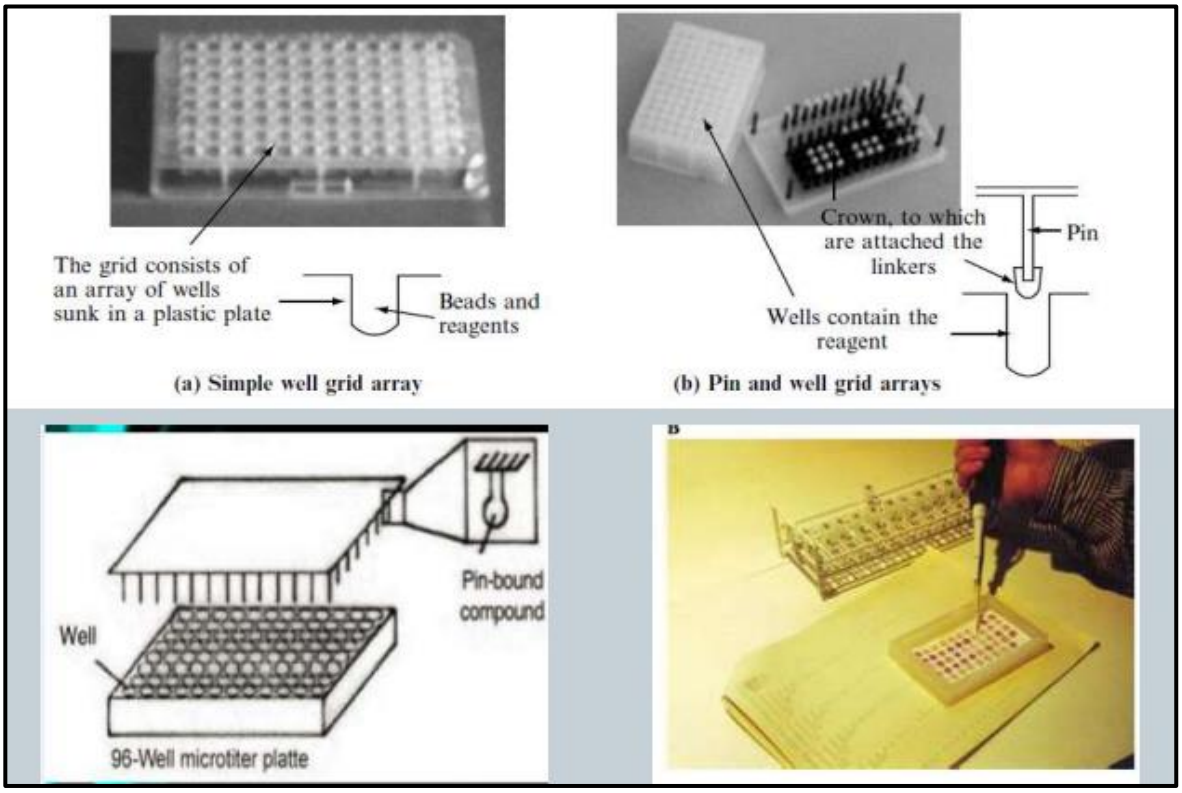
➔ **Advantages** –

- Easy to identify active hit as its position (X, Y coordinate) in the array encodes the reagents and thus structure of the product.
- New equipment, such as 'personal synthesizers' and 'multi vial apparatus,' allows parallel synthesis of many compounds simply and quickly by one chemist.
- Robotized technology.
- Greater quantity of each compound is available at once (structural characterization).
- Labeling of the tea bags leads to easier identification of each compound.

➔ **Disadvantages –**

- Maximum impurities can occur unless the reactions are very clean.
- Most useful for one to three step reactions only.
- Can only be used for making smaller (more focused) libraries.

- **Automated Parallel synthesis:**



1) Parallel synthesis is possible when it advances in automation.
2) Automated synthesizers are available with 42, 96, 144 reaction vessels or wells.
3) Beads or pins are used for solid support.
4) Reactions and work ups are carried out automatically.
5) Same synthetic route for each vessel, but different reagents.
6) Different product obtained per vessel.

Steps:

1) In this method, each starting material is reacted with each building block separately (i.e. in separate vessel), without remixing.
2) This is not like a split synthesis because it requires a solid support.
3) It can be done without solid support or in a solution.
4) A 96 well micro titer plate is commonly used format for parallel synthesis.
5) After each reaction step the product is split into 'n' portions before it is reacted with n new building blocks.
6) In following figure spheres represents resin beads, A, B & C represent the sets of building block and borders represents the reaction vessels. In the case, when three building blocks are used, in

each coupling step after three stages, a total number of 27 different compounds, one on each resin bead, are formed using 9 individual reactions (ignoring deprotection).

7) Like the split and pool method, it results in the production of multiple compounds at the same time.
8) However, unlike split and pool, parallel synthesis gives individual compounds, not a mixture. Thus deconvolution is not an issue in this method.

C. MIXED COMBINATORIAL SYNTHESIS:

1) The aim is to use a standard synthetic route to produce a large variety of different analogues where each reaction vessel or tube contains a mixture of products.
2) The identities of the structures in each vessel are not known with certainty.
3) It is useful for finding a lead compound.
4) It is capable of synthesizing large numbers of compounds quickly.
5) Each mixture is tested for activity as the mixture.
6) Inactive mixtures are stored in combinatorial libraries.
7) Inactive mixtures are studied further to identify active component.

❖ **Screening of Combinatorial Library:**

1) Can be done in 2 ways: Virtual screening and Experimental real screening.

Virtual screening:

2) Virtual screening uses computational methods to predict or simulate how a particular compound interacts with a given target protein.
3) The 3 virtual screening methods used in modern drug discovery include
   - Molecular Docking,
   - Pharmacophore Mapping
   - QSAR/QSPR
4) Disadvantages of virtual screening:
   - Cannot replace real screening
   - Generated hits may be very difficult to chemically synthesize

Experimental real screening:

5) Real screening approaches, such as high-throughput screening (HTS), can test the activity of hundreds of thousands of compounds experimentally, providing real results;
6) Disadvantage:
   - These methods are far more expensive and
   - Slower than virtual screening methods.

7) Most common assay to screen a combinatorial library is to determine the binding of the library compounds to the target protein.
8) Other common assays are functional assays such as biochemical and enzymatic assays, or cell-based assays.
9) Cell-based assays can be direct cytotoxic assays, receptor-binding assays, or cell-signaling assays using cell lines with specific genetic reporter systems.
10) Selection of screening methods greatly depends on:
    - The nature of the combinatorial libraries to be screened.
    - Position-addressable soluble libraries prepared from parallel synthesis can be screened with automated HTS methods in 96-, 384-, and 1536-well plates.
    - Libraries on solid supports (e.g. OBOC library) can be easily screened against a variety of biological targets (proteins, cells, viruses, etc.) for binding or functional activities or released in situ for solution phase functional assays.
11) Phage-display peptide libraries can be screened with bio-panning or limited cell-based functional assays, such as cell-binding and cellular uptake assays.
12) Structure-based virtual libraries are screened in silico.

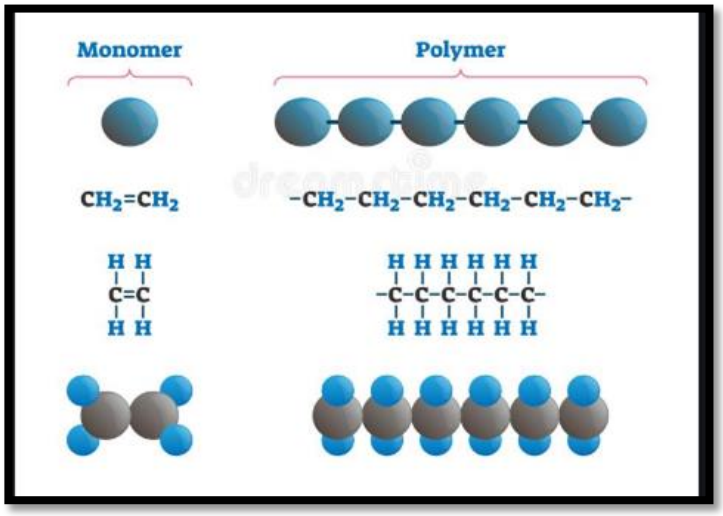
❖ **Applications of Combinatorial Chemistry:**

1) Application of combinatorial library methods in cancer research and drug discovery
2) Building synthetic gene circuits from combinatorial libraries: screening and selection strategies
3) Combinatorial library approaches for improving soluble protein expression in *Escherichia coli*
4) Combinatorial library-based strategies to optimize proteins
5) A Combinatorial Library Strategy for the Rapid Humanization of Anticarcinoma BR96 Fab
6) Generation and use of synthetic peptide combinatorial libraries for basic research and drug discovery.
7) Used in anti-viral research.

❖ **Strategies for Library Designing:**

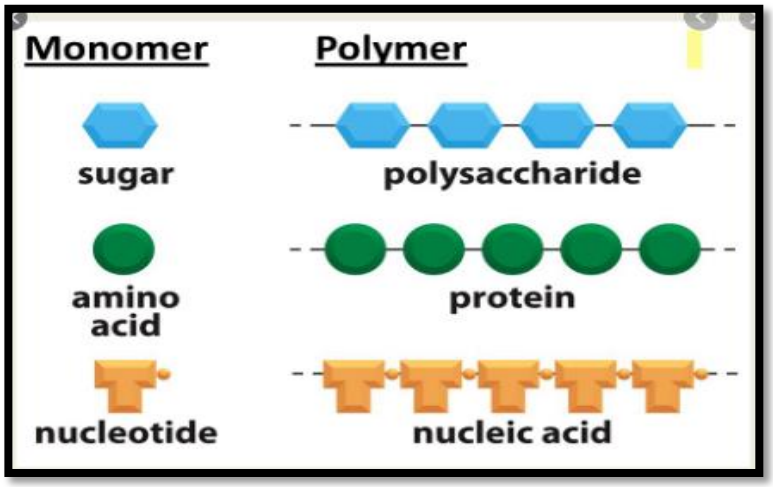
Two main strategies:

1. **Monomer-based selection:**



1) The small individual repeating units/molecules are known as monomers (means single part).
2) Imagine that a monomer can be represented by the letter A. Then a polymer made of that monomer would have the structure.

-A-A-A-A-A-A-A-A-A-A-A-A-A-A-A-A-A-A-A-A-A-A-A-A-A-A-A

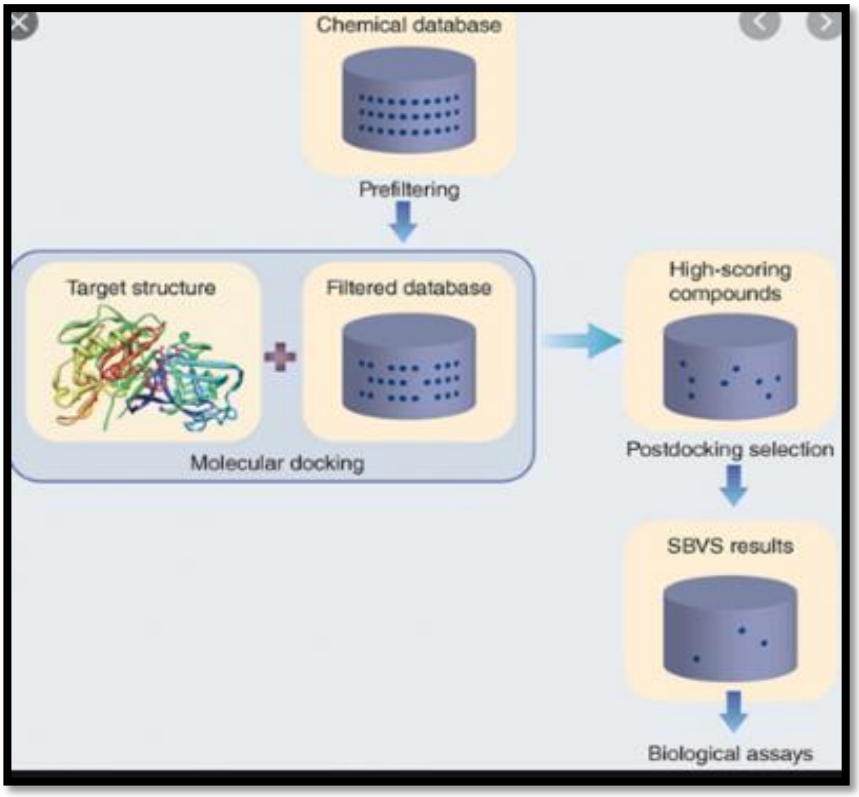


3) In monomer-based selection optimized subsets of monomers are selected without consideration of the products that will result.
4) Consider a hypothetical three-component library with 100 monomers available at each position of variability, where the aim is to synthesize a diverse 10 × 10 × 10 combinatorial library.

5) In monomer-based selection this would involve selecting the 10 most diverse monomers from each set of monomers i.e. there are subsets of size n contained within a larger set of N compounds.
6) Eg: more than 1013 different subsets of size 10 from a pool of 100 monomers.
7) It is not possible to examine all of these.
8) The subset selection problem can be solved in the context of selecting compounds for screening where the techniques of dissimilarity-based compound selection, clustering and partitioning were introduced, together with related optimization methods.

$$\frac{N!}{n!(N-n)!}$$

2. Product-based selection:



1) Product-based library design involves a more complex optimization procedure that we term 'combinatorial optimization' where the reagent selection is optimized against the properties of the corresponding products.
2) In product-based selection, the properties of the resulting product molecules are taken into account when selecting the monomers.

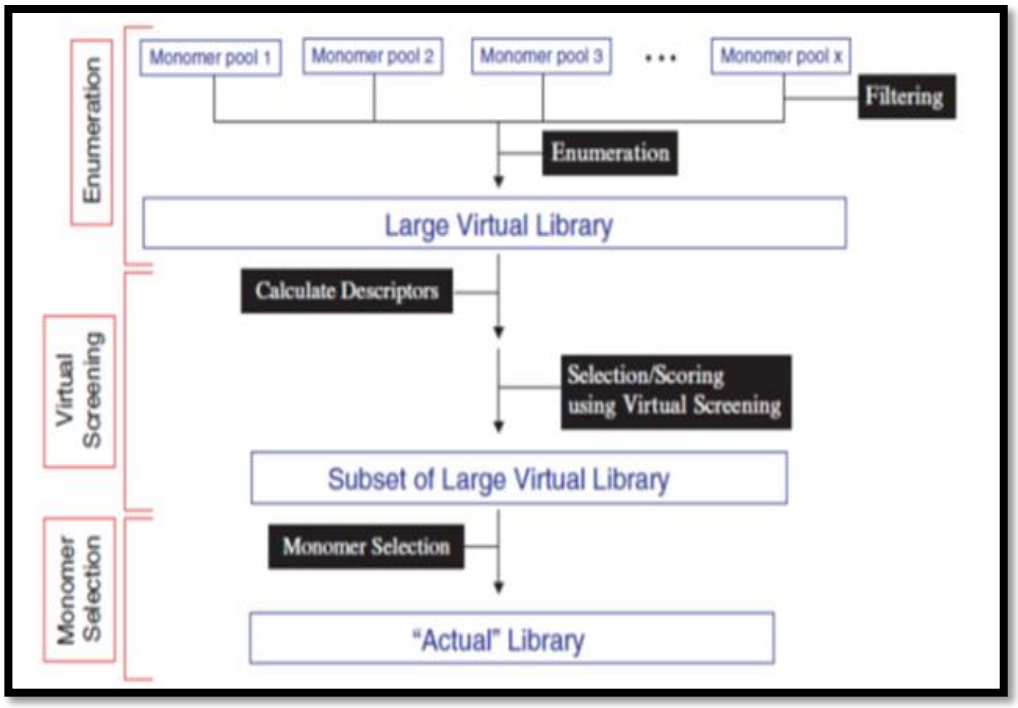
3) Having enumerated the virtual library any of the subset selection methods could then be applied.
4) This process is generally referred to as cherry-picking but it is synthetically inefficient in so far as combinatorial synthesis is concerned.
5) Synthetic efficiency is maximized by taking the combinatorial constraint into account and selecting a combinatorial subset such that every reagent selected at each point of variation reacts with every other reagent selected at the other positions.
6) It is much more computationally demanding than monomer-based selection, but can be more effective while optimizing the properties of a library as a whole.
7) The number of combinatorial subsets in this case is given by the following equation: where R is the number of positions of variability and there are ni monomers to be selected from a possible Ni at each substitution position.
8) Thus, there are almost 1040 different 10×10×10 libraries that could be synthesized from a 100×100×100 virtual library.
9) The selection of combinatorial subsets has been tackled using optimization techniques such as simulated annealing and genetic algorithms.
10) Despite the greater computational complexity of performing product-based selection compared to monomer-based selection it can be a more effective method when the aim is to optimize the properties of a library as a whole, such as diversity or the distribution of physicochemical properties.

$$\prod_{i=1}^{R} \frac{N_i!}{n_i!(N_i - n_i)!}$$

Approaches to Product-based Library Design:

- A general strategy for product-based library design involves the following 3 steps. They are as follows:

1) Lists of potential reagents are identified (e.g., by searching relevant databases), filtered them as needed, and the virtual library is enumerated.
2) The virtual library is subjected to virtual screening to evaluate and score each of the structures.
3) The reagents to be used in the actual library for synthesis are selected using the results from the virtual screening together with any additional criteria (such as the degree of structural diversity required, degree of similarity or dissimilarity to existing collections).



4) It is important to note that it may be possible to reduce significantly the size of the virtual library by eliminating from consideration monomers that can be unambiguously identified as being inappropriate.
5) The final, monomer selection stage is typically implemented using optimization techniques such as GAs or simulated annealing.
6) Assume a two component combinatorial synthesis in which nA of a possible NA first monomers are to be reacted with nB of a possible NB second monomers.
7) The chromosome of the GA thus contains *nA + nB* elements, each position specifying one possible monomer.
8) Then, the fitness function quantifies the "*goodness*" of the combinatorial subset encoded in the chromosome and the GA evolves new potential subsets in an attempt to maximize this quantity.
9) In some cases the virtual library is too large to allow full enumeration and descriptor calculation, making product-based combinatorial subset selection unfeasible.
10) A number of methods have been proposed to try to overcome this problem.
11) Alternative approaches to product-based library design have been developed that do not require enumeration of the entire virtual library.
12) These methods have been termed molecule-based methods to distinguish them from library based methods and they are appropriate for the design of targeted or focused libraries.

13) The molecule-based method is a relatively fast procedure, especially when optimization is based on 2D properties, since the fitness function involves a pairwise molecular comparison rather than the analysis of an entire library, as is the case in library-based methods.

14) In these approaches, however, there is no guarantee that building libraries from frequently occurring monomers will result in optimized libraries, nor is it possible to optimize properties of the library as a whole.

❖ **Encoding Combinatorial Libraries:**

1) By the process of screening the number of libraries that has "desirable properties" are sorted out.
2) It is now very important to learn the identity of "winning" library member.
3) The process of identification of active compound in a mixture of compounds is known as Encoding.
4) Chemical structure of individual compounds in conventional addressable combinatorial libraries or planar microarray libraries are known, there is no need to encode and decode the chemical hits.
5) For mixture libraries in solution, such as positional-scanning libraries, purification is needed to determine the identity of the hits.
6) Biological-displayed peptide libraries (e.g., phage, yeast or mRNA-display) are genetically encoded and can be decoded with PCR, DNA barcoding, DNA sequencing, Edman microsequencing , NGS, mass spectroscopy of released coding tags, fluorescence-based encoding method ,etc..
7) More than one million codes can be generated by using combinations of different methods, which are highly stable and reliable under bioassay conditions.
8) For identification of active compound following types of encoding methods are used:
   a. **Positional encoding or deconvolution (iterative resynthesis and rescreening).**
   b. **Chemical encoding (Tagging)**
   c. **Electronic encoding**

a. Positional encoding or deconvolution (iterative resynthesis and rescreening):

➔ In this method, the resynthesis and rescreening is carried out to know the identity of the active compound. In other terms, it is a process of optimizing an activity of interest by fractionating (normally by resynthesis, or by elaborating a partial library) a pool with some level of the desired activity to give a set of smaller pools.

➔ Repeating this strategy leads to single members with (ideally) a high level of activity and is termed iterative deconvolution.

b. Chemical encoding (Tagging):

➔ The most common approach to encoding solid phase libraries is to attach a chemical tag to the resin beads as the target molecule gets synthesized.

➔ Typically, at each step in the reaction, a tag is attached that is unique for the given step.

➔ For example, if we are creating a tripeptide and we have 10 possible amino acids at each position, we need to attach either a single tag that says "the tripeptide on this bead has amino acid Ala atposition 1, Phe at position 2 and Gly at position 3" or we need to attach three different tags, one for each position.

c. Electronic encoding:

➔ This technique uses a micro electronic device called a radio frequency (RF) memory tag.

➔ The tag measuring 13×3 mm is encased in heavy walled glass and contains the following:

- A silicon chip ,onto which laser etched a binary code,
- A rectifying circuit with which absorbed RF energy is converted to D.C. electrical energy,
- A transmitter/receiver circuit,
- An antenna, through which energy is received and RF signals are both received and sent.

6) Library Enumeration:

➔ Process by which the molecular graphs of the product molecules are generated automatically from lists of reagents (using connection tables or SMILES strings**).**

1) Fragment marking

- Central core template and one or more R groups.

2) Reaction transform approach

- Transform is a computer-readable representation of the reaction mechanism: atom mapping.

Advantages / Disadvantages:

- Fragment marking generally a very fast enumeration once core template and R group fragments are defined.
- May be difficult to generate the core and to generate fragments automatically.

3) Markush-based approaches to enumeration:

- Ideally suited when a common core can be identified.
- Certain subsets of the product structures may have features in common.

7) Identification of Active Ingredient Major Challenge in developing library of compounds:

➔ Major challenge in developing library of compounds is screening the library for the activity of the chemical species responsible.

➔ The goal of producing molecular libraries is to discover compounds that have some desired properties to serve as a drug.

1. **Analytical techniques:**

The resin bead mix and split method can be used to generate hundreds, thousands or even millions of different products.

As an example, a four step synthesis employing 10 building blocks at each step would afford 10 000 different compounds in only 10*4 chemical steps.

Although synthesis is rapid, the power of combinatorial libraries is only evident if structural information on active components may be easily obtained.

The iterative re-synthesis and rescreening offers a solution, but as it can be slow and requires a further dedication of synthetic and screening resource, there have been a number of new methods devised where information concerning the active compound may be carried on the bead in the form of a "tag".

The synthetic efficiency of the split synthesis technique can be contrasted with the technical difficulties encountered when analyzing the resulting libraries. For example, the simple split synthesis scenario outline above results in a library consisting of 10 pools of 1 000 compounds each. These compounds can be cleaved into solution and screened as soluble pools, or the ligands can remain attached to the beads and screened in immobilized form. Neither scenario is ideal for several reasons. Because of limitations on solubility, the concentration of the individual compounds present in soluble pools must be correspondingly diminished as the pool size increase – perhaps below a desirable threshold for screening. Biological screens performed on such large mixtures of soluble compounds can be ambiguous since the observed activity could be due to a single compound or due to a collection of compounds acting either collectively or synergistically. The subsequent identification of specific biologically active members is challenging, since the number of compounds present in the pools and their often-limited concentration deter their isolation and erase. Because of this, biologically active pools are often iteratively re-synthesized and re-assayed as increasingly smaller subsets until activity data are obtained on homogenous compounds.

In some instances, bead-based split synthesis libraries can be successfully assayed with the ligands still immobilized to the beads.

In this process, a reporter system is employed in the biological assay such that beads displaying active ligands can be physically distinguished from those displaying inactive compounds.

Suitable reporter system includes the use of fluorescently labeled receptors, or anti-receptors antibodies similarly labeled with a reporter molecule, that can be employed to "label" active beads.

Beads thus marked are physically removed and analyzed to identify the attached ligand.

This technique is limited by the capacity of the biological screen to detect immobilized ligands, as well as the sensitivity of the analytical methods employed to unambiguously identify the attached compounds.

2. **DNA based encoding**:

   One of the first reported successful ligand encoding strategies exploited oligo-deoxyribonucleic acid (DNA) as the surrogate analyte. This DNA encoding concept had in fact been demonstrated in some of the first combinatorial library preparation methods ever reported – those utilizing filamentous phage particles. In this approach, libraries of peptides are prepared biochemically from the cloning and expression of random sequence oligonucleotides. Pools of oligonucleotides encoding the peptides of interest are interested into an appropriate expression system, where upon translation the resulting peptides are synthesized as fusion proteins. One of the common expression systems fuses these sequences to the gene III or the gene VIII coat protein of filamentous phage particles. Each viral particle contains a unique DNA sequence that encodes only a simple peptide. After screening a library in a given biological system, any viral particles displaying active peptides are isolated and the structure of the active peptides is elucidated by sequencing their encoding DNAs. A distinct disadvantage with this approach is that the molecular diversity of such systems is limited to peptides, and amino acids that compose these peptides are restricted to the 20 encoded by genes.

   DNA encoded peptide prepared in a 1:1 correspondence on a linker capable of anchoring the synthesis of both oligomers. The structure of the peptides is determined by sequencing their accompanying unique DNA sequence.

3. **Mass encoding:**

   The entire reported single bead encoding schemes require the co-synthesis of a suitable tagging moiety to record the synthetic history of each compound prepared in the library. This is inherently inefficient, since each unique compound could encode for itself if appropriate analytical techniques such as 1H, 13C NMR could be used to assign structures to ligands present in the amounts provided by single beads.

   It can be seen that in each of these cases above, the use of a tagging group allows the synthesis of any type of compound within the library. The tagging molecules can encode for any building block and any synthetic transformation. Furthermore, given the uncertainties of much synthetic chemistry, the tag may be looked upon as not so much encoding a specific compound structure, but encoding instead a synthetic procedure. Thus, even if the intended compound was not made but biological activity was detected, the tagging system facilitates a replication of the synthetic steps employed in producing the active compound, and thus aids structure determination.

4. **Peptide tag:**

It has been recognised that peptides could be employed as tag since their information content could be extracted with high sensitivity via Edman degradation and sequencing. Since the Edman degradation requires a free N-terminus, this peptide as code strategy could also be used to encode other peptide by acylating the N-terminus of the binding peptide strand, and leaving a free amine at the coding peptide terminus. To accommodate the parallel synthesis of both binding and coding peptides, an orthogonally protected bifunctional linker was employed that contained both acid and base sensitive protecting groups. This bifunctional linker resided on the cleavable Rink amide linker, such that peptide-encoded peptide conjugates would be released into solution upon treatment of the Rink linker with 95% TFA.

The ligand and its associated tag are synthesized on a 1:1 correspondence on a cleavable linker and realized into free solution. Affinity selection techniques are employed to isolate conjugates that bind to the receptor, enzyme, or antibody target of interest.

The above peptide and DNA encoding techniques are not ideal because of the chemical liability of these oligomers. This places a severe restriction on the scope of the synthetic techniques that may be applied during library synthesis, and restricts the synthesis of more pharmaceutically attractive small organic molecules.

5. **Hard tag (Haloaromatic tag):**

the first encoding method utilizing such chemically stable tagging moieties. The tag consisted of haloaromatic reagent linker to a carboxylic acid though an internal photochemically cleavable linker. Amide bond chemistry served to attach the tag to the beads. These haloaromatic reagents acylated the same synthesis sites used for ligand synthesis (Figure), but due to the sensitivity of tag detection this competition could be minimised. Once the haloaromatic analyte was attached to the bead it could be selectively detached into solution upon photolysis with UV light. The liberated tag could then be resolved and detected as subpicomole concentrations using electron capture capillary gas chromatography EC-GC.

Chemically "hard" haloaromatic tag suitable for encoding applications where the beads will be exposed to rigorous synthetic conditions the tag are released photochemically and then detected via EC-GC.

A: haloaromatic tags incorporated via amine bond chemistry at the expense of the ligand synthesis sites.

B: tags incorporated via carbene insertion.

In both A & B tag concentration are minimised to prevent chemically derivation of the encoded ligands or the quenching of their synthesis sites.

While hard tagging strategies have been successfully used to encode a variety of different synthetic chemistries, a common limitation remains – the requirement for parallel synthesis (ligand and encoding tags). Since the robust preparation of a large combinatorial library is frequently a difficult synthesis

challenge, it would be desirable to obviate the need for tag cosynthesis and instead delineate individual compounds by other physical means.

6. **Radio frequency encoding:**

Radio frequency (RF) encoding techniques physically encapsulate an RF encodable microchip with the synthesis resin, such that the RF transponder can be scanned post-synthesis to identify its associated product.

RF encoding successfully avoids the need to cosynthesise surrogate analytes, and also permits the larger scale synthesis of compounds since each microcapsule can hold tens milligrams of synthesis beads.