COMPUTATIONAL CHEMISTRY – UNIT II

Scope of Computational Modelling

Due to the advancement of technologies the speed of research work is supplemented by the usage of various tools and software. In many cases the computational research is done at the initial stage followed by the experimental work. In order to do interdisciplinary research, the expertise from different categories is essential. A background of computing with domain knowledge is always beneficial. The challenge is always how do we incorporate the existing knowledge/skill and build up new knowledge/skills to contribute effectively.

At one time, computational chemistry techniques were used only by experts extremely experienced in using tools that were for the most part difficult to understand and apply. Today, advances in software have produced programs that are easily used by any chemist. Along with new software comes new literature on the subject. There are now books that describe the fundamental principles of computational chemistry at almost any level of detail. A number of books also exist that explain how to apply computational chemistry techniques to simple calculations appropriate for student assignments. There are, in addition, many detailed research papers on advanced topics that are intended to be read only by professional theorists.

The group that has the most difficulty finding appropriate literature are working chemists, not theorists. These are experienced researchers who know chemistry and now have computational tools available. These are people who want to use computational chemistry to address real-world research problems and are bound to run into significant difficulties. This unit is chosen to cover a large number of topics, with an emphasis on when and how to apply computational techniques rather than focusing on theory. It gives a clear description with just the amount of technical depth typically necessary to be able to apply the techniques to computational problems. There are many good books describing the fundamental theory on which computational chemistry is built. The description of that theory as given here is very minimal. We have chosen to include just enough theory to explain the terminology used in computing.

Many computational chemistry techniques are extremely computer-intensive. Depending on the type of calculation desired, it could take anywhere from seconds to weeks to do a single calculation. There are many calculations, such as ab initio analysis of biomolecules, that cannot be done on the largest computers in existence. Likewise, calculations can take very large amounts of computer memory and hard disk space. In order to complete work in a reasonable amount of time, it is necessary to understand what factors contribute to the computer resource requirements. Ideally, the user should be able to predict in advance how much computing power will be needed.

There are often trade-offs between equivalent ways of doing the same calculation. For example, many ab initio programs use hard disk space to store numbers that are computed once and used several times during the course of the calculation. These are the integrals that describe the overlap between various basis functions. Instead of the above method, called conventional integral evaluation, it is possible to use direct integral evaluation in which the numbers are recomputed as needed. Direct integral evaluation algorithms use less disk space at the expense of requiring more CPU time to do the calculation. An in-core algorithm is one that stores all the integrals in RAM memory, thus saving on disk space at the expense of requiring a computer with a very large amount of memory. Many programs use a semidirect

algorithm, which uses some disk space and a bit more CPU time to obtain the optimal balance of both.

Cost and Efficiency

Chemistry's impact on modern society is most readily perceived in the creation of mate-rials, be they foods, textiles, circuit boards, fuels, drugs, packaging, etc. Thus, even the most ardent theoretician would be unlikely to suggest that theory could ever supplant experiment. Rather, most would opine that opportunities exist for combining theory with experiment so as to take advantage of synergies between them.

With that in mind, one can categorize efficient combinations of theory and experiment into three classes. In the first category, theory is applied *post facto* to a situation where some ambiguity exists in the interpretation of existing experimental results. For example, photolysis of a compound in an inert matrix may lead to a single product species analyzed by spectroscopy. However, the identity of this unique product may not be obvious given a number of plausible alternatives. A calculation of the energies and spectra for *all* of the postulated products provides an opportunity for comparison and may prove to be definitive. In the second category, theory may be employed in a simultaneous fashion to optimize the design and progress of an experimental program. Continuing the above analogy, *a priori* calculation of spectra for plausible products may assist in choosing experimental parameters to permit the observation of minor components which might otherwise be missed in a complicated mixture (e.g., theory may allow the experimental instrument to be tuned properly to observe a signal whose location would not otherwise be predictable).

Finally, theory may be used to predict properties which might be especially difficult or dangerous (i.e., costly) to measure experimentally. In the difficult category are such data as rate constants for the reactions of trace, upper-atmospheric constituents that might play an important role in the ozone cycle. For sufficiently small systems, levels of quantum mechanical theory can now be brought to bear that have accuracies comparable to the best modern experimental techniques, and computationally derived rate constants may find use in complex kinetic models until such time as experimental data are available. As for dangerous experiments, theoretical pre-screening of a series of toxic or explosive compounds for desirable (or undesirable) properties may assist in prioritizing the order in which they are prepared, thereby increasing the probability that an acceptable product will be arrived at in a maximally efficient manner.

Molecular Interactions:

Molecular interactions are attractive or repulsive forces *between* molecules and between non-bonded atoms. Molecular interactions are important in all aspects of chemistry, biochemistry and biophysics, including protein folding, drug design, pathogen detection, material science, sensors, gecko feet, nanotechnology, separations, and origins of life. Molecular interactions are also known as noncovalent interactions, intermolecular interactions, non-bonding interactions, noncovalent forces and intermolecular forces. All of five of these phrases mean the same thing.

Non-Bonding Interactions. Molecular Interactions are *between* molecules, or between atoms that are not linked by bonds. Molecular interactions include cohesive (attraction between like), adhesive (attraction between unlike) and repulsive forces between molecules. Molecular interactions change (and bonds remain intact) when (a) ice melts, (b) water boils, (c) carbon dioxide sublimes, (d) proteins unfold, (e) RNA unfolds, (f) DNA strands separate

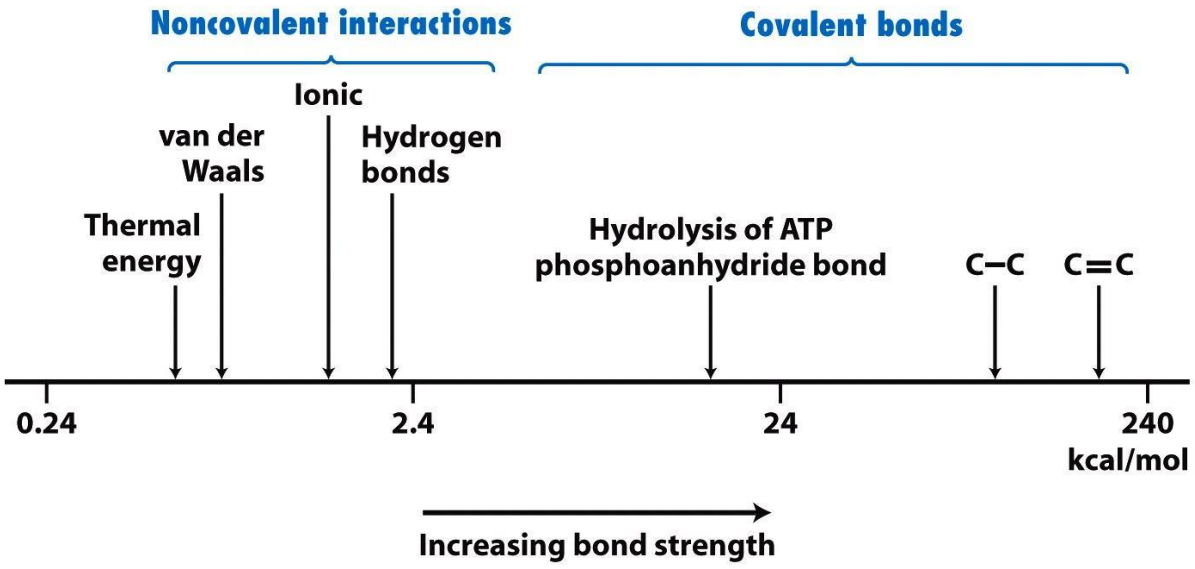
and (g) membranes disassemble. The enthalpy of a given molecular interaction, between two non-bonded atoms, is 1 - 10 kcal/mole (4 - 42 kjoule/mole), which in the lower limit is on the order of RT and in the upper limit is significantly less than a covalent bond.

Bonding Interactions. Bonds hold atoms together *within* molecules. A molecule is a group of atoms that associates strongly enough that it does not dissociate or lose structure when it interacts with its environment. At room temperature two nitrogen atoms can be bonded (N_2). Bonds break and form during chemical reactions. In the chemical reaction called fire, bonds of cellulose break while bonds of carbon dioxide and water form. Bond enthalpies are on the order of 100 kcal/mole (400 kjoule/mole), which is much greater than RT at room temperature; bonds do not break at room temperature.

Boiling Points. When a molecule transitions from the liquid to the gas phase (as during boiling), ideally all molecular interactions are disrupted. Ideal gases are the ONLY systems where there are no molecular interactions. Differences in boiling temperatures give good qualitative indications of strengths of molecular interactions in the liquid phase. High boiling liquids have strong molecular interactions. The boiling point of H_2O is hundreds of degrees greater than the boiling point of N_2 because of stronger molecular interactions in H_2O (liq) than in N_2(liq). The forces between molecules in H_2O (liq) are greater than those in N_2(liq).

Bond Strength

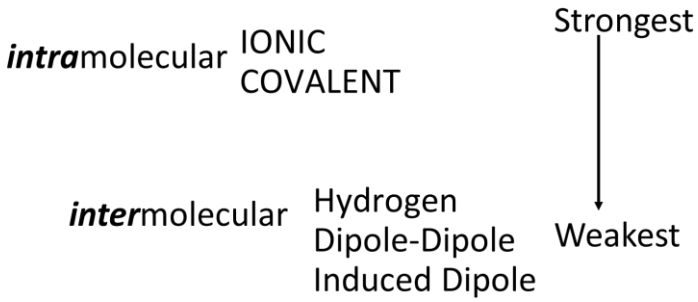
Noncovalent interactions are weak electrical bonds between molecules. Noncovalent interactions (1-5 kcal/mol) are typically ~100-fold weaker than covalent bonds.



Intermolecular Forces

- Intermolecular forces are interactions that exist between molecules. Functional groups determine the type and strength of these interactions.
- There are several types of intermolecular interactions.
- Ionic compounds contain oppositely charged particles held together by extremely strong electrostatic inter-actions. These ionic inter-actions are much stronger than the intermolecular forces present between covalent molecules.

- Covalent compounds are composed of discrete molecules.
- The nature of the forces between molecules depends on the functional group present. There are three different types of interactions, shown below in order of increasing strength:
    - van der Waals forces
    - dipole-dipole interactions
    - hydrogen bonding

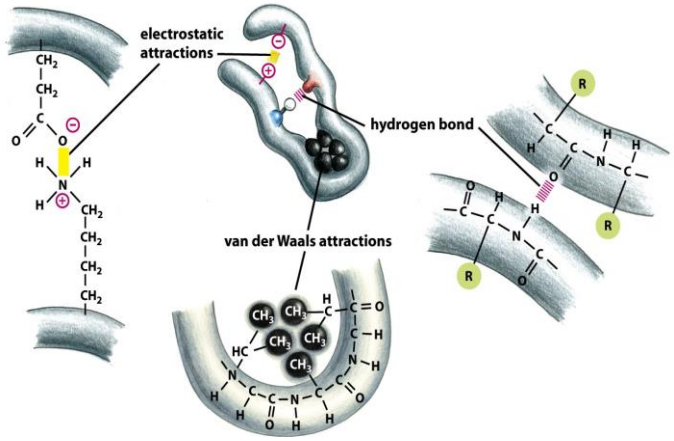


Noncovalent interactions determine protein structure

- Amino acids are connected by *covalent* bonds called peptide bonds.
- Four types of *noncovalent* interactions between amino acids affect protein structure:

van der Waals interactions

- Electrostatic interactions (salt bridges)
- Hydrogen bonds
- Hydrophobic forces



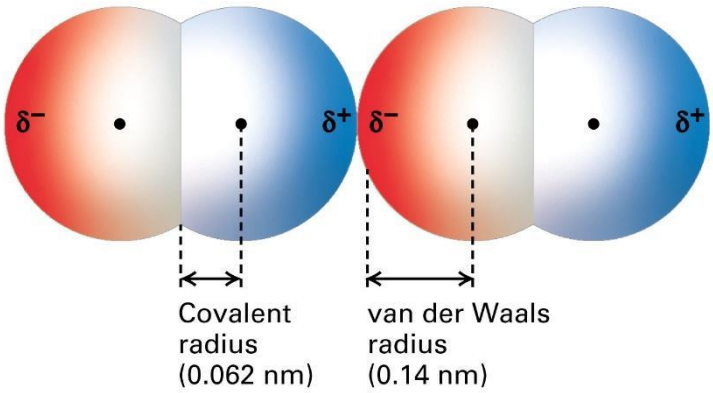
Short range repulsion

Atoms take space. Force two atoms together and they will push back. When two atoms are close together, the occupied orbitals on the atom surfaces overlap, causing electrostatic repulsion between surface electrons. This repulsive force between atoms acts over a very short range, but is very large when distances are short.

The repulsive energy goes up as $(d_i / R)^{12}$, where R is the distance between the atoms and d_i is the distance threshold below which the energy becomes repulsive. d_i depends on the types of atoms. The large exponent means that when $R < d_i$ then small decreases in R cause large increases in repulsion. Short range repulsion only matters when atoms are in very close proximity ($R < d_i$), but at close range it dominates other interactions. Because this repulsion rises so sharply as distance decreases it is often useful to pretend that atoms are hard spheres, like very small pool balls, with hard surfaces (called van der Waals surfaces) and well-defined radii (called van der Waals radii).

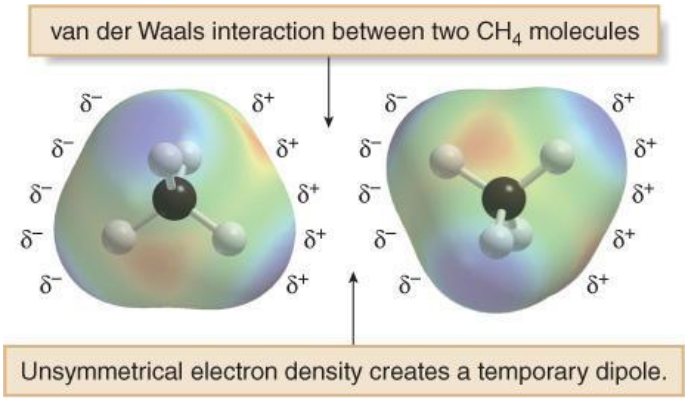
Van der Waals Forces

Van der Waals interactions are bonds between fluctuating, induced dipoles within the electron clouds of interacting molecules. These bonds can occur between nonpolar or polar molecules. van der Waals bonds are extremely dependent on the distance of separation between molecules, and are significant only when the electron clouds of the molecules are just touching. van der Waals interactions are demonstrated for two O_2 molecules and the covalent and van der Waals radii are shown.



- Van der Waals forces are also known as London forces.
- They are weak interactions caused by momentary changes in electron density in a molecule.
- They are the only attractive forces present in nonpolar compounds.
- All compounds exhibit van der Waals forces.

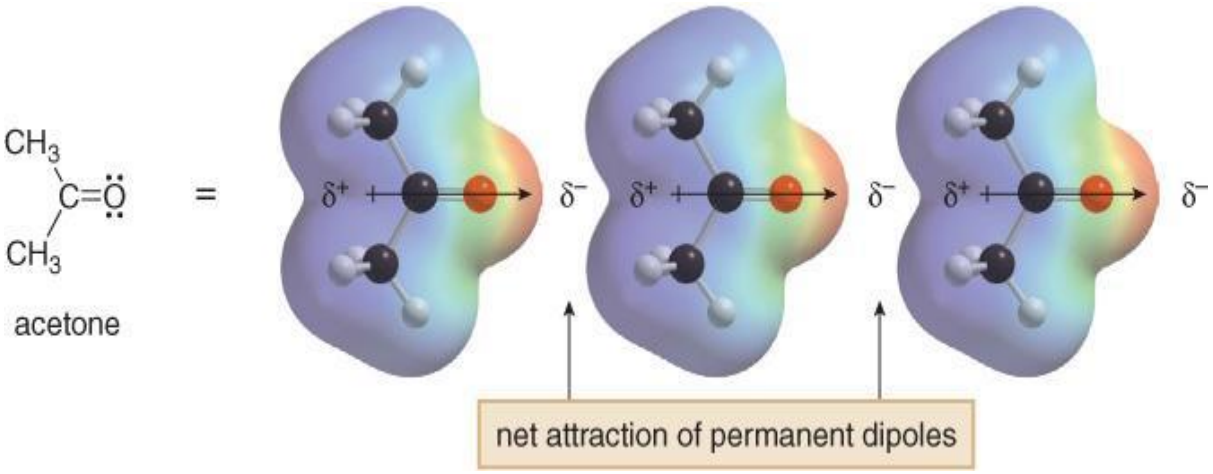
The surface area of a molecule determines the strength of the van der Waals interactions between molecules. The larger the surface area, the larger the attractive force between two molecules, and the stronger the intermolecular forces



Even though CH_4 has no net dipole, at any one instant its electron density may not be completely symmetrical, resulting in a temporary dipole. This can induce a temporary dipole in another molecule. The weak interaction of these temporary dipoles constitutes van der Waals forces.

Dipole-Dipole Interactions

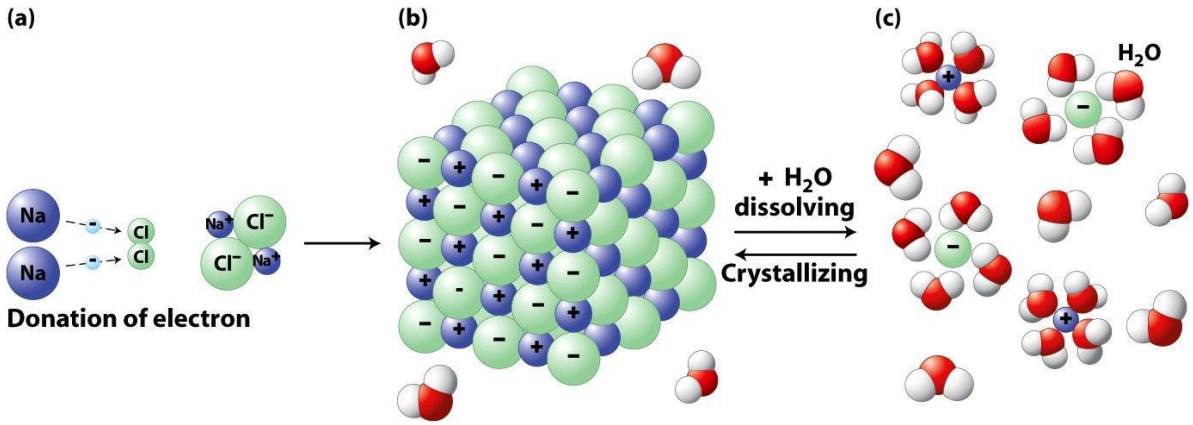
Dipole—dipole interactions are the attractive forces between the permanent dipoles of two polar molecules.



Consider acetone. The dipoles in adjacent molecules align so that the partial positive and partial negative charges are in close proximity. These attractive forces caused by permanent dipoles are much stronger than weak van der Waals forces.

Electrostatic Interactions

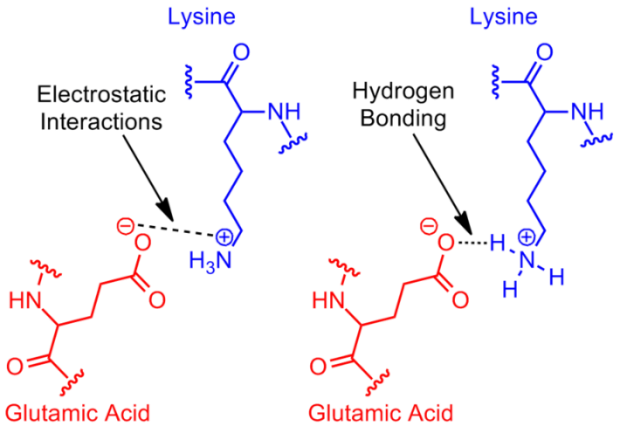
Ionic compounds such as NaCl are readily dissolved n water. Solvation spheres of water molecules surround ions in solutions. Water molecules orient so that the negative ends of their dipoles contact cations and the positive ends contact anions in solution.



Favorable electrostatic interactions cause the vapor pressure of sodium chloride and other salts to be very low. If you leave crystals of table salt (NaCl; Na^+=cation, Cl^-=anion) on a hot pan, how long does it take before they vaporize and sublime away? A very very long time; electrostatic interactions are very very strong. The electrostatic interactions within a sodium chloride crystal are called ionic bonds. But when a single cation and a single anion are close together, within a protein, or within a folded RNA, those interactions are considered to be non-covalent electrostatic interactions.

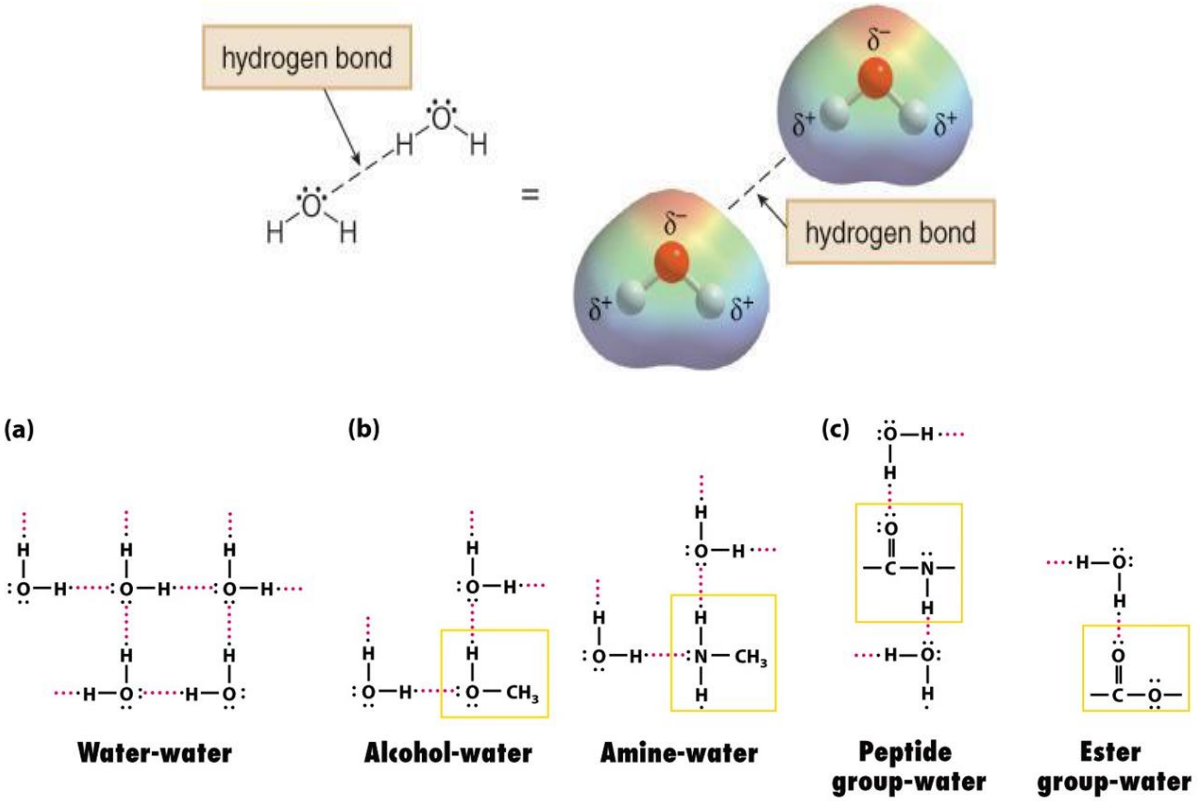
Electrostatic Interactions (Salt Bridge)

A salt bridge is *a non-covalent interaction between two ionized sites*. It has two components: a hydrogen bond and an electrostatic interaction. Salt bridges in proteins are bonds between oppositely charged residues that are sufficiently close to each other to experience electrostatic attraction.



Hydrogen Bonding

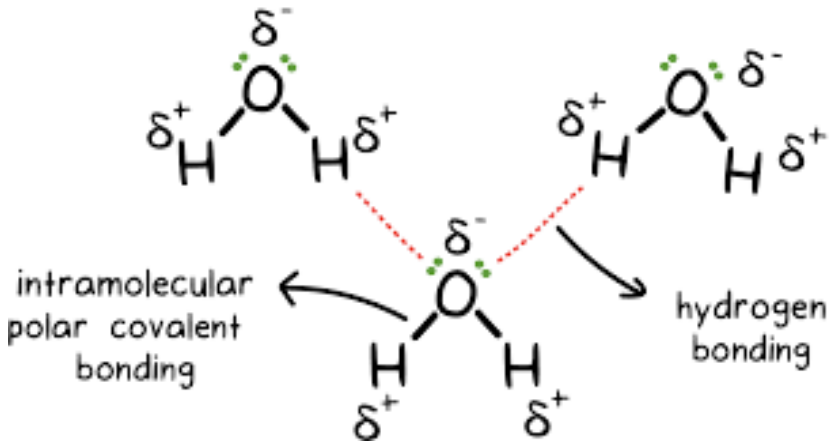
Hydrogen bonding typically occurs when a hydrogen atom bonded to O, N, or F, is electrostatically attracted to a lone pair of electrons on an O, N, or F atom in another molecule.



Why hydrogen? Hydrogen is special because it is the *only* atom that (i) forms covalent sigma bonds with electronegative atoms like N, O and S, *and* (ii) uses the inner shell (1S) electron(s) in that covalent bond. When its electronegative bonding partner pulls the bonding electrons away from hydrogen, the hydrogen nucleus (a proton) is exposed on the back side (distal from the bonding partner). The unshielded face of the proton is exposed, attracting the partial negative charge of an electron lone pair. Hydrogen is the only atom that exposes its nucleus this way. Other atoms have inner shell non-bonding electrons that shield the nucleus.

Water is a **POLAR** molecule, there are unshared pairs of electrons on the central atom.

More on ***inter***molecular forces Hydrogen "Bonding". STRONG ***inter***molecular force Like magnets. Occurs *ONLY* between H of one molecule and N, O, F of another molecule



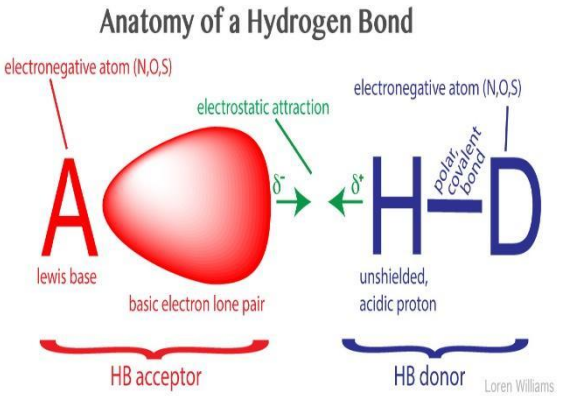


Figure 5 illustrates the elements of a hydrogen bond, including the HB acceptor and HB donor, the lone pair and the exposed proton. N, O, S are the predominant hydrogen bonding atoms (A & D) in biological systems.

A hydrogen bond is *not* an acid-base reaction, where the proton (H^+) is fully transferred from H-D to A to form D^- and HA^+. However, the strength of a hydrogen bond correlates well with the acidity of donor H-D and the basicity of acceptor A. In a hydrogen bond, the H^+ is partially transferred from H-D to A, but H^+ remains covalently attached to D. The H-D bond remains intact.

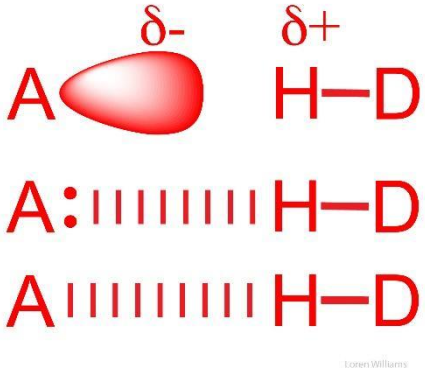


Figure 6 illustrates three different styles for representing a hydrogen bond. Atom A is the Lewis base (for example the N in NH_3 or the O in H_2O) and the atom D is electronegative (for example O, N or S). The conventional nomenclature is confusing: a hydrogen bond is not a covalent bond.

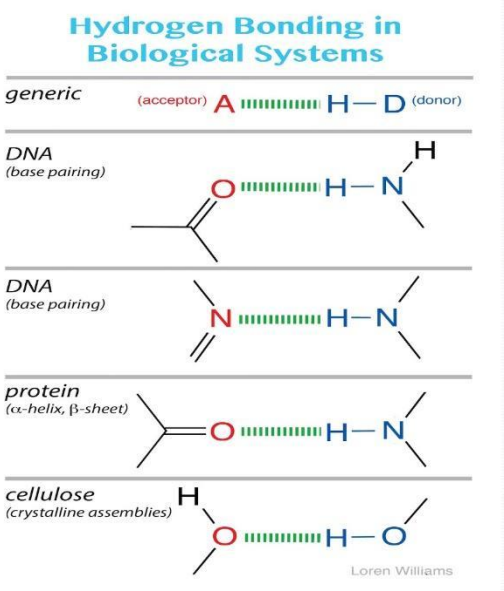


Figure 7 shows the most common hydrogen bond acceptors and donors in biological macromolecules.

he most common hydrogen bonds in biological systems involve oxygen and nitrogen atoms as A and D. Keto groups (=O), amines (R_3N), imines (R=N-R) and hydroxyl groups (-OH) are the most common hydrogen bond acceptors in DNA, RNA, proteins and complex carbohydrates. Hydroxyl groups and amines/imines are the most common hydrogen bond donors. Hydroxyls and amines/imines can both donate and accept hydrogen bonds.

In traversing the Period Table, increasing the electronegativity of atom D strips electron density from the proton (in H-D), increasing its partial positive charge, and increasing the strength of any hydrogen bond. Thiols (-SH) can both donate and accept hydrogen bonds but these are generally weak, because sulfur is not sufficiently electronegative. Hydrogen bonds involving carbon, where H-D equals H-C, are observed, although these are weak and infrequent. C is insufficiently electronegative to form good hydrogen bonds. Hydrogen bonds are essentially electrostatic in nature, although the energy can be decomposed into additional contributions from polarization, exchange repulsion, charge transfer, and mixing.

Hydrogen bond strengths form a continuum. Strong hydrogen bonds of 20-40 kcal/mole (82 to 164 kjoule/mole), generally formed between charged donors and acceptors, are nearly as strong as covalent bonds, Weak hydrogen bonds of 1-5 kcal/mole (4 - 21 kjoule/mole), sometimes formed with carbon as the proton donor, are no stronger than conventional dipole-dipole interactions. Moderate hydrogen bonds, which are the most common, are formed between neutral donors and acceptors are from 3 - 12 kcal/mole (12 - 50 kjoule/mole)).

A hydrogen bond is not a bond. It is a molecular interaction (a non-bonding interaction).

Molecular Topology:

One property of molecules appears to be very close to a binary relation: that is two atoms in a given molecule are either bonded or not bonded. Therefore, molecules can be represented by graphs when the only property considered is the existence or not of a chemical bond. This property is called molecular topology.

- In chemistry, topology provides a way *of describing and predicting the molecular structure within the constraints of* three-dimensional (3-D) space.

These graphs represent different chemical objects: molecules, reactions, crystals, polymers, clusters, etc. The common feature of chemical systems is the presence of *sites* and *connections* between them. Sites may be atoms, electrons, molecules, molecular fragments, groups of atoms, inter- mediates, orbitals, etc. The connections between sites may represent bonds of any kind, bonded and nonbonded interactions, elementary reaction steps, rearrangements, van der Waals forces, etc. Chemical systems may be depicted by *chemical graphs* using a simple conversion rule:

Site ↔ vertex

connection ↔-» edge

A special class of chemical graphs are *molecular graphs.* Molecular graphs are chemical graphs which represent the *constitution* of molecules. They are also called *constitutional graphs.*

- In these graphs vertices correspond to individual atoms and edges to chemical bonds between them.

Molecular graphs are necessarily connected graphs. As examples the molecular graphs corresponding to propane and cyclopropane are shown in Figure 10.

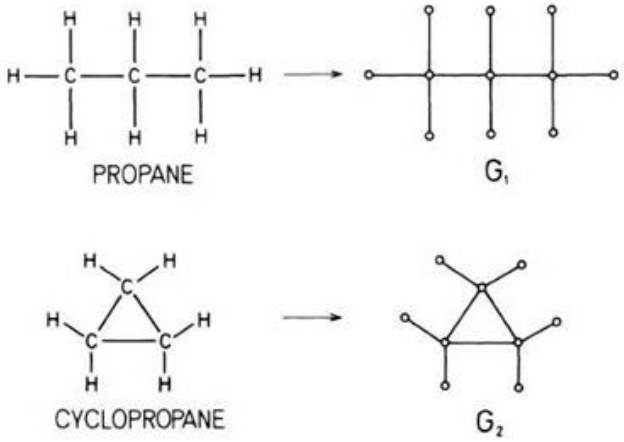


Figure 10 The molecular graphs corresponding to propane and cyclopropane

- In order to simplify the handling of molecular graphs, *hydrogen-suppressed graphs,* i.e., graphs depicting only molecular skeletons without hydrogen atoms and their bonds, are often used. They are also called *skeleton graphs.*

The hydrogen-suppressed graphs are almost universally used in chemical graph theory, because the neglect of the hydrogen atoms and their bonds in most cases cannot be the cause of any ambiguity. The hydrogen- suppressed graphs corresponding to butane and cyclobutane are given in Figure 11.

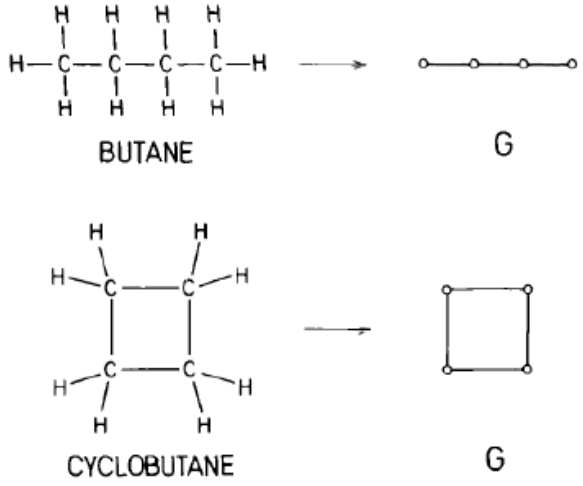


Figure 11 The hydrogen suppressed molecular graphs depicting butane and cyclobutene

The molecular graph grossly simplifies the complex picture of a molecule by depicting only its constitution (i.e., the chemical bonds between the various pairs of atoms in the molecule) and neglecting other structural features (e.g., geometry, stereochemistry, chirality). Even so, a simple picture of a molecule as the molecular graph can enable one to make useful predictions about physical and chemical properties of molecules. Since the predictions of properties and reactivities of molecules are of prime interest to chemists, the development of chemical graph theory is, thus, justified.

Molecular graphs depicting constitutional formulae of molecules represent their *topology.* This is a chemist's view of molecular topology. However, a more precise definition of molecular topology may also be given using the concept of the molecular graph. A *topological space* is formed by a set and the topological structure defined upon the set. A simple connected (molecular) graph can be associated with a topological space if it can be shown that a topological structure is defined upon its vertex-set.

Graph theoretical matrices

The chemical graphs can be handled in many different representations for calculations to draw meaningful information. In mathematics matrix algebra is very easy to do such calculations. One of the ways to represent chemical graphs for computational purpose is graph theoretical matrix. Graphs, adequately labeled, may be associated with several matrices.

- A graph G is *labeled* if a certain numbering of vertices of G is introduced. Here two graph-theoretical matrices, i.e., the adjacency matrix and the distance matrix will be discussed. They are also sometimes referred to as topological matrices. These matrices may be used for identifying certain properties of graphs, which would not otherwise easily emerge.

The Adjacency Matrix

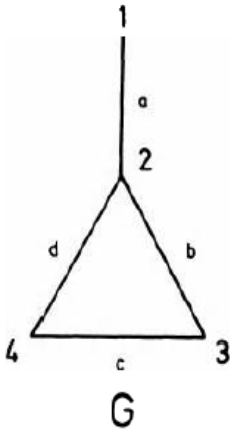
The most important matrix representation of a graph G is the *vertex- adjacency matrix* A = A(G). This matrix is also of importance in chemistry and physics. The vertex-adjacency matrix A(G) of a labeled connected graph G with N vertices is the square N x N symmetric matrix which contains information about the internal connectivity of vertices in G.

- Vertex-adjacency matrix is defined as,

$$A_{ij} = \{1\ if, and\ only\ if (i,j) \in E(G)\ 0\ otherwise \quad (1)$$

$$A_{ii} = 0 \quad (2)$$

Therefore, a nonzero entry appears in A(G) only if an edge connects vertices *i* and *j*. For example, the following vertex-adjacency matrix can be constructed for a labeled graph G (Figure 12).



$$A(G) = \begin{array}{c} \\ ① \\ ② \\ ③ \\ ④ \end{array} \begin{array}{c} \begin{array}{cccc} ① & ② & ③ & ④ \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

Figure 12 A vertex and edge labelled graph G

The adjacency matrix is symmetrical about the principal diagonal. Therefore, the transpose of the adjacency matrix A leaves the adjacency matrix unchanged,

$$A^T(G) = A(G) \quad (3)$$

This transpose A^T is formed by interchanging rows and columns of the matrix A.

The *edge-adjacency matrix* of a graph G, EA = EA(G), is determined by the adjacencies of edges in G. It is very rarely used.

- The edge-adjacency matrix is defined as

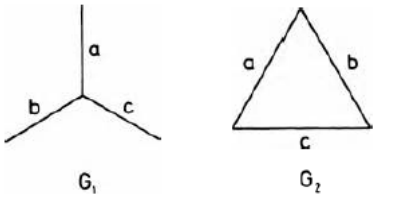
$$({}^EA)_{ij} = \{1\ if\ edges\ e_i and\ e_j\ are\ adjacent\ \ 0\ otherwise \quad (4)$$

$$({}^EA)_{ii} = 0 \quad (5)$$

For example, the following edge-adjacency matrix can be constructed for a labeled graph G in Figure 12:

$$
{}^{E}\mathbf{A}(G) = \begin{array}{c} \\ \text{a} \\ \text{b} \\ \text{c} \\ \text{d} \end{array}
\begin{array}{c}
\begin{array}{cccc} \text{a} & \text{b} & \text{c} & \text{d} \end{array} \\
\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}
\end{array}
$$

Although both the vertex-adjacency matrix and the edge-adjacency matrix reflect the topology of a molecule, they differ in their structure. However, it should be noted while the vertex-adjacency matrix uniquely determines the graph, the edge-adjacency matrix does not. In other words, there are known non-isomorphic graphs with identical edge-adjacency matrices. A pair of such non-isomorphic graphs is shown in Figure 13. The corresponding edge-adjacency matrix is given by



$$
{}^{E}\mathbf{A}(G_1) = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = {}^{E}\mathbf{A}(G_2)
$$

Figure14 A pair of nonisomorphic graphs (G_1 and G_2) which possess the identical edge-adjacency matrix.

Graphs Gi and G2 have obviously different vertex-adjacency matrices.

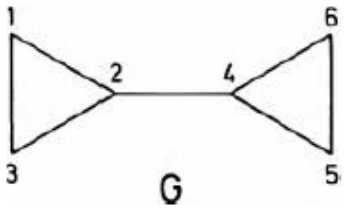
The Distance Matrix

The *distance matrix* (which is also sometimes called the metrics matrix) is, in a sense, a more complicated and also a richer structure than the adjacency matrix. It is a graph-theoretical (topological) matrix less common than the adjacency matrix, but it has been increasingly used in the last two decades. in many different areas of chemistry and physics. It has been pointed out an interesting fact that the distance matrix has also found considerable use in the areas of research which are relatively remote from chemistry and physics and to a great extent non mathematical, such as anthropology, geography, geology, ornithology, philology, and psychology.

- The distance matrix D = D(G) of a labeled connected graph G is a real symmetric N x N matrix whose elements $(D)_{ij}$ are defined as follows:

$$(D)_{ij} = \{e_{ij,}\ if\ i \neq j\ \ 0, if\ i = j \qquad (6)$$

where $e_{ij,}$ is the length of the shortest path (i.e., the minimum number of edges) between the vertices v_i , and v_j. The length is also called the *distance* between the vertices v_i, and v_j thence the term distance matrix. For example, the following distance matrix can be constructed for a labeled graph G (Figure 15):



$$D(G) = \begin{array}{c} \\ ① \\ ② \\ ③ \\ ④ \\ ⑤ \\ ⑥ \end{array} \begin{array}{c} \begin{array}{cccccc} ① & ② & ③ & ④ & ⑤ & ⑥ \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 2 & 3 & 3 \\ 1 & 0 & 1 & 1 & 2 & 2 \\ 1 & 1 & 0 & 2 & 3 & 3 \\ 2 & 1 & 2 & 0 & 1 & 1 \\ 3 & 2 & 3 & 1 & 0 & 1 \\ 3 & 2 & 3 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

Figure 15 A label graph G

The distance matrix has found a widespread application in chemistry in both explicit and implicit forms. The first explicit use of the distance matrix was employed for studying the permutational isomers of stereo chemically nonrigid molecules. The distance matrix in explicit form is also used to generate the ***distance polynomial*** and the ***distance spectrum*****.**

Topological indices

- A single number that can be used to characterize the graph of a molecule is called a *topological index.* (The term *graph-theoretical index* would be more accurate than topological index, but the latter is more common in the chemical literature.)

A topological index, thus, appears to be a convenient device for converting chemical constitution into a number. Evidently, this number must have the same value for a given molecule regardless of ways in which the corresponding graph is drawn or labeled. Such a number is referred to by graph theorists as a *graph invariant*. For example, one of the simplest graph invariants (topological indices) is the number of vertices in the graph (the number of atoms in the molecule). Hence, it could be simply said that topological indices are graph invariants. It should also be pointed out that topological indices do not generally allow the reconstruction of the molecular graph, implying that a certain loss of information has occurred during their creation.

Topological indices were introduced 150 years ago, and the very fact that they are still in use today is demonstration of their durability and versatility. There are more than 120 topological indices (including information-theoretic indices) available to date in the literature, with no sign that their proliferation will stop in the near future. This large (and every increasing) number of topological indices indicates that perhaps a clear and unambiguous criterion for their selection and verification is still missing, although some attempts along these lines have been reported. Moreover, a large number of topological indices also lead to a question to what extent are they orthogonal? In other words, is it possible that some topological indices

express predominantly the same type of constitutional information: the difference residing in the scaling factor? Several analyses on the example of alkane trees with up to 12 vertices indicate that a number of topological indices are strongly intercorrelated i.e., that many of them contain to a great extent the same type of structural information.

One of the ultimate targets of theoretical chemists is to build schemes that would allow accurate predictions of the bulk properties of matter from the knowledge of molecular structure. We are still far away from this ideal., but one way of trying to achieve this goal is by means of topological indexes since they serve as convenient descriptors of molecular structure.

Most of the proposed topological indices are related to either a vertex adjacency relationship (connectivity) in the molecular graph G or to graph- theoretical (topological) distances in G. Therefore, the origin of topological indices can be traced either to the adjacency matrix of a molecular graph or to the distance matrix of a molecular graph. Furthermore, since the distance matrix can be generated from the adjacency matrix, most of the topological indices are really related to the latter matrix.

The interest in topological indices is mainly related to their use in *nonempirical* quantitative structure-property relationships (QSPR) and quantitative structure-activity relationships (QSAR). The latter use in such areas as pharmacology, toxicology, environmental chemistry, and drug design is intensively studied by many researchers.

Examples of topological indices

Zagreb Indices

The first and second Zagreb indices (M1 and M2) are another set of classic vertex-based descriptors developed in 1972 and 1975, respectively. They were called the Zagreb group indices as their authors were members of the "Rudjer Bošković" Institute in Zagreb, Croatia. In these indices one counts the connections from each vertex (node, carbon).

- The first Zagreb index M1(G) is equal to the sum of squares of the degrees of the vertices, and the second Zagreb index M2(G) is equal to the sum of the products of the degrees of pairs of adjacent vertices of the underlying molecular graph G.

$$M_1 = \sum_{i=1}^{n} \delta_i^2$$

$$M_2 = \sum \delta_i \delta_j$$

For pentane, each would be calculated as:

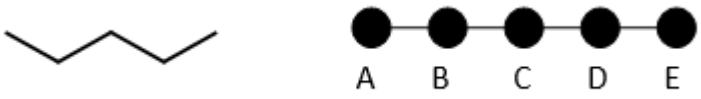


Figure 16 A label graph of pentane

$M_1 = 1^2 + 2^2 + 2^2 + 2^2 + 1^2 = 1 + 4 + 4 + 4 + 1 = 14$

$M_2 = 1x2 + 2x2 + 2x2 + 2x1 = 2+4+4+2 = 12$

For 2-methylpentane, each would be calculated as:

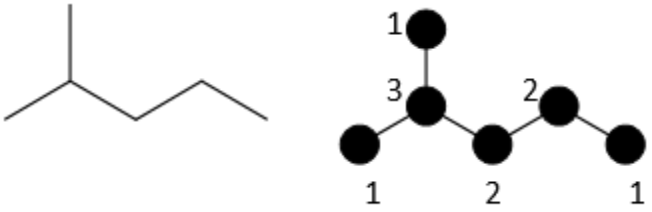


Figure 17 A label graph of 2-methylpentane

$M_1 = 1^2 + 1^2 + 3^2 + 2^2 + 2^2 + 1^2 = 1 + 1 + 9 + 4 + 4 + 1 = 20$

$M_2 = 1x3 + 1x3 + 3x2 + 2x2 + 2x1 = 3+3+6+4+2 = 18$

There are thousands of 2D descriptors that are frequently applied in modeling or predicting properties or biological functions. What is interesting is that these graphs are often descriptors that are reduced to a single value that can be used to make meaning of the physical world. Zagreb group indices were introduced to *characterize branching*.

Wiener Index

One of the first mathematical representations of chemical structure used for prediction of properties was developed in 1947 by Harold Weiner. It is defined at the sum of distances between any two carbon atoms (pairs of nodes) in the molecule.

- Mathematically it is represented as:

$$W(G) = \frac{1}{2} \sum_{u,v \in G} d(u,v)$$

Where G represents the total atoms in the molecule, u and v are individual carbon atoms and *d*(u,v) is the distance in bonds between any two carbon atoms in the shortest path between any two atoms. In using this index, Weiner showed that the index value is closely correlated with the *boiling point of a series of alkanes*. Further work also showed that it correlated with other physical properties such as *density, surface tension and viscosity.*

To calculate the Wiener index for a molecule, for each pair of atoms in the structure, count the distance between atoms. Take the sum of all distances and divide by two. For example in the case of ethane, which only has two nodes:

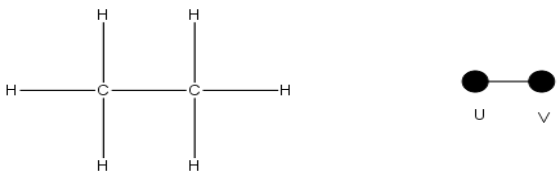


Figure 18 A label graph of ethane

	u	**v**
u	0	1
v	1	0

$$W(G) = \frac{1}{2}(1+1) = \frac{1}{2}(2) = 1$$

Pentane has 5 nodes, and distances between each node are calculated and summed.

	A	**B**	**C**	**D**	**E**	**total**
A	0	1	2	3	4	10
B	1	0	1	2	3	7
C	2	1	0	1	2	6
D	3	2	1	0	1	7
E	4	3	2	1	0	10

$$W(G) = \frac{1}{2}(10+7+6+7+10) = \frac{1}{2}(40) = 20$$

The Platt Number

Platt was also interested in devising a scheme for predicting physical parameters *(molar volumes, boiling points, heats of formation, heats of vaporization*) of alkanes. He introduced an index F = F(G), which is equal to the total sum of edge-degrees in a graph G. The *edge-degree* of an edge *e*, D(*e*), is the number of its adjacent edges. This index was named the *Platt number*.

- The Platt number of a graph G is defined by

F(G) =$\sum_{i=1}^{M} D(e_i)$, The Platt number, thus represents the first neighbor's sum.

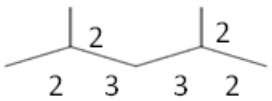


Figure 19 A label graph of 2, 4 di-methyl pentane

The Largest Eigenvalue

- The characteristic (spectral) polynomial P(G;x) of a graph G is the characteristic polynomial of its adjacency matrix,

P(G;x) = det |xI - A|

where A and I are, respectively, the adjacency matrix of a graph G with N vertices and the N x N unit matrix. A graph eigenvalue xi is a zero of the characteristic polynomials. P(G;xi) = 0

for i = 1, 2 , N. The complete set of graph eigenvalues {x1, x2, . . . xN} forms the spectrum of the graph. The eigenvalues are all real and the interval in which they lie is bounded.

- According to the Frobenius theorem, the limits of the graph spectrum are determined by the maximum valency of a vertex Dmax in a graph: $-Dmax \leq xi \leq Dmax$

The largest eigenvalue, xi, in the graph spectrum may be used as a topological index. For example, it has been found that xi can be employed as a *measure of branching* and that (alkane) trees can be well ordered according to xi. In Figure 20 as an example, the ordering of alkane trees with seven vertices is shown. The smallest value of xi belongs to C7 chain and the largest value of xi to the most branched C7 alkane tree. The largest eigenvalue is not a very discriminative index, because in many cases the same xi value belongs to two (or more) non-isomorphic molecular graphs. One such degenerate pair appears also in the alkane trees shown in Figure 20.

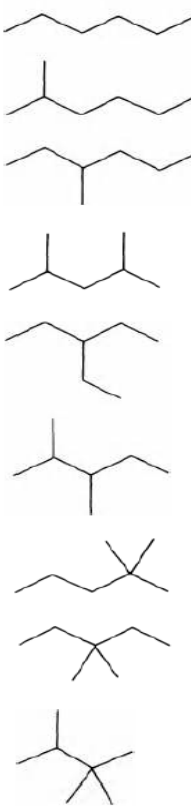


Figure 20 The ordering of alkane trees with seven vertices according to the increasing value of xi. This order follows the intuitive notion of branching

QSAR/QSPC concept for Insilco prediction of properties

- Quantitative structure property relationships (QSPR) and, when applied to biological activity, quantitative structure activity relationships (QSAR) are methods for determining properties due to very sophisticated mechanisms purely by a curve fit of that property to aspects of the molecular structure.

This allows a property to be predicted independent of having a complete knowledge of its origin. For example, drug activity can be predicted without knowing the nature of the binding site for that drug. Structure–property relationships are qualitative or quantitative empirically

defined relationships between molecular structure and observed properties. In some cases, this may seem to duplicate statistical mechanical or quantum mechanical results. However, structure-property relationships need not be based on any rigorous theoretical principles. The simplest case of structure-property relationships a qualitative rule of thumb. For example, the statement that branched polymers are generally more biodegradable than straight-chain polymers is a qualitative structure–property relationship. When structure-property relationships are mentioned in the current literature, it usually implies a quantitative mathematical relationship. Such relationships are most often derived by using curve-fitting software to find the linear combination of molecular properties that best predicts the property for a set of known compounds. This prediction equation can be used for either the interpolation or extrapolation of test set results. Interpolation is usually more accurate than extrapolation. When the property being described is a physical property, such as the boiling point, this is referred to as a quantitative structure–property relationship (QSPR). When the property being described is a type of biological activity, such as drug activity, this is referred to as a quantitative structure–activity relationship (QSAR). Our discussion will first address QSPR. All the points covered in the QSPR section are also applicable to QSAR, which is discussed next.

QSPR/QSAR

A stepwise procedure of QSPR is given below.

- Step 1- Compilation of list of compounds and their experimental property value

Compilation of list of compounds for which the experimentally determined property is known. Ideally, this list should be very large. Often, thousands of compounds are used in a QSPR study. If there are fewer compounds on the list than parameters to be fitted in the equation, then the curve fit will fail. If the same number exists for both, then an exact fit will be obtained. This exact fit is misleading because it fits the equation to all the anomalies in the data, it does not necessarily reflect all the correct trends necessary for a predictive method. In order to ensure that the method will be predictive, there should ideally be 10 times as many test compounds as fitted parameters. The choice of compounds is also important. For example, if the equation is only fitted with hydrocarbon data, it will only be reliable for predicting hydrocarbon properties.

- Step 2- Geometry of molecule

To obtain geometries for the molecules. Crystal structure geometries can be used; however, it is better to use theoretically optimized geometries. By using the theoretical geometries, any systematic errors in the computation will cancel out. Furthermore, the method will predict as yet un-synthesized compounds using theoretical geometries. Some of the simpler methods require connectivity only.

- Step 3- Molecular Descriptor

Molecular descriptors must then be computed. Any numerical value that describes the molecule could be used. Many descriptors are obtained from molecular mechanics or semiempirical calculations. Energies, population analysis, and vibrational frequency analysis with its associated thermodynamic quantities are often obtained this way. Ab initio results

can be used reliably, but are often avoided due to the large amount of computation necessary. The largest percentage of descriptors are easily determined values, such as molecular weights, topological indexes, moments of inertia, and so on.

- Step 4- Correlation

Once the descriptors have been computed, is necessary to decide which ones will be used. This is usually done by computing correlation coefficients. Correlation coefficients are a measure of how closely two values (descriptor and property) are related to one another by a linear relationship. If a descriptor has a correlation coefficient of 1, it describes the property exactly. A correlation coefficient of zero means the descriptor has no relevance. The descriptors with the largest correlation coefficients are used in the curve fit to create a property prediction equation. There is no rigorous way to determine how large a correlation coefficient is acceptable.

Intercorrelation coefficients are then computed. These tell when one descriptor is redundant with another. Using redundant descriptors increases the amount of fitting work to be done, does not improve the results, and results in unstable fitting calculations that can fail completely (due to dividing by zero or some other mathematical error). Usually, the descriptor with the lowest correlation coefficient is discarded from a pair of redundant descriptors.
A curve fit is then done to create a linear equation, such as Property = c0 + c1d1 + c2d2 +
where ci are the fitted parameters and di the descriptors. Most often, the equation being fitted is a linear equation like the one above. This is because the use of correlation coefficients and linear equations together is an easily automated process. Introductory descriptions cite linear regression as the algorithm for determining coefficients of best fit, but the mathematically equivalent matrix least- squares method is actually more efficient and easier to implement. Occasionally, a nonlinear parameter, such as the square root or log of a quantity, is used. This is done when a researcher is aware of such nonlinear relationships in advance.

- QSAR is also called traditional QSAR or Hansch QSAR to distinguish it from the 3D QSAR method. This is the application of the technique described above to biological activities, such as environmental toxicology or drug activity.

In order to parameterize a QSAR equation, a quantified activity for a set of compounds must be known. These are called lead compounds, at least in the pharmaceutical industry. Typically, test results are available for only a small number of compounds. Because of this, it can be difficult to choose a number of descriptors that will give useful results without fitting to anomalies in the test set. Three to five lead compounds per descriptor in the QSAR equation are normally considered an adequate number. If two descriptors are nearly col- linear with one another, then one should be omitted even though it may have a large correlation coefficient.
In the case of drug design, it may be desirable to use parabolic functions in place of linear functions. The descriptor for an ideal drug candidate often has an optimum value. Drug activity will decrease when the value is either larger or smaller than optimum. This functional form is described by a parabola, not a linear relationship.
The advantage of using QSAR over other modelling techniques is that it takes into account the full complexity of the biological system without re- quiring any information about the binding site. The disadvantage is that the method will not distinguish between the contribution of binding and trans- port properties in determining drug activity. QSAR is very

useful for deter- mining general criteria for activity, but it does not readily yield detailed structural predictions.

Predicting Molecular Geometry

Computing the geometry of a molecule is one of the most basic functions of a computational chemistry program. However, it is not trivial process. The user of the program will be able to get their work done more quickly if they have some understanding of the various algorithms within the software. The user must first describe the geometry of the molecule. Then the program computes the energies and gradients of the energy to find the molecular geometry corresponding to the lowest energy.

- **Specifying Molecular Geometry**

One way to define the geometry of a molecule is as a set of Cartesian coordinates for each atom. Graphic interface programs often generate Cartesian coordinates since this is the most convenient way to write those programs. It is becoming more common to uses programs that have a graphical builder in which the user can essentially draw the molecule. There are several ways in which such programs work. Some programs allow the molecule to be built as a two-dimensional stick structure and then convert it into a three-dimensional structure. Some programs have the user draw the three-dimensional backbone and then automatically add the hydrogens. This works well for organic molecules. Some programs build up the molecule in three dimensions starting from a list of elements and hybridizations, which can be most convenient for inorganic molecules. Many programs include a library of commonly used functional groups, which is convenient if it has the functional groups needed for a particular project. A number of programs have specialized building modes for certain classes of molecules, such as proteins, nucleotides, or carbohydrates.

- **Geometry optimization**

Many computational chemistry programs will do the geometry optimization in Cartesian coordinates. This is often the only way to optimize geometry in molecular mechanics programs and an optional method in orbital-based programs. A Cartesian coordinate optimization may be more efficient. Cartesian coordinates are often preferable when simulating more than one molecule since they allow complete freedom of motion between separate molecules. Geometry optimizations that run poorly either take a large number of iterations or fail to find an optimized geometry. There are many different algorithms for finding the set of coordinates corresponding to the minimum energy. These are called *optimization algorithms* because they can be used equally well for finding the minimum or maximum of a function.

Example of softwares for coordinate generation and geometry optimization are Discovery studio, Avogadro, CORINA etc

The materials in Unit-II are compiled from different reference sources.

1. COMPUTATIONAL CHEMISTRY, A Practical Guide for Applying Techniques to Real-World Problems, Darid C. Young
2. Essentials of Computational Chemistry by Christopher J. Cramer
3. Molecular Interactions and the Behaviors of Biological Macromolecules, by Loren Dean Williams https://williams.chemistry.gatech.edu/structure/molecular_interactions/mol_int.html
4. Chemical graph theory, Nenad Trinajstic, Ph.D., Professor of Chemistry, The Rugjer Boskovic Institute Zagreb, The Republic of Croatia

5. Topological Index calculator:
   https://www.cs.gordon.edu/courses/organic/topo/manual-v3/manual.html

Syllabus

Computational chemistry: Scope, cost and efficiency of computational modeling. Stabilizing interactions: Bonded and non-bonded interactions. Molecular topology, topological matrix representation, topological indices, QSAR/QSPC concept for insilico prediction of properties. 3D co-ordinate generation for small molecules, geometry optimization.