Chapter 1

Introduction
The discovery and development of new drug is a complex process and mainly based on trial and error method in pharmaceutical research. It is reported that for a new drug from lead identification to clinical trials needs about fourteen years of time\(^1\) on cost of about 800 million US dollars\(^2\). Therapeutic effects and hazards to health are assessed using a series of experimental and \textit{in vivo} tests. Moreover, usage of animal models is often subject to ethical as well as financial consideration. Therefore, alternative methods are being developed to reduce the requirement of animals testing, cost and time. Chemometric methods are often implemented in drug discovery pipeline due to their lower cost and time; an added bonus is their significant contribution to identification and development of effective drugs from new chemical entities (NCE). Computational techniques combine with chemistry, which describe chemical phenomena and development of more intuitive informatics interface that provide a new set of cheminformatics tools to the experimental scientists. These tools are used in conjunction with traditional research techniques to predict properties and activities for NCE. The technique which used in development of molecular modeling program and successfully applied in drug discovery research is known as chemometric technique. Chemometric approach is a relatively new field with outstanding opportunity mainly virtual design of new useful compounds with well defined properties that reduce high cost of experimental research in the drug development strategy. In this context, the characterization of drug–target interactions (e.g., molecular docking), and the assessment and optimization of drug activity using quantitative structure activity relationships (QSARs) along with pharmacophore modeling to visualize the important 3D representation of molecular features are already proved useful in the drug discovery process\(^3\). This is also crucial to emphasize that chemometric tool can not replace the laboratory experiment but it may be considered as additional supporting tool to gain a better insight into the chemistry and biology of the problem at hand by generating and analyzing a large volume of information in comparatively lesser period of time. Due to massive advantages of chemometric technique over traditional drug discovery method, this approach is used in wide range of diseases to develop potent and safer drug candidate molecules\(^4\).

In a report of World Health Organization (WHO), cancer, the leading life threatening diseases worldwide, accounted for 7.4 million deaths (~ 13% of all deaths) in 2004. In
addition to breast cancer, lung, stomach, liver and colon cancer are the most common form of cancer deaths in each year. Estrogen, an essential endocrine hormone plays crucial role in the post-menopausal diseases that include hormone responsive breast cancer, cardio-vascular diseases and osteoporosis. Chemometric technique can be applied to the estrogenic activity of the chemicals to design potential drug molecules for the estrogen related diseases.

Estrogens comprise essential biochemical components in the female reproductive system and also in the maintenance of diverse range of non-reproductive tissues. It is now proved that ovarian hormones, particularly estrogen is mainly responsible for the development of breast cancer. Laboratory evidence identified estrogen as the trophic hormone in estrogen target tissues (e.g. the uterus and some breast cancers), so “anti-estrogen” therapy becomes a central theme for the treatment and prevention of breast cancer. Drug substances mimicking or antagonizing effects of endogenous estrogens are being developed with a view to either alter the physiological process of reproduction or to achieve protective effects on certain vital organs and organ systems which may assist in improve quality of life through preventing degeneration or excessive proliferation of organ system. Use of chemometric techniques for the bioactivity and subsequent design and development of efficacious compounds based on that knowledge might prove to be of significant asset towards benefit of women community.

1.1 Estrogen: Estrogen is a group of chemically steroidal hormone secreted by the ovaries and testis with involvement of placenta, adipose tissue, and adrenal glands. The biosynthesis of estrogen is shown in Figure 1.1. It is reported that estrogen develops and maintains the reproductive functions, plays an important role in bone formation and reduction of bone resorption, retention of salt (sodium) and water. It has also beneficial effects on cardiovascular system by increasing the level of high-density lipoprotein (HDL) and decreasing the level of low-density lipoprotein (LDL). The level of estrogen decreases in women’s body due to mainly two reasons, viz., one is in menopause and secondly the women who are likely to live a substantial part of their lives in a state of estrogen deficiency due to increasing life expectancy. At low level of estrogen in women body several post-menopausal symptoms like hot-flushes, vaginal
atrophy and sleeping disturbance arise, and also rise of LDL that progressively increases the chance of coronary and osteoporosis diseases. Among the several structurally related forms, 17β-estradiol (E2) is found as predominant. The hormone replacement therapy (HRT) in which synthetic estrogens are administered into the body that reduce osteoporotic fractures, improve severe menopausal symptoms and use as complement to urogenital atrophy during menopause. Moreover, the protective effect of HRT on enhancing bone density is clearly demonstrated in the post menopausal/progestin intervention (PEPI) trial. But on other hand malevolent aspect of HRT is increasing chance of breast and uterine cancers. The most common uses of exogenous estrogen agonists are for HRT and contraception, while anti-estrogens are used in the treatment of hormone responsive breast cancer and female infertility. It is now became accepted that ovarian hormones, particularly estrogens are mainly responsible for the development of breast cancer. Thus it is desirable to block estrogen action for cancer treatment and prevention. Estrogen also exerts positive effects upon overall health, including prevention of osteoporotic fracture and reduced cardiovascular disease. Therefore, it is desirable to either supply estrogens, or supply a non-steroidal drug molecule that can also mimic the estrogen action.

Figure 1.1: Biosynthesis of estrogen
Estrogen agonists are compounds that mimic the effect of endogenous estrogens, while estrogen antagonists or anti-estrogens block the binding of $E_2$ (1.1) to the estrogen receptor\textsuperscript{20}. The ICI182,780\textsuperscript{21} (1.2) is a major steroidal anti-estrogen that has been developed with pure anti-estrogen activity in all tissues, while MER25 (1.3) is the first reported nonsteroidal anti-estrogen\textsuperscript{22}. The MER25 could not enter the clinics due to its low potency and serious side effects\textsuperscript{23}. However, the tamoxifen (1.4) is the first model anti-estrogen for the therapy of breast cancer and having structural similarity to MER25, and is used in the painless treatment of advanced breast cancer since several decades\textsuperscript{24,25}. Tamoxifen had been studied expansively in clinical trials for 20 years; consequently the drug became the agent of preference to attest\textsuperscript{26} the significance of an anti-estrogen for the prevention of breast cancer in high-risk women.

![Chemical Structures]

1.1.1 Estrogen Receptor: The biochemical effect of estrogen at the level of gene regulation is mediated by estrogen receptor (ER), which is a ligand-inducible intracellular transcription factor. It is reported that ER predominantly constitutes one of the important member of the nuclear receptor super family\textsuperscript{13,27}. Nuclear hormone receptors (NHR) are a family of hormone activated transcription factors that can initiate or enhance the genes containing specific hormone response elements. The human ER (hER), which belongs to this nuclear-receptor family and was successfully cloned and sequenced from MCF-7 human breast cancer cell\textsuperscript{28,29}. In the nucleus, the ER up- or down-regulates the expression of the target genes by interacting through its site-specific DNA and with other coregulatory proteins that include co-activators and co-repressors\textsuperscript{30,32}. Like other members of the NHR, ER comprises several domains (Figure 1.2) that serve specific roles\textsuperscript{33}. Starting from amine
The functional domains are (i) ligand independent N-terminal domain (NTD), (ii) central DNA binding domain (DBD) and (iii) the ligand-binding domain (LBD). The activation function (AF) domains, AF-1 and AF-2, located within NTD and LBD respectively. These AFs are responsible for regulating the transcriptional activity of ER\textsuperscript{34}. Full transcription activity of the ER is thought to be achieved by synergism between the two AFs, and their activities are promoter and cell specific\textsuperscript{35}. AF-1 functions as hormone independent, whereas AF-2 function requires the presence of hormone/steroid\textsuperscript{34, 36}. The LBD alters conformation on binding of an agonist that subsequently permits recruitment of one or more coactivators. This complex activates DNA transcription by binding to an estrogen response element (ERE)\textsuperscript{37-40}. The agonist is entirely enclosed by ER and formed part of the hydrophobic core of the protein. The orientation of helix-12, located at the carboxy-terminus of the LBD, is fundamental in distinguishing the functions between agonist and antagonist. Antagonist blocks access to a groove located between helices 3, 4 and 5, the binding site for coactivators during transcription (AF-2 site)\textsuperscript{41}. In point of fact, the AF-2 undergoes a distinct conformational transformation in presence of several ligands and determines the subsequent binding of coactivators (augmenting the activity of receptors) or co-repressors (mediating the repressive effects of receptors)\textsuperscript{42, 43}. It has been suggested that ER action modulates the rate of transcription initiation through interactions with the basal transcription machinery and changes in the state of chromatin arrangement at the promoter of target genes via the recruitment of a variety of co-activators\textsuperscript{44}. Although the experimental facts suggest that activating functions can be disallowed in the anti-estrogen-ER complex by stopping co-activator binding. In another hypothesis, an increased binding of proteins that only interact with the anti-estrogen-ER complex with co-repressors neutralize an anti-estrogen-ER complex\textsuperscript{45}.

![Figure 1.2: 1°, 2° and 3° structure of nuclear receptor](image-url)
Until 1995, it was assumed that there was only one type of ER that existed and was responsible for mediating all of the physiological and pharmacological effects of natural and synthetic estrogens and antiestrogens. In 1995, Kuiper cloned a second type of ER from a rat prostate cDNA library and coined it as ERβ, while earlier it was termed as ERα. Both ERα and ERβ have a similar modular architecture (Figure 1.3) to the other members of the steroid/thyroid receptor family. In terms of sequence homology, the DBD in ERβ shows a high homology to ERα with more than 95% amino acid identity and in the LBD about 55% amino acid identity. However, the NTD of ERβ is shorter than that of ERα with a very poor sequence homology of only about 15% compared to that of ERα. The three-dimensional (3D) structures of the independently expressed DBD and LBD have been solved and show overall folds that represent globular proteins with natively ordered conformations. To date, no 3D natively folded structure for the NTD is available not only for the ER but for the entire NHR superfamily. It was known for some time that both ERα and ERβ can promote neuronal survival by activating estrogen mechanisms in rat hippocampal neurons. But growing evidence suggests that ERβ is basically the fundamental requirement for activation of mechanisms that bring about estrogen inducible neuronal morphological change, brain development, and cognition. As such, these facts institute a possible therapeutic application for ERβ as a target to promote memory function and neuronal defense mechanisms against old age neurodegeneration such as Alzheimer's disease (AD). Other therapeutic advantages related with ERβ are regulation of estrogen vasculoprotective action and development of interventions targeting diseases, such as depression, colon and prostate cancer, obesity, leukemia, and infertility. However, a disadvantage of an ERβ-selective ligand could be lack of activation of ERα in bone, as ERα has been demonstrated to mediate estrogen regulation of bone density.

Figure 1.3: The domain structures of the ERα and ERβ (adopted from Nilsson et al.)
Thus, it may be possible that selective binding to and/or activation of one or the other of these two ER subtypes could attain some of the tissue-selective effects of estrogens. This prospect has initiated efforts in research laboratories to identify ER subtype-specific ligands\textsuperscript{62-65}. The ability of the two subtypes to form heterodimers suggest that ER may function through different dimeric states, and it may also be possible that the dimers could be activated by selective ligands\textsuperscript{66}. A second binding site has been identified within the LBD of both ER\textsubscript{a} and ER\textsubscript{b}. This site is suggested to be located in close proximity to the steroid binding site of both ER subtypes\textsuperscript{41,67}.

1.1.2 Estrogen receptor modulators: Estrogen receptor modulators (ERMs) bind with ER that induces conformational changes in the receptor. These conformations may interact with cell and tissue-specific coactivating or corepressing proteins or even estrogen response elements (EREs) at receptor site, which lead to produce diverse biological effects\textsuperscript{68}. Basically ERMs represent a group of chemical compounds that exert agonist or antagonist effects on various estrogen target tissues\textsuperscript{69}. ERMs being are evaluated for a number of estrogen related disorders, including post-menopausal osteoporosis, hormone-responsive breast cancer and cardiovascular diseases\textsuperscript{70, 71}. In the United States, approximately 90 million prescriptions for HRT were dispensed annually from 1999 through 2002\textsuperscript{72}. Indeed, records suggest that HRT was the most commonly prescribed medicine in the world during the late 1990s and early 2000s\textsuperscript{73}. Therefore, as part of the Women's Health Initiative (WHI), a large randomized controlled primary prevention trial to determine the risk benefit ratio of HRT in postmenopausal women was undertaken. In July'2002, the WHI study for examining the effects of HRT are reported\textsuperscript{14} that indicated approximately 16,000 women are treated either with estrogen/progesterone combination HRT or placebo in which an approximately 26% increase in the incidence of breast cancer is detected. In Million Women Study\textsuperscript{74}, which is not a randomized prospective clinical trial but followed a group of post-menopausal women during the same time frame as the WHI, collected information about the use of HRT. The overall conclusion is that users of HRT are more likely than never users of HRT to develop breast cancer and die from it\textsuperscript{74}.
The ICI-164,184 (1.5), which is known as agonist ERM and used to block the actions of E$_2$. The Schering Plough evaluated the EM80076 (1.6) and its metabolite EM652 (1.7) as preventive agents in breast cancer. Anti-estrogen ICI-182,780 (1.8) developed by AstraZeneca Pharmaceuticals is the derivative of E$_2$ with a long hydrophobic side chain and demonstrates exciting anti-tumor activity in pre-clinical models. Phytoestrogens are a diverse group of estrogenic compounds obtained from plants. The presence of such compounds in the human diet proves to be beneficial and may even confer reduced risk of hormone-dependent breast cancer and heart disease and improve symptoms associated with menopause. Genistein (1.9), an isoflavonoid phytoestrogen that is found at significant levels in soya beans and soy products, binds to both ER isoforms with moderate affinity but exhibits a preference for ER$_{B}$. The DT56a (Femarelle), a natural ERM is shown to activate ER in human cultured female derived osteoblasts. It is also shown to relieve menopausal symptoms and to increase bone mineral density with no effect on sex steroid hormone levels and on the endometrial thickness. DT56a, similarly to E$_2$, stimulated the specific activity of creatine kinase (CK) in skeletal and vascular tissues of female rats, as a marker of ER activation. Deoxybenzoins, (1.10, 1.11) found in natural sources are being used as antiestrogen in breast cancer cells, exerting estrogenic effect on osteoblasts without causing stimulation of endometrial cells. It has the potential to be considered as breast anticancer agent and an alternative to hormone replacement therapy.
1.1.3 Selective estrogen receptor modulators: Opportunities in the selective therapeutics are developed on advancement of the idea of multifunctional medicines. Selective estrogen receptor modulators (SERMs) are most successful therapeutic agents used for treatment of both post-menopausal symptoms and hormone-responsive breast cancer. It is evident that SERMs, with activities ranging from nearly full estrogenic activity to almost pure anti-estrogenic activity, can be developed and used for specific therapeutic purposes. SERMs elicit biological response by changing a unique conformation in the receptor-ligand dependent transcriptional activation function. According to Sun et al., SERMs may exhibit preference over the two isoforms of ER. Some successful SERMs include tamoxifen (1.4), raloxifene (1.12) and toremifene (1.13) are classified in generation, suggesting a progressive development in a process intended to improve the beneficial effects while reducing the harmful side effects associated with the earlier ERMs. These are implemented clinically in the therapy of osteoporosis and hormone dependent breast cancer. Raloxifene is a greatly effective anti-estrogen in the reproductive tissues, but acts as partial ER agonist in bone and also lowers blood cholesterol. Tamoxifen is hydroxylated by CYP2D6 to 4-hydroxytamoxifen (1.14), which has a higher binding for the ER than the parent drug. Derivatives of tamoxifen and their metabolites are in clinical testing for the treatment of advanced breast cancer in post-menopausal women.
1.2 Molecular modeling: Traditional drug discovery includes several uncertain phases like synthesis, biological testing and analysis. It is an expensive process, requiring about 800 million US dollars$^2$ and takes about 14 years$^1$ for a drug to reach the market from its initial discovery stage. In the early 1990s, rapid development in the fields of combinatorial chemistry and high throughput screening (HTS) technologies have created an environment for expediting the discovery process by enabling huge libraries of compounds to be synthesized and screened in short period of time$^9$.$^8$. Thus rational approach to drug discovery emerged in the pharmaceutical industry and has contributed to the rapid development of molecular modeling.

To find out a lead compound for a particular target, the screening of compound databases is presently the most popular and useful cheminformatics application in pharmaceutical discovery. Chemometric technique can be used to design compounds with interesting physicochemical characteristics, as well as systematic assessment of potential lead candidates before they are synthesized and tested$^9$.$^8$. The foundations of chemometric method were established in the early 1970s with the use of structural biology to modify the biological activity of insulin$^9$.$^9$ and to guide the synthesis of human haemoglobin ligands$^{10}$.$^{100}$. Combination of X-ray crystallography and comparative modeling based on natural structure with advances in combinatorial chemistry, HTS technologies and computational infrastructures have rapidly bridged the gap between theoretical modeling and medicinal chemistry. Chemometric technique now plays a critical role in search of new molecular entities$^{10}$.$^{101}$-$^{103}$, and also focuses on improved design and management of data sources, creation of computer programs to generate libraries of pharmacologically interesting compounds by development of new algorithms to assess the potency and selectivity of lead candidates, and design of predictive tools to identify potential ADME/Tox liabilities. Chemometric techniques can be categorized into two broad ways, direct or structure-based method and indirect or ligand-based method. In structure-based approach, the 3D structure of the receptor is considered and find out the correct orientation of the ligand. In ligand-based approach, the design is based on the comparative analyses of the structural features of known active and inactive molecules that are interpreted in terms of their complementarity with a hypothetical receptor site.
model. The schematic representation of chemometric drug design approach is depicted in Figure 1.4.

Based on “biological activity of the molecules is the function of chemical structures”, principle structure activity relationship (SAR) is established qualitatively or quantitatively between biological activity and chemical structure of molecules. It is not based on any rigorous theoretical principles, rather it is the visualization of 3D structure of molecule with associated properties, such as geometries, electrostatic and hydrophobicity, which provide useful means in understanding some of the driving forces operating in life processes. Such information allows to rationalize SAR and has proven to reduce the role of empiricism in the design of new prototype drugs.

![Figure 1.4: Schematic representation of chemometric drug design approach](image)

In 1960s Hansch and co-workers\textsuperscript{104} successfully applied quantitative SAR (QSAR) models in various areas in chemistry and biology\textsuperscript{105}. In QSAR study molecular descriptors are generated followed by development of statistical relationship between descriptors and biological activity is established.

Hansch model is based on the determination of mathematical equation expressing biological activity in terms of molecular parameters such as hydrophobicity, steric
substituent constant and molar refractivity. These parameters have been expanded to the use of structural indexes obtained by quantum chemical treatments (HOMO and LUMO energies, total dipole moments, molecular polarizability and frontier orbital indices). Since this technique is limited to 2D frame of the structure, 3D properties are being calculated in which the stereo-chemical features are explored. Now SAR is represented by 3D visualization of molecular structure and not solely by mathematical models. Various pharmacophore perception algorithms have been developed. Structure-based pharmacophore (direct method) is developed when structure of receptor is known. Structure-based focusing (SBF) technique is an approach based on (i) calculation of interaction sites using the algorithms defined in the LUDI program, (ii) clustering of the vectors for H-bond donating and accepting groups, and hydrophobic regions, and (iii) transformation of the obtained clusters into a feature-based pharmacophore hypothesis. Ligand Scout program is used for generating feature-based pharmacophore from a ligand-target complex structure. The above two approaches can be combined depending on the availability of data for successful model generation.

1.2.1 Ligand-based method: Ligand-based drug design represents an important research field in the cheminformatics drug discovery. If 3D structure of target protein is not available, the chemometric approach applied on the ligands with known biological activity to generate potential model is known as ligand-based drug design. It is well established that binding energy of the ligand is the function of physiochemical properties of the different parts of the ligand, and molecular properties can be visualize as a hybrid of topologic, geometric, and electronic features. The molecular topology can be explained as the number and type of atoms in the molecule and the way they are interconnected to each other, whereas molecular geometry is the disposition of atoms in a 3D space, and it is characterized by three main structural parameters: bond lengths, bond angles and dihedral angles, which define completely the 3D structure of the molecule. On the other hand, electronic molecular structure is characterized through the molecular wave function from which the electronic distribution and electronic properties of the molecule can be known. Accordingly, some topological patterns are responsible for some geometrical features of the molecule, and they determine most of
the electronic parameters. Estrada et al. substantiated the fact through mathematical approaches that 2D molecular descriptors, such as those derived from a graph-theoretical representation of molecules, can be directly used to describe 3D structural parameters of molecules. The flow diagram of ligand-based chemometric technique is shown in Figure 1.5.

![Flow diagram of ligand-based chemometric technique](image)

Figure 1.5: Flow diagram of ligand-based chemometric technique

1.2.1.1 2D QSAR: QSAR models are regarded as a scientifically credible tool for predicting and classifying biological activities of untested chemicals. The fundamental assumption of QSAR is that variations in the biological activity of a series of chemicals that target a common mechanism of action are correlated with variations in their physicochemical properties. Since presumably the structurally related properties of a chemical can be determined by experimental or computational means much more efficiently than its biological activity using in vitro or in vivo approaches. A statistically validated robust QSAR model can predict the biological activity of a new chemical within the same analogs in less amounts of time, money and human resources.

In drug designing approach structural representation and manipulation of chemical structures have lead to the generation of varied methods for representing entire molecular structure. Hansch model, Free Wilson analysis, topological models, quantum mechanical methods and 3D representation of molecular geometry are approached to determine the whole molecular representation.
1.2.1.1 Hansch analysis: Crown Hansch is known as founder of modern QSAR and published a classical article on QSAR study\textsuperscript{118} of plant growth regulators and their dependency on Hammet constant\textsuperscript{119} and hydrophobicity. The partition coefficient of the chemical was measured using n-octanol/water system, and new hydrophobic scale (Eq. 1.1) is introduced. The parameter $\pi$, which is the relative hydrophobicity of a substituent, defined in a manner analogous to the definition of Hammet electronic descriptor ($\sigma$).

\[
\Pi = \log P_x - \log P_H
\]  

(1.1)

$P_x$ and $P_H$ represent the partition coefficient of a derivative and parent molecule respectively. In 1964, Hansch illustrated that biological activity can be correlated linearly by free energy related terms (different physicochemical substituent constant). This approach is originally known as linear free energy relationship (LFER) and expressed as

\[
\log \left( \frac{1}{C} \right) = a\pi + b\sigma + cE_s + \ldots + \text{constant}
\]

(1.2)

where, $C$ is molar dose that produces a certain biological response, $\pi$ is the hydrophobic contribution of the substituent, $\sigma$ is the Hammet electronic descriptor of the substituent\textsuperscript{120} represented by $\log K_x/\log K_H$. $E_s$ is the Taft’s steric parameters\textsuperscript{121} and $a$, $b$, $c$ are coefficients. $K_x$ and $K_H$ are the ionization constant of the meta- or para-substituted and un-substituted acid at 25°C respectively.

Realizing that biological activity of hydrophobic drugs started to decrease after reaching the optimum value, Hansch introduced parabolic model (Eq. 1.3) for analogous of chemical dataset given a second order relationship of hydrophobicity with biological activity \textsuperscript{118}.

\[
\log \left( \frac{1}{C} \right) = a\log P + b(\log P)^2 + c\sigma_1 + C_2
\]

(1.3)

\[
\log \left( \frac{1}{C} \right) = a\log P - b(\log P)^2 + c\sigma + e
\]

(1.4)

Hansch and Fujita included steric, electronic and hydrophobic properties in QSAR equation (Eq. 1.5).

\[
\log \left( \frac{1}{C} \right) = a(\text{parameters}) + b(\text{electronic parameters}) + c(\text{steric parameters})
\]

(1.5)

\[
+ d(\text{other parameter}) + e
\]

Where, $a$, $b$, $c$, $d$ and $e$ are the regression constant determined by least square analysis.
In Hansch model the squared lipophilicity term represents a straightforward way to express nonlinear dependences, but disadvantage is the symmetry of the paraboloid. This limitation is overcome by Kubiny's bilinear model, in which the ascending and descending parts of the functions have different slopes which facilitate more accurate modeling of the observed data. The bilinear model that describes the nonlinear dependence of biological activity of drugs on hydrophobicity is expressed as

$$\log\left(\frac{1}{C}\right) = a\log P - b\log(\beta P + 1) + c \quad (1.6)$$

The terms a and c are linear in nature and can be calculated by multiple regression analysis, whereas b is a nonlinear term and must be calculated by an iterative method. Thus the linear model introduced by Hansch and bilinear model by Kubiny have impact on the mechanistic understanding of chemical structure towards biological activity.

1.2.1.2 Free-Wilson analysis: Free-Wilson approach is a de novo mathematical approach to find out contribution of various substituents and parent ring to the biological activity. The Free-Wilson expression addresses structure-activity studies in a chemical analogs described as

$$BA = \sum a_i x_i + u \quad (1.7)$$

Where, BA is the biological activity, u is the average contribution of parent molecule and $$a_i$$ is the contribution of each structural feature; $$x_i$$ is the indicator, denotes the presence ($$x_i=1$$) or absence ($$x_i=0$$) of a specific structural fragments. The limitation of this approach is that it cannot predict contribution of any other substituent that is present in the original dataset.

Free-Wilson analysis is not as simple in its original formation. No reference compound is selected and so-called symmetry equations are generated to avoid the problem of linear dependence between the variables. This leads to development of Fujita and Ban approach.

1.2.1.3 Fujita-Ban approach: In 1970, Fujita and Ban introduced a modified mathematical model (Eq. 1.8) using logarithmic activity, which is a free energy related term and additive in nature.

$$\log\left(\frac{1}{C}\right) = \sum a_g + \mu \quad (1.8)$$
Where, $a_j$ is the group contribution of a substituent $X_j$ in the position $j$ and $\mu$ is the biological activity value of a reference compound. Fujita-Ban has the same advantages over Free-Wilson approach.

i) Regression analysis matrix can be easily generated.

ii) Addition or elimination of compounds is simple and does not significantly change the values of other regression coefficients.

iii) Any compound may be chosen as the reference compound – singularity problem is avoided.

The values of the group contributions are directly related to Hansch analysis. Introduction of substituent variable in Hansch analysis describes a mixed approach.

1.2.1.1.4 Topological methods: Development of QSAR model using topological approach is one of important method because topological parameters are easily calculated from the graphical representation of the molecules and do not require estimation of any physicochemical property. An atom in a molecule is part of a field of information with regard to electronic influences and topological surrounding\textsuperscript{124, 125}. Influence of this field on any atom can correlate to the biological performance of a molecule. This quantification is based upon three components: (i) Intrinsic state attribute that is associated with each atom, which quantifies the organization, hybrid state, topology and hydride state of the atoms or groups in isolation, (ii) The quantification of the field effect that influences one atom on another within the molecule, and (iii) Distance or separation between any two atoms in a convenient matrix. Most of these indices are based on chemical graph theory. Among the various topological schemes, molecular connectivity indices of Kier and Hall\textsuperscript{124}, and Ghosh and Crippen\textsuperscript{126} have gained maximum popularity, and have been successfully used in quantitative structure property relationship (QSPR) of different physicochemical properties of organic molecules and in the QSAR studies of different bioactive compounds.

1.2.1.1.4.1 Electrotopological method: Kier and Hall proposed electrotopological hypothesis\textsuperscript{124} where combination of electronic and topological attributes develop electrotopological state (E-state) index of an atom. The E-state indices are generated using the chemical graph (hydrogen-suppressed skeleton), which is based on graph theory. The index is also based on the electronic effect of each atom on the other atoms in
the molecule as modified by molecular topology\textsuperscript{124, 125, 127}. Each atom has an assigned intrinsic state value $I_i$ calculated as per Eq. 1.9.

$$I_i = \frac{(2/N)^2 \delta' + 1}{\delta} \quad (1.9)$$

Where, $N$ is the principal quantum number of the atom $i$, $\delta'$ the number or valence electrons in the skeleton ($Z' - h$), and $\delta$ the number of $\sigma$ electrons in the skeleton ($\sigma-h$). For a skeleton atom, $Z'$ is the number of valence electrons, $\sigma$ the number of electrons in $\sigma$ orbitals, and $h$ the number of bounded hydrogen atoms. The E-state, $S(A_i)$ for the atom $A_i$ is the modified intrinsic value (Eq. 1.10).

$$S(A_i) = I_i + \Delta I_i \quad (1.10)$$

Where, $\Delta I_i$ quantifies the perturbing effect on the intrinsic atom value. This perturbation is assumed to be a function of the difference in the intrinsic values $I_i$ and $I_j$ (Eq. 1.11).

$$\Delta I_i = \sum_{j=1}^{N} (I_i - I_j) / r_i^2 \quad (1.11)$$

Where, $r_{ij}$ is the number of atoms in the shortest path between atoms $i$ and $j$ including both $i$ and $j$. The difference in intrinsic values, $\Delta I_i$ for a pair of skeletal atoms encodes both electronic and topological attributes that arise from electronegativity differences and skeletal connectivity. Derived from this electronegativity difference, the E-state value for an atom is related but not limited to the concept of atomic partial charge.

1.2.1.1.4.2 Refractotopological Method: The refractotopological state index (R-state)\textsuperscript{109} is developed using concept of chemical graph theory and the partition of the molar refractivity defined by Ghose and Crippen\textsuperscript{126}. The index is based on the influence of dispersive forces of each atom on other atoms in the molecule, modified by molecular topology. The R-state ($\Re_i$) for atom $i$ is defined as Eq. 1.12.

$$\Re_i = AR_i + \Delta AR_i \quad (1.12)$$

Where, $AR_i$ is the intrinsic refractivity value of the atom $i$, and $\Delta AR_i$ is a perturbation term defined by Eq. 1.13.

$$AR_i = \sum_{j=1}^{N} (AR_i - AR_j) / r_{ij}^2 \quad (1.13)$$

where, the sum is all over $j$ adjacent vertices in the graph, $AR_i$ and $AR_j$ are the intrinsic refractivity values of the atoms $i$ and $j$ respectively, and $r_{ij}^2$ is the number of atoms in the
shortest path between atoms $i$ and $j$ including both $i$ and $j$. As in the E-state, this quadratic topological distance, indicates that there must be a decrease in the interaction effect with the separation distanced between atoms. The intrinsic values of the refractivity for each one of the heavy atoms include the atomic refractivity value of the hydrogen atoms bonded to them.

1.2.1.1.5 Quantum mechanical method: Molecular mechanics is based on the laws of classical physics and deals with electronic interaction by highly simplified approximations, assume that electronic interactions can be adequately accounted for by parameterization. These electromagnetic interactions can be described rigorously by contemporary theories of ‘quantum mechanics’ and ‘quantum electrodynamics’. Quantum mechanics have at least three essential roles to play in drug design: charge approximations, characterization of molecular electrostatic potentials and parameter development for molecular mechanics. Electronic descriptors are derived from molecular wave function. This approach utilizes the Hartree-Fock self consistent field method to solve electronic Schrödinger equation. The Schrödinger equation is solved by two different approaches – ab initio and semiempirical.

In molecular orbital theory, each molecular orbital ($\psi$) is represented as linear combinations of atomic orbital.

$$\psi_I = C_a \psi_a + C_b \psi_b + \ldots + C_n \psi_n$$

(1.14)

$c$ represents contributions of atomic orbital to molecular orbital.

Prediction of preferred conformation of molecule based on molecular orbital theory is now an active area of modeling research. Preferred conformation is a function of interactions of atoms within the molecule. Attainment of minimum energy conformation (MEC), which is resultant of attraction and repulsion among the atoms are the driving forces of conformational changes. MEC is the function of bond angles, bond length and torsional angles, and it can be obtained by varying the parameters and calculating total energy as a sum of orbital energies.

$$E_{tot} = E_{str} + E_{bend} + E_{oop} + E_{tors} + E_{vdw} + [E_{ele} + E_{dist} + E_{ang} + E_{tors} + E_{range} + E_{multi} + E_{field}]$$

(1.15)

where, $E_{str}$ is bond stretching energy, $E_{bend}$ is angle bending energy, $E_{oop}$ out of plane bending energy, $E_{tors}$ is torsional energy, $E_{vdw}$ van der Waals energy, $E_{ele}$ is electrostatic
energy, $E_{dist,c}$ is distance constraint energy, $E_{ang,c}$ is angle constraint energy, $E_{tor,c}$ is torsional constraint energy, $E_{range,c}$ is range constraint energy, $E_{mult}$ is multifit energy and $E_{fieldfit}$ is the energy associated with field fit.

1.2.1.2 3D QSAR: Since the era of classical 2D QSAR proposed by Hansch in 1960s, a variety of QSAR approaches have been reported. Among them first applicable 3D QSAR method was proposed by Cramer et al. in 1988. Molecular field analysis (MFA) was a major breakthrough in the field of 3D QSAR. The main objective of 3D QSAR method is to establish a correlation of biological activities of a series of structurally and biologically characterized compounds with the spatial fingerprints of numerous field properties of each molecule, such as steric demand, lipophilicity and electrostatic interactions. Typically a 3D QSAR analysis allows the identification of the pharmacophoric arrangement of molecular features in space and provides guidelines for the design of next-generation compounds with enhanced bioactivity or selectivity. Since 3D properties of molecule govern biological activity, it is especially informative to analyze a 3D structure to observe how structural changes influence biological activity. Approaches that do provide such a graphical representation are often attractive to the scientific community for easy understanding. The 3D QSAR is more advantageous over 2D QSAR as the former method takes into account the 3D structure of ligands, and is applicable to sets of structurally diverse compounds. 3D QSAR methods are currently used as standard tools in drug design, since they are computationally feasible and afford fast generation of models from which the biological activity of newly synthesized molecules can be predicted.

Approaches of 3D QSAR are mainly include comparative molecular field analysis (CoMFA), comparative molecular similarity indices analysis (CoMSIA), and the GRID/GOLPE program (generating optimal linear PLS estimations). Researchers are devoted to deal with the basic theory, the pitfalls, and the application of 3D-QSAR approaches.

Molecular alignment of 3D structures of the investigated ligands is an important prerequisite for several methodologies in drug design, e.g., 3D field and similarity analyses. The ligands are expected to be aligned against each other to maximize the
overlap of the features to generate molecular fields correctly. An alignment generation procedure usually considers two steps: superimposing the molecules and scoring of the resulting alignments. Superposition techniques may either utilize information obtained from a binding site of a target protein (direct target-based methods) or be based solely on information obtained from the ligands themselves (indirect ligand-based methods). If the crystal structure of target proteins is available then molecular docking is the solution of alignment of molecules for development of 3D QSAR models. It is reported that docking conformers are appropriately aligning the ligands and developing reliable QSAR models.

1.2.1.2.1 CoMFA: CoMFA is the first method that implemented the concept into a 3D QSAR method, i.e., the biological activity of a ligand can be predicted from its 3D structure. It was used as synonym of 3D QSAR for many years. Till now CoMFA is most widely used 3D QSAR method. A CoMFA model normally starts with traditional pharmacophore modeling in order to suggest a bioactive conformation of each molecule by ways to superimpose the molecules under study. The idea behind CoMFA study is that differences in a target property, e.g., biological activity is often closely related to equivalent changes in shapes and strengths of non-covalent interaction fields surrounding the molecules, or the steric and electrostatic fields can provide all information necessary for understanding the biological properties of a set of compounds. The interaction energies between the molecule and defined probe are calculated for each grid points after placing molecule in a cubic grid. In CoMFA method generally two potentials, namely a steric potential in the form of a Lennard-Jones function and an electrostatic potential in the form of a simple Coulomb function, are used for model generation. Only enthalpic contributions of the free energy of binding are provided by the potentials used in this approach.

Cramer et al. in original study, field values were systematically calculated for ligands at each grid point of a regularly sampled 3D grid box that extended 4 Å beyond the dimension of all molecules in the data set, using a sp$^3$ carbon atom with +1 charge as probe. The grid resolution should be in a range to produce the field information that is necessary to describe variations in biological activity. Involvement of too much irrelevant data to statistical analysis may result in a low predictivity of the model. Typically a
resolution of 2 Å is utilized. Often superior results are derived using a grid spacing of 2 Å as opposed to the more accurate 1 Å spacing\textsuperscript{135}. In addition, the CoMFA provides a variety of other parameters (probe atoms, charges, energy scaling, energy cut-offs, etc.) which can be critically controlled.

\textbf{1.2.1.2 CoMSIA:} Klebe \textit{et al.}\textsuperscript{137} first time introduced similarity indices in field analyses after analyzing the problems associated with the functional form of the Lennard-Jones potential used in most CoMFA study. CoMSIA basically based on similarity indices instead of grid-based fields that are obtained using a functional form that is adapted from the SEAL algorithm\textsuperscript{150}. Three different indices related to steric, electrostatic, and hydrophobic potentials were used in their study of the classical steroid benchmark data set. Additionally, hydrogen bond (HB) acceptor (a) and donor (d) are also incorporated. The advantage of this method lies in the functions used to describe the molecules studied, as well as the resulting contour maps. The contour maps obtained from CoMSIA are generally easier to interpret, compared to CoMFA approach. CoMSIA also avoids cut-off values used in CoMFA to restrict potential functions by assuming unacceptably large values.

\textbf{1.2.1.2.3 GRID/GOLPE:} GRID\textsuperscript{151} is used as an alternative of CoMFA method for calculating interaction fields. One of the major advantages of the GRID approach is the use of a 6-4 potential function, which is smoother than the 6-12 form of the Lennard-Jones type, for calculating the interaction energies at the grid lattice points. In a study by Cruciani \textit{et al.}\textsuperscript{152}, use of GRID interaction fields in combination with the GOLPE program\textsuperscript{152} accomplish the necessary chemometrical analysis. The GOLPE approach is developed in order to identify variables those are meaningful for the prediction of the biological activity and to remove those have no impact on predictivity. Within this approach, fractional factorial design (FFD) is applied initially to test multiple combinations of variables. For each combination, a PLS model is generated and only variables which significantly increase the predictivity are considered. Variables are then classified considering their contribution to predictivity. A further advance in GOLPE is the implementation of the smart region definition (SRD) procedure that aims to select the cluster of variables mainly responsible for activity rather than a single
variable. The SRD technique is found to be less prone to change correlation than any single variable selection, and improves the interpretability of the models.

1.2.1.2.4 Hologram QSAR (HQSAR): HQSAR technique avoids many of the problems associated with classical or 3D QSAR approaches\textsuperscript{153}, like involvement of rigorous steps of 3D model generation and mutual alignment in 3D space. In addition, HQSAR's molecular holograms typically consist of only 50 to 200 variables compared to approximately 1,000 variables for molecular fields, thereby enabling considerably faster model development. Only 2D structures and activity are required as input — no selection of descriptors or 3D molecular alignment is needed. HQSAR converts the molecules of a data set into counts of their constituent fragments. These fragment counts are then related to biological data using PLS analysis. Both steps, fragment counting and PLS analysis are very fast. Nevertheless the method is robust and highly predictive for many data sets.

1.2.1.3 Pharmacophore: The term ‘pharmacophore’ was first introduced by Ehrlich\textsuperscript{154} in 1909, and defined as “a molecular framework that carries (phoros) the essential features responsible for a drug’s (pharmacon) biological activity”. Till now the basic pharmacophore concept still remains unchanged, but its intentional meaning and application range have been expanded considerably. In 2000, IUPAC\textsuperscript{155} defined pharmacophore “an ensemble of steric and electronic features that are necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response”. It can be also explain as a set of functional group/fragment types in a spatial arrangement that represents the interaction made in common by a set of small molecular ligands with a protein molecule. The pharmacophore concept is based on the kind of interactions observed in molecular recognition, i.e., hydrogen bonding, charge, and hydrophobic interactions.
A pharmacophore model can be established in two ways: i) A ligand-based manner that superpose a set of active molecules and extracting common chemical features essential for their bioactivity, ii) A structure-based manner by proving possible interaction points between the macromolecular target and ligands. Pharmacophore approaches have been used extensively in virtual screening, de novo design and other applications, such as lead optimization and multi target-drug design. A variety of automated tools for
pharmacophore modeling and applications appeared constantly after the advances in computational chemistry in the past few decades. Many successful stories of pharmacophore approaches in facilitating drug discovery have been reported in recent years\textsuperscript{155,156}. The pharmacophore approach, however, still faces many challenges that limit its capability to reach its expected potential, particularly with the demand for reducing the current high cost associated with the discovery and development of a new drug. In pharmacophore generation, two problems are considered, i) the conformational properties of the rigid molecule are taken into consideration, and ii) determines a pharmacophore which is common to the set of molecules as well as similar orientation of these molecules in space. There are mainly four methods available for pharmacophore generation.

1.2.1.3.1 Constrained systematic search: The constrained systematic is introduced by Dammkoehler \textit{et al.}\textsuperscript{157} to generate the pharmacophore model. This procedure determines the existence of common, 3D orientation of specified functional groups in a series of compounds using critical aspect of the Active Analog Approach\textsuperscript{158}. In medicinal chemistry, the chemical compounds considered are often those which are active at a particular receptor, and the desired geometry is that of the pharmacophore or that of the binding groups of the receptor. The procedure has an exponential dependence on the number of rotatable bonds in the molecule. Due to elimination of high energy conformations, the algorithm is considerably improved by using the tree pruning algorithm\textsuperscript{159}. The method further improves the efficiency by deriving additional constraints on the torsional angle. Initially, the pharmacophoric groups in each molecule that will be overlaid for final pharmacophore are identified and the most rigid molecule is considered for exploring its conformational space. During the conformational search, the distances between all pairs of the selected pharmacophoric groups are recorded. Considering the second most rigid molecule and using the inter-pharmacophore distance ranges derived from the first molecule, constraint on the values permitted to each of its torsional angles are derived. The distance ranges become more restricted as more molecules are considered. In case of flexible compounds, very limited distance ranges are possible on each of its torsional angle which makes the conformational search very efficient.
The basic drawback is that user has to manually specify the important pharmacophoric groups in each molecule that are involved in interaction with the receptor and the correspondences between these groups. It may be difficult to do this when there are many potential pharmacophores giving rise to many possible matches in the molecules.

1.2.1.3.2 Clique detection method: Clique detection\textsuperscript{160} is derived from graph theory, and a ligand can be regarded as a graph in which both the nodes and the edges have labels corresponding to the features (e.g. atom types) and the relations (e.g. inter-atomic distances) respectively. The graph representation enables the graph-theoretical methods to the identification of pharmacophoric patterns. A clique is a subgraph in which every node is connected to other nodes, and the detection algorithm finds the largest clique in a reference graph, which is contained in every other graph in the set. In case of pharmacophore, nodes are the feature elements and connections are the feature-to-feature distances. The graph of a molecule matches the reference if, for any of the molecule's conformers, it has a set of elements of the same class as that of the reference and the inter-element distances match those of the reference within a specified tolerance. This tolerance or acceptable deviation in distance arises primarily from the fact that no fixed set of conformers is likely to contain exactly the correct binding conformation.

1.2.1.3.3 Maximum likelihood method: The method\textsuperscript{161} uses a pre-calculated set of low energy conformations. These conformations are obtained from poling conformational search method\textsuperscript{162}, which is designed to generate a relatively small set of conformations that cover pharmacophore space\textsuperscript{162}. This conformational search method adds an additional penalty term to the energy function during the minimization part of the conformational analysis. This penalty term has the effect of pushing the conformations away from those found previously. The unique feature is its use of location constraints to define the pharmacophore rather than distance ranges between features\textsuperscript{163} that differentiate it from constrained search and clique method. The location of the constraints usually corresponds to a spherical region in 3D space or centered on a point within which the relevant pharmacophoric feature should be positioned. The radius of the spherical region may vary; the different values reflect differences in the dependence of interaction energy on distance for different types of interaction.
1.2.1.3.4 Genetic algorithm: A Genetic Algorithm (GA) is a general randomized optimization technique and useful for solving combinatorial problems with search spaces that are too large for exploration by deterministic search algorithms. This is simulates the process of natural selection by manipulating a population of data-structures called chromosomes. Each chromosome represents a potential solution to a considered problem. Starting from an initial population of chromosomes, each generation undergoes different changes via genetic operators like mutation and crossover, which simulate the evolution. Each time a new generation of the population is created according to the ‘survival of the fittest’ principle, which ensures that over time the population should move toward the optimum solution. This algorithm also can use to find out pharmacophore model from a set of compounds in reasonable amount of time.

1.2.2 Structure-based method: Molecular recognition plays a key role in promoting fundamental biomolecular events, such as enzyme-substrate, drug-protein and drug-nucleic acid interactions. Detailed understanding of the general principles that govern the nature of the interactions (van der Waals, hydrogen bonding, electrostatic) between the ligands and their protein or nucleic acid targets may provide a conceptual framework for designing the desired potency and specificity of potential drug leads. Practical application of this knowledge requires structural data for the target of interest and a procedure for evaluating candidate ligands. Structure-based drug design is one of several methods in the rational drug design toolbox in which chemists are finding the right orientation of ligand and ligand-receptor complex. This method has adopted a growing importance in pharmaceutical research, especially in search for new drugs. There are many receptor dependent (RD) methods, approaches, and programs available to achieve the objective. Binding constant determination by the experimental procedure of a large number of molecules is costly and time-consuming. Molecular docking and scoring methods have been developed for multifarious applications to minimize the problems. RD modeling has an advantage over receptor independent (RI) methods. In QSAR models, an assumption is made that the enthalpy can replace the total free energy of binding. A severe limitation of the techniques, which use only ligand structures, might fail in predicting the activity of a new compound, if many of its 3D properties reside out of the
space occupied by the molecules, or if the predicted compound is structurally different from those used to derive the model. Compared with ligand-based 3D-QSAR approaches, structure-based modeling methods are particularly attractive for their explicit representation of the interactions involved in the protein-ligand binding process. The flow chart of the structure-based chemometric approach in drug design is depicted in Figure 1.6.

The structure-based approach uses the 3D geometrical shape or structure of proteins to assist in the development of new drug compounds. The 3D structure of protein targets is most often derived from x-ray crystallography or nuclear magnetic resonance (NMR) techniques. X-ray and NMR methods can resolve the structure of proteins to a resolution of a few angstroms. For a real understanding of SAR at the receptor level, a direct study of the forces and properties involved in drug receptor interaction is necessary. The term receptor mapping refers to a variety of methods employed to evaluate the structure of a receptor (binding site) by regarding it as complimentary to the drug fitting the receptor. Drug receptor interactions proceed through three steps: recognition of the right features of the compound by the receptor, binding of the drug molecule with receptor and specific perturbation of the three dimensional receptor structures.

![Figure 1.6: Flow chart of receptor-based chemometric drug design approach](image)

Figure 1.6: Flow chart of receptor-based chemometric drug design approach
1.2.2.1 Docking: Computational docking of a small molecule to a biological target involves efficient sampling of possible poses of the ligand in the specified binding pocket of the target in order to identify the optimal binding geometry, as measured by a user-defined fitness or scoring function. The crystalline structure of ligand binds to the target receptor is one of the most important sources to gain information about the basic mechanisms of interaction between the parts constituting the 3D complex structure. The detailed understanding of the general principles that govern the nature of interactions between the ligand and its protein or nucleic acid target may provide a conceptual framework for designing the desired potency and specificity of potential drug leads for a given therapeutic target. Docking procedures aim to identify correct poses of ligands in the binding pocket of a protein and to predict the affinity between the ligand and protein. A number of algorithms can be used for docking which include matching ligand and receptor complementary surfaces or the calculation of the ligand-receptor interaction energies. The method validates ligand-receptor interacting ability by the calculation of scoring functions. Docking can be performed by placing rigid molecules or fragments into the proteins active site using different approaches like clique-search, geometric hashing or pose clustering.

1.2.2.2 De novo design: De novo drug design aims at building a complete molecule from molecular bricks (building blocks) to chemically fill the binding sites of a target molecule. This is an iterative process in which the 3D structure of the receptor is used to design newer molecules. The complete chemical entries could be constructed through linking the “building blocks” together, or growing from an “embryo” molecule with guidance of evaluation of binding affinity. The building block may be atoms or fragments. Using atoms as building blocks is thought to be inefficient whereas fragment-linking approach, the binding site is mapped to identify the possible anchor points for functional groups. These groups are then linked together and form a complete molecule. It is also identifies potential novel ligands by screening a library of small molecules to find those that are complementary to a target receptor. Complementarity is defined as an appropriate spatial orientation of hydrogen-bonding and hydrophobic functional groups. There are two basic types of de novo design approaches.
1.2.2.2.1 **Outside-in approach:** In outside-in method, binding site is first analyzed to determine where specific functional groups are connected together to give molecular skeletons, which are then converted into 'real' molecules.

1.2.2.2.2 **Inside-out approach:** In inside-out, molecules are grown within the binding site under the control of an appropriate search algorithm with each suggestion being evaluated using an energy function.

*De novo* method uses the Ludi algorithm\(^\text{173}\) which works in three steps.

1. Interaction sites within a defined search sphere inside the target receptor are calculated. Typically the search sphere definition is based on the location of a set of known ligands, which bind within the receptor cavity.
2. Libraries are searched for fragments that can fit inside the sphere while forming favorable interactions with the interaction sites.
3. An alignment or linking for the fragments is formulated.

1.3 **Molecular modeling of SERMs:** The identification of potent SERMs is an greatest opportunity to the researchers for beneficial of human being as well as refine the target site-specific effects of novel molecules to explain the selective actions of the modulators in different estrogen target tissues of the same host. It is suggested that the ER-complex could be interpreted as a stimulatory signal at one site and inhibitory signal at another. This idea is consolidated with experimental evidences\(^\text{174}\) and the observations are translated to clinical medicine to develop SERMs to their full potential that could maintain bone density and protect against CHD, but at the same time prevent breast and endometrial cancer. Screening of drug-like compound databases is presently the most popular and useful cheminformatics applications in pharmaceutical discovery. The virtual screening techniques are roughly segregated into structure- and ligand-based applications leading to interesting outcomes for novel drug development.

1.3.1 **Ligand-based studies:** A set of natural ligands and xenoestrogens\(^\text{175}\) are selected for SAR study using Molecular Dynamics (MD) simulations and free energy calculations for ER\(_\alpha\) subtype. The study concluded that diethylstilbestrol (DES) (1.15) like derivatives with slightly larger substituents (in position of ethyl groups of DES) could
make closer fit with receptor. Henkel et al.\textsuperscript{76} worked out on systematic comparisons between characteristics of natural products and synthetic molecules by statistical analysis of structural fragments. The study illustrated that amides and halogens are found more frequently in synthetic compounds, while naturally occurring molecules are explored to be regularly richer in oxygen (e.g., alcohol or ester groups). Using Shannon Entropy (SE) analysis\textsuperscript{177} the molecular descriptors are identified that correctly distinguished, in binary QSAR calculations, between naturally occurring molecules and synthetic compounds.

\begin{center}
\includegraphics[scale=0.5]{molecule.png}
\end{center}

Using PCA a set of 60 diverse environmental estrogens\textsuperscript{178} are classified with to Hierarchical Cluster Analysis (HCA). Good correlation is obtained with hydrophobicity, melting point and molecular energy for cluster of compounds comprising small mono- and bicyclic compounds including butyl phenols, polychlorinated benzenes, DDTs, certain PCB analogs, pesticides and DES-like molecules. Fang et al.\textsuperscript{179} reported SAR based on a total of 230 chemicals (natural and xenoestrogens). Beside molecular modeling simulations using 3D docking approaches, correlation analyses are also performed involving physico-chemical descriptors.

Gallegos Saliner et al.\textsuperscript{180} considered 120 chemicals binding to the ER\textsubscript{a} for SAR study, based on Molecular Similarity Indices (MSI), to predict the binding affinity of estrogenic compounds. The SAR models are constructed by correlating the structural information, described by quantum-chemical indices, physicochemical properties and indicator variables with the biological activity. The observation concluded that presence of a phenolic hydroxyl group resembling the 3-hydroxyl group of the E\textsubscript{2} (1.1) molecule seemed to be essential for the effective binding to ER. In addition, the number of carbon atoms are also found to be significant, indicating the hydrophobic contribution from the ligand.

A set of 232 compounds belonging from estrogen and non-estrogen receptor-binding classes\textsuperscript{181} are considered to determine activity of the ligands using Decision Tree methods. One of the major advantage of this method is speed of model development and prediction\textsuperscript{182, 183}. Chemicals predicted to be active with probability > 0.7 are shown to
have 100% compliance with experimental data, thus demonstrating its use for lead selection.

A set of 313 structurally diverse ligands\textsuperscript{184} are used to develop QSAR models using MultiCASE expert system to screen chemicals with ER binding potential. Sub-structural features associated with ER binding activity (biophores) and features that prevent receptor binding (biophobes) are identified. It is found that the phenolic hydroxyl group is the most prominent sub-structural feature for the ER ligands. Additionally, lipophilicity of the molecule and distance between two hydroxyl groups are also demonstrated to be important for increased activity of the phenolic derivatives.

A set of E\textsubscript{2} derivatives\textsuperscript{185} is considered to develop QSAR models with respect to RBA (relative binding affinity) based on atom-level E-state indices\textsuperscript{124} and the results compared\textsuperscript{186} with Molecular Orbital (MO) derived atom-level parameters for super delocalizability and atomic charges.

The predictive ability of consensus kNN QSAR\textsuperscript{187} method is explored using a diverse set of 245 estrogenic compounds. The key scheme is to predict a molecule’s activity by calculating the weighted mean of the activities of \( k \) most comparable molecules. kNN QSAR method provided good and superior correlation - predictivity pattern as compared to other methods, including CoMFA, HQSAR, CODESSA (Comprehensive Descriptors for Structural and Statistical Analysis), and FRED/SKEYS (Fast Random Elimination of Descriptors/ substructure keys), investigated earlier on same data sets\textsuperscript{87-89}.

Waller \textit{et al.}\textsuperscript{188} considered a set of structurally diverse compounds for comparative study using CoMFA, HQSAR, and FRED/SKEYS paradigms for ER binding affinities. Through this study, it is explained that the 3D nature of the CoMFA technique necessitated that much time be spent generating reasonable structures in an appropriate alignment with information for various molecular fields parsed onto a grid. In comparison, the HQSAR and FRED/SKEYS approaches do not require 3D structures, alignments, or molecular fields. The generation of molecular descriptors is rapid for both of these techniques. The regression component of the FRED/SKEYS technique is the most intense of the trio since numerous models must be developed and evaluated per generation. However, due to the innovations in the variable selection routine implemented in FRED, the number of generations required in this particular evolutionary
algorithm is much fewer than similar programs. The regression requirements of the 
HQSAR technique are similar to those of CoMFA. Therefore, in terms of the time 
required to build and validate QSAR models using the methods described, CoMFA ranks 
at the top being the most intensive followed by FRED/SKEYS and HQSAR. 
Wang et al.\textsuperscript{189} considered 127 ER\textsubscript{a} modulators to develop QSAR models using multiple 
linear regression (MLR) approach. The model describes the importance of connectivity 
valence chain indices. Its negative coefficient may be interpreted as that low value of the 
connectivity valence chain indices can lead to increased binding affinity for a compound. 
Moreover presence of more kinds of elements may lead to an increase of the binding 
affinity for antagonism. Except for C, H atoms, elements O, N and S are also found to be 
crucial for the compounds of the data sets. 
The continued interest in the development of new selective ER\textsubscript{b} ligands, Taha et al.\textsuperscript{190} 
considered 119 ER\textsubscript{b} inhibitors/activators to develop ligand-based 3D pharmacophore(s) 
integrated within self consistent QSAR model(s). The Catalyst used for 
pharmacophore\textsuperscript{191} model generation whereas Cerius\textsuperscript{2}\textsuperscript{192} is used for the QSAR model 
development. Pharmacophore models explain the importance of HB donor, hydrophobic 
and aromatic ring features are crucial for binding affinity towards ER\textsubscript{b}. The QSAR 
models revealed the importance of molecular rigidity, electropotential sum descriptor 
for aromatic oxygen atoms and connectivity indices along with hydrophilicity indicators. 
Tong et al.\textsuperscript{193} developed CoMFA model for 31 estrogenic chemicals with respect to both 
ER\textsubscript{a} and ER\textsubscript{b} subtypes binding affinity. The ER\textsubscript{a} and ER\textsubscript{b} contour plots suggest that 
substitution at the 3-position of the A-ring would affect binding to ER\textsubscript{a} and ER\textsubscript{b} 
equally. Further the RBAs for 2-hydroxy-E\textsubscript{2} and 4-hydroxy-E\textsubscript{2} are proportionately same 
for both ER\textsubscript{a} and ER\textsubscript{b} compared to their respective RBAs for E\textsubscript{2}. Both ER\textsubscript{a} and ER\textsubscript{b} are 
sensitive to adding steric bulk in the vicinity of the 17a-postion on the steroid ring. 
A set of molecules having relative binding affinities values with respect to at ER\textsubscript{a} and 
ER\textsubscript{b} are considered to establish QSAR models\textsuperscript{194} to explore binding affinity through 
QSAR study. From optimized molecular structures a number of theoretical descriptors 
are generated using Molecular Modeling Pro 5.1\textsuperscript{195}, which described 2D and 3D 
structural information as well as molecular physical/chemical properties. These included 
constitutional, steric, electronic, topological and chemical descriptors. The QSAR models
revealed the importance of simple molecular characteristics for differential ER binding by selecting 10 identical descriptors. These molecular descriptors encode molecular characteristics that are responsible for nonselective binding. The remaining descriptors were characteristic for selectivity pattern. Mutual descriptors included molecular size and shape, cyclic structures, solubility parameters, hydrogen bonding/donating potential, electrostatic parameters and the number of ether oxygens. The model confirmed that five distinguishing criteria are essential for nonselective ER activity of phytoestrogens: H-bonding ability of the phenolic ring mimicking the 3-OH, H-bond donor mimicking the 17β-OH, oxygen–oxygen distance between 3- and 17β-OH, precise steric hydrophobic centers at 7α- and 11β-substituents, hydrophobicity and ring structure. Furthermore, predominant molecular characteristics important for subtype selectivity for ERα are 17β substituents and substituted heterocyclic structure. Molecular size, polarity and electronic affect, lipophilic substituents are on the other hand important for ERβ selective binding.

1.3.2 Structure-based studies: Receptor-based approach assesses ligand conformations and find out ligand-receptor interactions that play an important role in modern drug discovery research. Molecular docking and scoring methods are the replacement of experimental determinations of the binding constants of a large number of molecules in reduced time and cost. Significant discoveries are made since few decades on application of virtual docking and superposition-based studies for development of ER-specific ligands and processed for quantifying the ligand-receptor interactions. Triphenyl acrylonitrile derivatives with a p-OH or p-CH₃ group on one or more of the phenyl rings are assessed for the relative influence of each position on binding to the ER. The study concluded that the right-handed helical of triphenyl acrylonitriles structure could logically be superimposed upon the phenolic E₂ ring. The findings on triphenyl acrylonitriles are further studied based on χ²-metrics computations for depicting field of molecules and ER on a single graph. It is observed that systemic substitutions of the triphenyl acrylonitrile skeleton introduced specialized activities, such as phenolic hydroxylation is found to be associated with agonistic activity; introduction of a hydrophobic group is found to emphasize antagonistic activity, whereas bulky and N-containing substituents introduced antagonism and cytotoxic properties.
In a study by Stauffer et al.\textsuperscript{199} investigated the SAR in tetra-substituted pyrazole group of compounds through the conformational search using MMFF94 (Merck Molecular Force Field) and docking studies in Flexidock\textsuperscript{200}. The studies demonstrated that mono-, di- and triphenolic pyrazoles - the skeleton binds to the ER with the C\textsubscript{3} phenol in the E\textsubscript{2} A-ring binding pocket.

Gust et al.\textsuperscript{201} performed docking simulations with a set of heterocyclic derivatives to ER. Based on the theory a phenolic carborane substituted with \(-\text{CH}_2\text{OH}\) at the distal terminus (in the polyborane) is found to be a good-fitting ligand in ER\textsubscript{\alpha} LBD\textsuperscript{202}. Additional evidences for the existence of Asp351 as the extra binding region in the LBD of hER\textsubscript{\alpha}, based on superposition and docking studies with piperazine, imidazoline and imidazole analogs, is observed.

A series of 1,1-diarylethylene moieties\textsuperscript{203} with bridged bicyclic or tricyclic cores are considered for docking study to observe the influence of size, shape and flexibility of ligand in binding pockets of ER\textsubscript{\alpha} and ER\textsubscript{\beta}. The studies revealed the binding conformations for full antagonism to ER\textsubscript{\beta} of most of the species.

A set of 17\textalpha\textsuperscript{\alpha}-20\textalpha\textsuperscript{\alpha}E-21-(4-substituted phenyl)-19-norpregna-1, 3, 5 (10), 20-tetraene-3, 17\beta-diols have been evaluated as high affinity ligands for the ER\textsubscript{\alpha}\textsuperscript{64}. The observed RBA is directly related to the calculated binding energies, and amino acids in the vicinity of p-position of phenyl ring play significant role in binding. Molecular modeling studies have further suggested that there may exist supplementary ligand accessible regions within ER\textsubscript{\alpha}-LBD. It is discovered that additionally Met421, 342, and 348 along with Phe404 and 425 may also be involved in binding ligands to ER\textsubscript{\alpha}-LBD.

A set of 160 ER\textsubscript{\alpha} ligands\textsuperscript{204} are considered for 2-tiered scoring scheme to predict the binding energies of diverse compounds. Ensembles of structures for multiple-conformation docking are generated through Molecular Dynamics (MD) or Monte Carlo (MC) simulation\textsuperscript{205, 206}. This approach led to substantial enrichment of the virtual screening conducted on mixtures of active and inactive ER\textsubscript{\alpha} compounds. A computational model is developed\textsuperscript{207} for predicting binding affinities of ER\textsubscript{\alpha} ligands (containing steroids, synthetic and natural non-steroidal estrogens) using MD simulations in combination with Linear Interaction Energy (LIE) approach\textsuperscript{208}. Good linear correlation
is obtained between experimental and calculated binding energy values for compounds
that bind to the ERα. Advantage of the LIE method lies in the fact that it includes solvent
in the model and takes into account all possible binding orientations. This model
substantiated to be a good predictive platform for xenoestrogens and potential
metabolites of estrogens that are hard to generate or isolate and therefore difficult to test
in vitro.

Yang et al. developed a pharmacophore-based evolutionary approach for virtual
screening using Generic Evolutionary Method for molecular DOCKing (GEMDOCK).
The GEMDOCK integrates discrete and continuous global search strategies with local
search strategies to expedite convergence, whereas the evolutionary approach integrating
an empirical-based energy function and pharmacological preferences, simultaneously
serves as the scoring function for both molecular docking and post docking analyses to
improve screening accuracy. Accuracy of the approach using ER and a ligand database
from the comparative studies are done by Bissantz et al.209. It is revealed that the average
goodness-of-hit (GH) score is high and the average false-positive rate is low for ER
antagonists. The performance of pharmacophore-based scoring function indeed is able to
reduce the number of false positives; moreover the resulting pharmacological interactions
at the binding site as well as ligand preferences are important to the screening accuracy in
the experiments.

Celik et al.210 considered multiple compounds including endogenous estrogens, clinically
utilized SERMs, and putative endocrine-disrupting compounds (EDCs) into the structure
of the ER LBD. The docked results show the significance of several protein
conformations to better understand the mechanism of EDCs interacting with the LBD.
The binding energies calculated in the docking scores are approximate; more
comprehensive calculations can be performed for interesting compounds211,212. The study
further reveals that the new quasi-stable conformations can indeed accommodate most of
the studied ligands, in many cases with better G-scores, especially the two conformations
derived from an apo structure, which give rise to the best docking scores computed for
many of the ligands. The availability of novel conformations of ER LBDs is great
importance to better comprehend the mechanism of interaction of EDCs and may be of
importance for future rational drug design efforts.