The main focus of the work reported in this thesis is on the development of topology based molecular descriptors (MDs) for prediction of biological properties of molecules. Accordingly, numerous new generation topology based MDs have been developed and evaluated. Subsequently the proposed MDs were successfully utilized along with the existing MDs for development of models for diverse biological activities.

Chapter 1 provides an introductory information on drug discovery process, structure activity relationships and significance of topology based MDs for development of models.

Drug discovery is a highly complex, expensive, and time-consuming task, as there is no single systematic way to automatically discover a drug even when the disease, targets and molecular mechanism(s) of drug activity are well understood (Jordheim et al., 2003). FDA estimates that it takes approximately eight-and-a-half years to study and test a new drug before it can be approved for the market. This estimate includes early laboratory and animal testing, as well as later clinical trials using human subjects (www.usfda.gov). The design of lead and drug-like molecules with expected desired properties and feasible chemical synthesis had been the only objectives of the medicinal chemists (Bonnet, 2012). But in today’s scenario medicinal chemists are facing the challenging task of preparing new drug molecules, having specific pharmacodynamics, pharmacokinetic and toxicological properties i.e. the better drugs in shorter times and cost effective way (Colombo and Peretto, 2008). Moreover, the involvement of ever-increasing numbers of compounds and druggable targets emphasise need for efficient screening tools to identify a large number of new drugs. In silico techniques have emerged as a promising alternative or complementary tool toward the effective screening of potential drugs and this approach has been encouraged by the US FDA, as a vital approach for expediting the development of new lead compounds (Helgueraa et al., 2008). These technologies have become well integrated in the modern drug discovery process and have profound impact on many drug design projects in pharmaceutical companies since it generates novel ideas to be exploited by medicinal chemists (Ivanenkov et al., 2009). The in silico techniques are becoming increasingly important for lead discovery and lead optimization. Lead optimization often involves (quantitative) structure-activity/property relationship
[(Q)SAR/QSPR] which focuses on predicting the biological activity/property of a compound from a vectorial representation of molecular structure (Bruce et al., 2007). (Q)SAR/QSPR constitute vital tools used in drug design and have gained extensive recognition in the correlation and prediction of various properties like ADME (Hansch et al., 2004), toxicity (Pasha et al., 2007), retention time (Zhao et al., 2007), stability (Srivastava et al., 2011) and physicochemical properties (Srivastava, 2009) other than the biological activity (Srivani and Sastry, 2009; Srivani et al., 2008; Srivastava et al., 2009; Bohari et al., 2011). A major goal of the (Q)SAR/QSPR studies is to find a mathematical relationship between the property of interest and one or more descriptive parameters (descriptors) derived from the structure of the molecule (Katritzky et al., 2010). However, an inherent problem in the development of a (Q)SAR model is the quantification of chemical structures, since a suitable correlation can only be developed if both the biological activity and chemical structure are quantified. Moreover, the structure-activity relationships need to be established in such a way that the behaviour of a series of molecules can be explained and, at the same time, the process can be inverted in order to obtain new structures that have the activity or property studied (Gupta and Singh, 1999). It has been shown in recent years that much valuable information in structure-property relations can be obtained from the set of connections in the structure (Hall and Mohney, 1991). The graph theoretical invariants (known popularly as topology based molecular descriptors) offer a useful pool of molecular descriptors when the molecular properties are dominated by atom-atom connectivity (Randic, 2001). The basic elements of chemical graph theory are counts of structure features in the molecular skeleton. Graph theoretical indices are developed from counting features such as atoms, bonds, rings, pairs of adjacent bonds, and so forth (Trinajstic, 1983). For the purpose of studying the relationship between chemical structure and property, the binding topology of a molecule is converted into as expression which may be a matrix, a polynomial, a sequence of numbers or numerical index (Rouvray, 1986). The objective is the development of graph invariants which represent the molecular structure and are independent of the manner in which the graph is numbered. Indices from graph theory have become the basis of (Q)SAR expressions whose statistics are often impressive (Sabljic and Trinajstic, 1981). Moreover, the topology based MDs
have the advantage of being true structural invariant, which means that their values are independent of molecular conformations (Kier, 1980). These are based upon molecular topology that considers the arrangements of atoms across the parent molecular skeleton, concepts of steric relations and molecular bulk, branchedness and relationships among various non-bonded parts of the molecule that would be useful to better understand relationships between molecular structure and their experimental properties (Kier and Hall, 1986). The computation of topology based MDs is very swift and can be easily calculated by transforming a hydrogen-suppressed molecular graph of the chemical structure being studied into a number by means of various molecular matrices (Gálvez and García-Doménech, 2010). Consequently MDs based on molecular topology have emerged as molecular descriptors of choice in studies of structure-property activity and rational drug design. They are in particular inescapable in the development of successful (Q)SAR models as well as in screening combinatorial libraries (Randic and Pompe, 2001). Significant progress has been reported during last decade in the development of various topological, geometric, electrostatic, and quantum chemical indices, to be used as molecular descriptors (Basak and Magnuson, 1983). Because of the simplicity of topological structural representation, the topology based MDs are preferred to more complicated geometric, electrostatic, and quantum chemical descriptors, especially in the cases where their use significantly reduces the computation time (Balaban, 1976). Topology based MDs have been classified according to their nature in first, second and third generations (Balaban, 1992). MDs having discriminating power ≥100 for all possible structure containing five vertices have been proposed as fourth generation topology based MDs (Dureja et al., 2008). These MDs found new applications in all research areas of drug development including lead discovery, lead optimization, virtual screening, drug design, combinatorial library design, structure pharmacokinetics and structure-toxicity relationships (Estrada Uriarte, 2001; Yang et al., 2003).

Hence, it is not surprising to see continual development of novel topology based descriptors (Randic and Pompe, 2001). The development of MDs possessing high discriminating power and devoid of any degeneracy still remains a challenge for the scientific community. Consequently, there is a strong need for the development of novel MDs with high sensitivity towards branching, exceptionally high discriminating
power amalgamated with negligible degeneracy (Hollas, 2006). A plethora of MDs have been proposed so far. The calculation of these MDs is well documented in the literature. But, many MDs exhibit considerable mutual correlation. This is a major problem when performing structure-activity studies as the employed statistical methods may fail or give little meaningful results on sets of correlated data (Motoc et al., 1982). In addition, strong correlations among a set of MDs raise doubt whether these indices describe different and meaningful biological, chemical or physical properties of molecules. The purpose of defining a MD is to represent each chemical structure with a numerical value, keeping it at the same time as discriminatory as possible (Hollas, 2006). Researchers are striving hard to develop MDs with not only high discriminating power but also devoid of any degeneracy. The availability of more and more number of diverse theoretical descriptors will be useful to better understand relationships between molecular structure and experimental evidence along with the advancement in computational algorithms.

Literature on various topology based MDs has been reviewed in Chapter 2. Numerous new generation MDs has been conceptualized in Chapter 4. These include Superaugmented eccentric distance sum connectivity indices (denoted by: \(S_{\text{cSED}}\xi_1, S_{\text{cSED}}\xi_2, S_{\text{cSED}}\xi_3\) and \(S_{\text{cSED}}\xi_4\)) and relative distance sum indices and relative distance product indices (denoted by \(S_{1}^{R}, S_{2}^{R}, S_{3}^{R}, S_{4}^{R}, S_{1}^{RP}\) and \(S_{2}^{RP}\)) and relative eccentric distance sum/product indices (denoted by \(R_{\xi_1}^{SV}, R_{\xi_2}^{SV}, R_{\xi_3}^{SV}, R_{\xi_2}^{RP}\) and \(R_{\xi_2}^{SV}\)) along with their topochemical versions have been conceptualized in the present study.

**Superaugmented eccentric distance sum connectivity indices**

Superaugmented eccentric distance sum connectivity indices, \(S_{\text{cSED}}\xi_{n}\), are defined as the inverse of the summation of quotients of the product of adjacent vertex degrees and product of the squared distance sum and eccentricity of the concerned vertex for all vertices in a hydrogen-suppressed molecular graph. These may be expressed as

\[
S_{\text{cSED}}\xi_{n} = \left[ \sum_{i,j=1}^{n} \frac{M_{ij}}{E_{i}^{n} + S_{j}^{n}} \right]^{-1}
\]
Where $M_i$ is the product of degrees of all the vertices ($v_j$), adjacent to vertex $i$ and can be easily obtained by multiplying all the non-zero row elements in augmentative adjacency matrix, $E_i$ is the eccentricity, $S_i$ is the distance sum of vertex $i$ and $n$ is the number of vertices in the graph and the $N$ is equal to 1, 2, 3, 4 for superaugmented eccentric distance sum connectivity indices-1, 2, 3, 4 respectively.

The value of $\text{SED}_{\xi^e_1}$ changes about twelve times (from 232.56 to 19.61), the value of $\text{SED}_{\xi^e_2}$ changes about thirty two times (1492.54 to 46.73) the value of $\text{SED}_{\xi^e_3}$ changes about eighty seven times (9090.91 to 104.17) and the value of $\text{SED}_{\xi^e_4}$ changes by two hundred fifty five times (55555.56 to 217.39) following branching of eleven membered linear carbon structure. The $\xi^e$ value changes by 2.4 times (150 to 62) and the $A\xi^e$ value changes by 3.7 times (5.12 to 19.3) the $\text{SA}_{\xi^e_1}$ value changes by 1.6 times (12 to 7.23) for identical changes. $\text{SED}_{\xi^e_1}$ is about five and three times more sensitive to changes in molecular structure when compared with $\xi^e$ and $A\xi^e$ respectively. $\text{SED}_{\xi^e_2}$ is about sixteen times and ten times more sensitive to changes in molecular structure when compared with $\xi^e$ and $A\xi^e$ respectively and 2.5 times more sensitive than $\text{SED}_{\xi^e_1}$. $\text{SED}_{\xi^e_3}$ is about forty times and twenty times more sensitive respectively in comparison to $\xi^e$ and $A\xi^e$ and about seven times and three times more sensitive in comparison to $\text{SED}_{\xi^e_1}$ and $\text{SED}_{\xi^e_2}$. $\text{SED}_{\xi^e_4}$ and about one hundred six times and sixty nine times more sensitive in comparison to $\xi^e$ and $A\xi^e$ respectively and about twelve times, eight times and three times more sensitive in comparison to $\text{SED}_{\xi^e_1}$, $\text{SED}_{\xi^e_2}$ and $\text{SED}_{\xi^e_3}$ respectively. Low degeneracy and high discriminating power are the most seeking properties for the topological indices. The discriminating power may be defined as the ratio of highest to lowest value for all possible structures of same number of vertices. The discriminating power for all possible structures containing only five vertices for $\text{SED}_{\xi^e_1}$ is 862 for $\text{SED}_{\xi^e_2}$ is 2033 for $\text{SED}_{\xi^e_3}$ is 4561 and for $\text{SED}_{\xi^e_4}$ is 9864 in comparison to 2.3, 295.4 and 755 for $\xi^e$, $A\xi^e$ and $\text{SA}_{\xi^e_1}$. Therefore, the proposed TIs ($\text{SA}_{\xi^e_1}$, $\text{SED}_{\xi^e_2}$, $\text{SED}_{\xi^e_3}$ and $\text{SED}_{\xi^e_4}$) were found to be far more sensitive towards branching using three isomers of undecane.
Summary and Conclusion

$\textit{cSED}_2$, $\textit{cSED}_3$, and $\textit{cSED}_4$ did not exhibit any degeneracy for all possible structures with three, four and five vertices whereas $\textit{cSED}_1$ had a very low degeneracy of one in case of all possible structures with five vertices. $\xi^c$ had 11 identical values out of 21 structures with five vertices, where as no degeneracy was found in case of $\textit{cA}_1$. $\textit{cSA}_1$ had exhibited very low degeneracy of one in case of all possible structures with five vertices. Extremely low degeneracy of the proposed indices ensures the enhanced sensitivity towards the minor changes in branching, connectivity and changes in the molecular structures. The proposed TIs ($\textit{cSED}_2$, $\textit{cSED}_3$ and $\textit{cSED}_4$) are not intercorrelated with some of the widely used MDs: Balaban’s index, Wiener’s index, molecular connectivity index and eccentric connectivity index. Absence of direct correlation with well known indices indicates that proposed $\textit{cSED}_n$ describe the structural properties in a different manner in comparison to other indices. High discriminating power amalgamated with negligible degeneracy offer proposed non-correlating indices a vast potential for use in the characterization of structures, similarity/dissimilarity studies, lead identification, lead optimization, combinatorial library design, and (Q)SAR/QSPR/QSTR for the prediction of various physicochemical, biological, and toxicological properties to facilitate drug design.

**Superaugmented eccentric connectivity topochemical indices**

In order to consider the presence and relative position of heteroatom(s), topochemical versions the above-mentioned four topostructural indices have also been proposed. These adjacency-cum-distance based topochemical indices collectively termed as superaugmented eccentric distance sum connectivity topochemical indices ($\textit{SED}_{\xi^c}$) are i.e. superaugmented eccentric distance sum connectivity topochemical indices-1 ($\textit{SED}_{\xi^c_1}$), superaugmented eccentric distance sum connectivity topochemical indices-2 ($\textit{SED}_{\xi^c_2}$), superaugmented eccentric distance sum connectivity topochemical indices-3 ($\textit{SED}_{\xi^c_3}$) and superaugmented eccentric distance sum connectivity topochemical indices-4 ($\textit{SED}_{\xi^c_4}$).
The superaugmented eccentric distance sum connectivity topochemical indices \( (\text{SED}_{cN}^\xi) \) are defined as the inverse of the summation of quotients of the product of adjacent vertex chemical degrees and product of the squared chemical distance sum and chemical eccentricity of the concerned vertex for all vertices in a hydrogen-suppressed molecular graph. These may be expressed as

\[
\text{SED}_{cN}^\xi = \sum_{i=1}^{n} \frac{M_{ic}}{E_{ic}^N \cdot S_i^2} \]

Where \( M_{ic} \) is the product of chemical degrees of all the vertices \((v_i)\), adjacent to vertex \(i\) and can be easily obtained by multiplying all the non-zero row elements in additive chemical adjacency matrix, \( E_{ic} \) is the chemical eccentricity, \( S_i \) is the chemical distance sum of vertex \(i\) and \(n\) is the number of vertices in the graph and the \(N\) is equal to 1, 2, 3, 4 for superaugmented eccentric distance sum connectivity topochemical indices-1, 2, 3, 4 respectively (denoted by \(\text{SED}_{cN}^\xi_1\), \(\text{SED}_{cN}^\xi_2\), \(\text{SED}_{cN}^\xi_3\) and \(\text{SED}_{cN}^\xi_4\)).

The value of \(\text{SED}_{cN}^\xi_1\) changes by about eleven times (from 238.801 to 20.804), the value of \(\text{SED}_{cN}^\xi_2\) changes by thirty times (1554.158 to 52.646), the value of \(\text{SED}_{cN}^\xi_3\) changes by about seventy seven times (9810.431 to 127.118) and the value of \(\text{SED}_{cN}^\xi_4\) changes by two hundred three times (60235.7 to 296.55) with a minor change in the branching of a eleven membered molecule containing one heteroatom. The discriminating power of \(\text{SED}_{cN}^\xi_1\), \(\text{SED}_{cN}^\xi_2\), \(\text{SED}_{cN}^\xi_3\) and \(\text{SED}_{cN}^\xi_4\) is 302.9, 643.31, 1301.54 and 2627.99 respectively for all possible structures containing only five vertices.

High discriminating power of proposed new descriptors renders them extremely sensitive towards any change in molecular structure. The indices having discriminating power \(\geq 100\) for structures containing only five vertices are treated as ‘fourth generation’ topological descriptors. \(\text{SED}_{cN}^\xi_1\), \(\text{SED}_{cN}^\xi_2\), \(\text{SED}_{cN}^\xi_3\) and \(\text{SED}_{cN}^\xi_4\) did not exhibit any degeneracy for all possible structures with three, four and five vertices.

High discriminatory power, absence of degeneracy and sensitivity towards presence and relative position of heteroatom(s) render proposed Superaugmented eccentric distance sum connectivity topochemical indices - novel molecular descriptors exhibited exceptionally high discriminating power, sensitivity towards both the presence as well as relative position of heteroatom amalgamated with low
Summary and Conclusion

degeneracy. Moreover these indices were found to be non-correlating with important topological descriptors. These qualities ensure their utility in drug design, quantitative structure activity/property relationships, combinatorial library design, isomer discrimination and similarity/dissimilarity studies.


Relative distance sum indices and relative distance product indices

Six detour cum distance matrix based topological indices (TIs) termed as relative distance sum indices and relative distance product indices (denoted by $S_1^R, S_2^R, S_3^R, S_4^R, S_1^{RP}$ and $S_2^{RP}$) as well as their topochemical versions (denoted by $S_1^{Rc}, S_2^{Rc}, S_3^{Rc}, S_4^{Rc}, S_1^{RPc}$ and $S_2^{RPc}$) have been conceptualized so as to take care of cyclic structures. Relative distance sum, denoted by $S_m^R$ may be defined as the summation of ratio of maximum path sum and distance sum of each vertex in a hydrogen suppressed molecular graph having $n$ vertices. It may be expressed as

$$S_m^R = \sum_{i=1}^{n} \left( \frac{\sigma_i}{S_i} \right)^m$$

Relative distance product, denoted by $S_m^{RP}$, may be defined as the product of ratio of maximum path sum and distance sum of each vertex in a hydrogen suppressed molecular graph having $n$ vertices. It may be expressed as

$$S_m^{RP} = \frac{1}{k_m} \left[ \prod_{i=1}^{n} \left( \frac{\sigma_i}{S_i} \right)^x \right]$$

Simple change in the position of ethyl group from ortho to either meta or para leads to steep change in index values of proposed TIs. In case of the $S_1^R$, the value changes from 19.79 to 17.83 as ethyl substituent is shifted from ortho to para position. The index value also changes from 19.79 to 18.69 as ethyl substituent is shifted from ortho to meta position. In case of $S_2^R$ the index value changes from
36.206 to 30.172 as ethyl substituent is shifted from ortho to para position. In case of $S_3^R$, the index value changes from 67.25 to 53.24 as ethyl substituent is shifted from ortho to para position. In case of $S_4^R$, the index value changes from 126.78 to 97.57 as ethyl substituent is shifted from ortho to para position. In case of the $S_1^{RP}$, the value changes by two times from 24.23 to 12.687 as ethyl substituent is shifted from ortho to para position. The index value also changes from 24.23 to 17.419 as ethyl substituent is shifted from ortho to meta position. In case of the $S_2^{RP}$, the value changes by four times from 0.587 to 0.1609 as ethyl substituent is simply shifted from ortho to para position. The index value also changes by two times from 0.587 to 0.303 as ethyl substituent is shifted from ortho to meta position. This major change in the index values without changing number of vertices reveals extremely high sensitivity of proposed indices.

The ratio of the highest to lowest value for all possible structures containing five vertices for $S_1^R$, $S_2^R$, $S_3^R$, $S_4^R$, $S_1^{RP}$ and $S_2^{RP}$ is 2.818, 7.866, 21.767, 59.731, 13.48 and 181.76 respectively which is very high. All the proposed indices ($S_1^R$, $S_2^R$, $S_3^R$, $S_4^R$, $S_1^{RP}$ and $S_2^{RP}$) did not exhibit any degeneracy for all possible cyclic structures with four and five vertices.

High discriminating power and extremely low degeneracy indicates the enhanced capability of these indices to differentiate and demonstrate slight variations in the molecular structure. Intercorrelation analysis of the proposed topological indices with other well-known and widely used TIs revealed that these are not correlated with Wiener’s index, molecular connectivity index, eccentric connectivity index and Balaban’s index. However, proposed indices are weakly correlated with Zagreb indices $M_1$ and $M_2$. The proposed descriptors have been communicated.

**Relative chemical distance sum indices and relative chemical distance product indices**

In order to account for the presence and relative position of heteroatom topochemical versions the above-mentioned five topostructural indices have also been proposed. These detour cum distance matrix based indices, collectively termed as relative chemical distance sum indices ($S_{m}^{Rc}$) and relative chemical distance product indices ($S_{m}^{RPc}$), are: relative chemical distance sum index-1($S_{1}^{Rc}$), relative chemical
distance sum index-2 ($S_{2}^R$), relative chemical distance sum index-3 ($S_{3}^R$),
relative chemical distance sum index-4 ($S_{4}^R$) and relative chemical distance
product index-1 ($S_{1}^{RP}$), relative chemical distance product index-2.

The relative chemical distance sum indices are defined as the product of ratio
of maximum chemical path sum and chemical distance sum of each vertex in a
hydrogen suppressed molecular graph having $n$ vertices. It may be expressed as

$$S_{m}^{R} = \sum_{i=1}^{n} \left( \frac{\sigma_{ic}}{S_{ic}} \right)^m$$

The relative chemical distance product indices are defined as the product of
ratio of maximum chemical path sum and chemical distance sum of each vertex in a
hydrogen suppressed molecular graph having $n$ vertices. It may be expressed as

$$S_{m}^{RP} = \left[ \prod_{i=1}^{n} \left( \frac{\sigma_{ic}}{S_{ic}} \right) \right]^x$$

Discriminating power of topochemical versions of TIs $S_{1}^R$, $S_{2}^R$, $S_{3}^R$, $S_{4}^R$, $S_{1}^{RP}$ and $S_{2}^{RP}$ was found to be 2.629, 6.932, 18.306, 48.406, 11.18,
124.985 respectively. TIs $S_{1}^R$, $S_{2}^R$, $S_{3}^R$, $S_{4}^R$ and $S_{1}^{RP}$ did not exhibit any
degeneracy for all possible cyclic structures with four and five vertices, whereas
$S_{2}^{RP}$ had a low degeneracy of one in the case of all possible cyclic structures with
four vertices having one heteroatom

Similar to topostructural versions, the relative chemical distance sum indices
and relative chemical distance product indices are not intercorrelated with some of
the widely used TIs: $W_c$, $\chi^4$ and $M_5$.

These TIs are exclusively meant for cyclic structures because in case of
noncyclic structures the values for maximum path sum and distance sum will be same
and in turn the index value will be one. Since majority of medicinal compounds
possesss cyclic structures, therefore the proposed indices may be easily utilized in the
lead optimization, quantitative structure-activity/property/toxicity/pharmacokinetic
relationship.

344
High discriminating power amalgamated with low degeneracy offer proposed TIs a vast potential in [Q]SAR/QSPR of cyclic compounds. These qualities ensure their utility in drug design, (quantitative) structure activity/property/toxicity relationships, combinatorial library design, isomer discrimination and similarity/dissimilarity studies.

**Relative eccentric distance sum/product indices**

Five detour/distance matrix based topological indices (TIs) termed as relative eccentric distance sum/product indices (denoted by $R_{\xi_1}^{SV}$, $R_{\xi_2}^{SV}$, $R_{\xi_3}^{SV}$, $RP_{\xi_1}^{SV}$ and $RP_{\xi_2}^{SV}$) have been conceptualized so as to take care of cyclic structures. The *relative eccentric distance sum indices*, ($R_{\xi_m}^{SV}$) are defined as the summation of ratio of the product of maximum path sum and path eccentricity and the product of distance sum and eccentricity of each vertex in a hydrogen suppressed molecular graph having $n$ vertices. It may be expressed as

$$R_{\xi_m}^{SV} = \frac{1}{k} \sum_{i=1}^{n} \left( \frac{\sigma_i}{S_i} \right) E_i$$

*Relative distance product* ($RP_{\xi_m}^{SV}$) are defined as the product of the ratio of product of the maximum path sum and path eccentricity and the product of distance sum and eccentricity of each vertex in a hydrogen suppressed molecular graph having $n$ vertices. It may be expressed as

$$RP_{\xi_m}^{SV} = \frac{1}{k} \prod_{i=1}^{n} \left( \frac{\sigma_i}{S_i} \right) E_i$$

Simple change in the position of ethyl group from ortho to either *meta* or *para* leads to steep change in index values of proposed TI’s. In case of the relative eccentric distance product index-1 ($RP_{\xi_1}^{SV}$), the value changes from 643.12 to 42.31 as ethyl substituent is simply shifted from ortho to *para* position. The index value also changes from 643.12 to 162.88 as ethyl substituent is shifted from ortho to *meta* position. In case of relative t eccentric distance product index-2 ($RP_{\xi_2}^{SV}$) the index value changes from 4136.02 to 17.9 as ethyl substituent is shifted from ortho to *para* position. The index value also changes from 4136.02 to 265.28 as ethyl substituent is
shifted from ortho to meta position. In case of relative eccentric distance sum index-1 ($R_{E_{1}}^{SV}$) the index value changes from 37.295 to 24.43 as ethyl substituent is shifted from ortho to para position. In case of relative eccentric distance sum index-2 ($R_{E_{2}}^{SV}$) the index value changes from 13.70 to 6.78 as ethyl substituent is shifted from ortho to para position. In case of the relative distance product index-3 ($R_{E_{3}}^{SV}$), the value changes by two times from 5.37 to 2.19 as ethyl substituent is shifted from ortho to para position. The index value also changes from 5.37 to 3.24 as ethyl substituent is shifted from ortho to meta position. This major change in the index value without changing number of vertices reveals exceptionally high sensitivity of proposed indices.

The ratio of the highest to lowest value for all possible structures containing five vertices for $R_{E_{1}}^{SV}$, $R_{E_{2}}^{SV}$, $R_{E_{1}}^{SV}$, $R_{E_{2}}^{SV}$, $R_{E_{3}}^{SV}$ is 229, 52429, 9, 74 and 621 respectively which is exceptionally high. All the proposed TIs did not exhibit any degeneracy for all possible cyclic structures with four and five vertices. Intercorrelation analysis of proposed TIs revealed that these are not correlated with Wiener’s index, molecular connectivity index, eccentric connectivity index and Balaban’s index. However, these proposed TIs are weakly correlated with Zagreb indices $M_{1}$ and $M_{2}$.

Relative eccentric distance sum/product topochemical indices

In order to account for the presence and relative position of heteroatom topochemical versions the above-mentioned five topostuctural indices termed as relative eccentric distance sum/product topochemical indices ($R_{E_{1}}^{SV}$, $R_{E_{2}}^{SV}$, $R_{E_{3}}^{SV}$, $R_{E_{1}}^{SV}$, $R_{E_{2}}^{SV}$) have also been proposed.

The relative eccentric distance sum topochemical indices ($R_{E_{m}}^{SV}$) are defined as the summation of ratio of the product of the maximum chemical path sum and chemical path eccentricity and the product of chemical distance sum and chemical eccentricity of each vertex in a hydrogen suppressed molecular graph having $n$ vertices. It may be expressed as
The relative eccentric distance product topochemical indices ($RP_{m} SV$) the product of the ratio of product of the maximum path sum and path eccentricity and the product of distance sum and eccentricity of each vertex in a hydrogen suppressed molecular graph having $n$ vertices. It may be expressed as

$$RP_{m} SV = \frac{1}{k_m} \prod_{i=1}^{n} \left( \frac{\sigma_i \ast \eta_i}{S_i \ast E_i} \right)^x$$

Discriminating power of topochemical versions of proposed $RP_{1} SV$, $RP_{2} SV$, $RP_{1} SV$, $RP_{2} SV$ and $RP_{3} SV$ was found to be 80.54, 64.87, 6.16, 41.87 and 299.3 respectively. All the proposed indices ($RP_{1} SV$, $RP_{2} SV$, $RP_{1} SV$, $RP_{2} SV$, $RP_{3} SV$) did not exhibit any degeneracy for all possible cyclic structures with four and five vertices. The topochemical versions of proposed TIs also did not exhibit any degeneracy for all possible cyclic structures with four and five vertices containing one heteroatom. Intercorrelation analysis reveals that proposed indices are not intercorrelated with some of the widely used TIs.

The proposed TIs are exclusively meant for cyclic structures because in case of noncyclic structures the values for product of maximum path sum and path eccentricity and product of distance sum and eccentricity will be same and in turn the index value will be only one. Exceptionally high discriminating power amalgamated with negligible degeneracy offer proposed TIs a vast potential in QSAR/QSPR. These qualities ensure their utility in drug design, quantitative structure activity/property relationships, combinatorial library design, isomer discrimination and similarity/dissimilarity studies.

The proposed descriptors have been communicated. (The revised manuscript has been submitted to ACS and Combinatorial Science)

Subsequent chapters (Chapter 5-13) represent investigations and applications (case studies) that have been carried out through utilization of proposed as well as existing MDs towards development of models for prediction of diverse biological activities.
Case study-I

Application of proposed MDs for the development of models for the prediction of chk2 inhibitory activities of arylbenzimidazoles

The DNA damage checkpoints are known to comprise signal transduction cascades that link the detection of DNA damage to several other processes i.e. inhibition of progression through the cell cycle from G1 to S, through S and from G2 into M, activation of DNA repair and initiation of apoptosis. Chk2 acts as mediator between DNA damage signalling and also act as barrier for tumorogenesis. Agents that target checkpoint kinases have demonstrated impressive evidence preclinically that this approach will provide tumor specific potentiating agents and may have broad therapeutic utility. There are number of evidences in favour of therapeutic value of Chk2 inhibitors. Chk2 inhibitors are reported to augment the effect of various cytotoxic drugs e.g. Doxorubicin, Cisplatin and Paclitaxel. Past one decade has witnessed the development of checkpoint kinase inhibitors for the treatment of cancer.

The relationship between novel superaugmented eccentric distance sum connectivity indices ($c_{SED}^{C3}$, $c_{SED}^{C4}$, and $c_{SED}^{C1}$) and chk2 inhibitory activities of arylbenimidazoles has been investigated. The values of these topochemical indices were computed for each of the 47 analogues constituting the data set using an in-house computer program. In the present study, DT, RF and MAA based models were developed for the prediction of checkpoint kinase (Chk2) inhibitory activity of 2-arylbenezimidazoles. The decision tree classified the 2-arylbenezimidazoles analogues in the training set with an accuracy of 96% and 10 fold cross-validations with an accuracy of 76.6%. The specificity and sensitivity of DT based model in training set was of the order of 96.5% and 94.4% respectively. The specificity and sensitivity of DT based model in cross-validated set with respect to inactive analogues was of the order of 82.7% and 66.6%. The values of MCC for DT based model in the training set and cross validated set are 0.9 and 0.03. The RF classified 2-arylbenezimidazoles analogues either as active or inactive with an accuracy of 83%. The specificity and sensitivity was of the order of 82.7 and 88.8% respectively and the value of MCC was found to be 0.098. Using a single index at a time, four independent MAA based models using $c_{SED}^{C3}$, $c_{SED}^{C4}$, $M_1^C$ and $W_c$ were developed. Resulting data was analyzed
and suitable models were developed after identification of the active ranges by maximization of moving average with regard to active derivatives. The average IC$_{50}$ (nm) values for each range and activity were also calculated. Subsequently, biological activity was assigned to each analogue using proposed models, which were then compared with the reported chk2 inhibitory activities. Accuracy of prediction was found to vary from a minimum of 90% for model based upon *superaugmented eccentric distance sum connectivity index*-3 to a maximum of ~99% for model based upon *Wiener’s topochemical index* with regard to chk2 inhibitory activity. High predictability amalgamated with high potency in the active ranges offer proposed models a vast potential for providing lead structures for development of potent and selective chk2 inhibitors.


**Case Study-II**

*Models for prediction of receptor tyrosine kinase inhibitory activity of substituted 3-aminoindazole analogues*

Angiogenesis is a process in which new blood vessels are formed from pre-existing vasculatures. It has been reported that the angiogenesis is a rate limiting step in tumor development. Tumors that lack adequate vasculature become necrotic or apoptotic and don’t grow beyond a limited size. Vascular endothelial growth factor (VEGF) is the primary endothelial cell specific angiogenic factor. VEGF activity is mediated by three higher affinity receptors belonging to the class-V subfamily of receptor tyrosine kinases (RTKs). These are widely known for regulating angiogenesis, vasculogenesis, and lymphangiogenesis. The VEGFR family includes VEGFR-I/FLT-1 (fms-like tyrosine kinase-I), VEGFR-2/FLK-I (fetal liver kinase-I)/KDR (kinase inert domain containing receptor) and VEGFR-3/FLT-4 (fms-like tyrosine kinase-4). VEGFR-I is required for endothelial organization during vascular development while VEGFR-2 is required for formation of blood islands and also hematopoiesis. Over activation of KDR by VEGFs has been linked to progression of variety of human cancers. Thus the inhibition of tumor angiogenesis has become a compelling approach in the development of anticancer drugs.
In the present study, an in silico approach using DT and MAA has been applied to a data set comprising of 42 analogues of substituted 3-aminoindazoles for development of models for prediction of KDR, FLT3 and cKIT inhibitory activities. The DT classified the analogues with an accuracy of 88%.

Three independent MAA based models were developed. The overall accuracy of prediction was found to be 88 to 91%. The precision and sensitivity of inactive analogues was found to be 93.93% and 91.11%, whereas the precision and sensitivity of active analogues was of the order of 66.6% and 75% respectively. The MCC value was found to be 0.633 suggesting the randomness in the distribution of data.

The active ranges of the proposed models also exhibited significant FLT3 and cKIT inhibitory activities. A combination of VEGFR and PDGFR inhibitory activities will naturally be more beneficial for the treatment of tumors. High degree of predictability of the proposed models can provide valuable lead structures for the development of potent receptor tyrosine kinase inhibitors (RTKs).


1

Case Study-III

Diverse models for prediction of HIV integrase inhibitory activity of substituted quinolone carboxylic acids

Human immunodeficiency virus (HIV), a retrovirus, is the primary cause of AIDS (acquired immunodeficiency syndrome), and one of the main medical and social problems nowadays. Integration of retroviral DNA into the genome of the host cell is an essential step in the viral replication cycle. Although highly active antiretroviral therapy (HAART) regimens combining three or more RT or protease inhibitors have proven to effectively suppress uncontrolled viral replication, drug resistance and drug-induced side-effects can hinder complete viral suppression. It thus follows that the identification of an inhibitor of HIV IN would add a valuable third component to the antiviral armamentarium.
In the present study both classification and correlation techniques of diverse nature were successfully employed for development of models for the prediction of HIV integrase inhibitory activity using a dataset comprising of 50 analogues of quinolone carboxylic acid. The values of various molecular descriptors (MDs) for each analogue in the dataset were computed using MDS V-life science QSAR plus module. The values of other MDs which are not the part of MDS V-life science were computed using an in-house computer program. A decision tree was constructed for the HIV integrase inhibitory activity to determine the importance of MDs. The decision tree learned the information from the input data with an accuracy of 98% and correctly predicted the cross-validated (10 fold) data with an accuracy of 96%. Three MDs - E-state contribution descriptor (SssOHE), molecular connectivity topochemical index ($\chi^4$) and eccentric connectivity topochemical index ($\xi^e$) were used to develop the models using moving average analysis (MAA). The accuracy of classification of single descriptor based models using MAA was found to vary from a minimum of 96% to a maximum of 98%. The statistical significance of models was assessed through specificity, sensitivity, overall accuracy, Mathew’s correlation coefficient (MCC) and intercorrelation analysis. The widely used methods like multiple linear regression (MLR), partial least square (PLSR) and principal component regression (PCR) were employed for development of correlation models. The models were generated on a training set of 36 molecules. The models had correlation coefficient ($r^2$) of 0.86 to 0.92, significant cross validated correlation coefficient ($q^2$) 0.79 to 0.85, F test from 63.2 to 93.06, $r^2$ for external test set (pred. $r^2$) from 0.69, coefficient of correlation of predicted dataset (pred. $r^2$Se) 0.77 and degree of freedom from 27 to 30. Alignment independent descriptors, SsOHE-index, SaaCHE index, SssCH2 and xlogP were found to be the most important descriptors for development of correlation models for prediction of HIV integrase inhibitory activity.

High accuracy of prediction of these models offers vast potential for providing lead structures for the development of potent but safe therapeutic agents for HIV integrase inhibition. Simultaneous use of diverse technologies for development of numerous models can be of immense use for accelerating drug discovery process.
The said models have been published. (Gupta, M. and Madan, A. K., 2012, Diverse Models for the Prediction of HIV Integrase Inhibitory Activity of Substituted Quinolone Carboxylic Acids, Arch. Pharm. Chem. Life Sci., 345, 989–1000)

Case Study-IV

Development of diverse models for prediction of CDK4 inhibitory activity of substituted 4-aminomethylene isoquinoline-1, 3-diones

Dysregulation of cell-cycle leading to an uncontrolled cellular proliferation is a universal characteristic of cancer. Progression of cells through the cell-cycle is dependent on the formation of specific protein kinase complexes called as cyclin dependent kinases (CDKs) which form in a cyclical fashion. Alterations in the cascade involving CDK4, CDK6, cyclin D, Ink4, pRB and E2F have been observed in more than 80% of human cancers. Therefore, the development of selective CDK4 inhibitors is a promising approach for cancer therapy.

In this study, both classification and correlation approaches have been successfully employed for development of models for the prediction of CDK4 inhibitory activity of aminomethylene isoquinoline-1,3(2H,4H)-diones. A wide variety of MDs have been utilized to develop suitable models through DT, RF, MAA, multiple linear regression (MLR), partial least square regression (PLSR) and principal component regression (PCR). The DT classified the analogues with an accuracy of 96%. The specificity and sensitivity of the training set was found to be of the order of 100% and 86.6% respectively. In cross-validated set, 69% of aminomethylene isoquinoline-1, 3-diones analogues were correctly classified with regard to biological activity. The specificity and sensitivity of cross validated set was found to be 72.9% and 60% respectively. The RF classified aminomethylene isoquinoline-1,3-diones analogues as inactive and active with an accuracy of 96.4% with respect to CDK4 inhibitory activity. In addition to DT and RF analysis, four MAA based models were developed. The overall accuracy of prediction for MAA based models varied from 90.6% to 96.2%. The QSAR model using MLR, PLSR and PCR had a correlation coefficient ($r^2$) of 0.72, significant cross validated correlation coefficient ($q^2$) of 0.63, F test of 15.59, $r^2$ for external test set (pred.$r^2$) 0.59, and degree of freedom 30. The major group of contributing descriptors involved subgroups like rotatable bond count, SsOHE-index, SssOE-index and alignment independent descriptors. A relationship between the descriptor and activity was also studied.
Investigations reveal excellent correlation of MDs with CDK4 inhibitory activities. This excellent correlation led to development of individual models derived through DT, RF, MAA, MLR, PLSR and PCR for prediction of CDK4 inhibitory activity. Proposed models offer vast potential for providing lead structures for development of potent therapeutic agents for CDK4 inhibition.

The models have been published. (Gupta, M. and Madan, A. K., 2013, Diverse models for the prediction of CDK4 inhibitory activity of substituted 4-aminomethylene isoquinoline-1, 3-diones, J. Chem. Sci., 125 (3), 483-493.)

Case Study-V
Models for the prediction of melanocortin-4 receptor agonist activity of 4-substituted piperidin-4-ol
The melanocortins have been involved in a diverse number of physiological functions, including pigmentation, energy homeostasis, steroidogenesis, exocrine secretion, sexual function, analgesia, immunomodulation, inflammation, temperature control, cardiovascular regulation, and neuromuscular regeneration. The involvement of melanocortin-4 receptor (MC4 R) in the sexual functions has drawn considerable attention in recent years as a potential therapeutic target for the treatment erectile dysfunction.

In the present study, DT, RF, MAA and MLR has been applied to a data set comprising of 56 analogues of piperidine-4-ol derivatives for the development of models for prediction of MC4 R agonist activity. A wide variety of 2D and 3D descriptors utilized in the present study were calculated using on-line available E-Dragon software (version 1.0) and an in-house computer program for characterization of each analogue of the dataset.

The DT classified analogues of piperidine-4-ol in the training set with an accuracy of 98.2% and the cross validated set with an accuracy of 96.4% with regard to MC4 R agonist activity. The value of sensitivity, specificity and Matthews correlation coefficient (MCC) for DT based model was found to be 97.4% 100% and 0.9 respectively for training set and 94.7%, 94.4% and 0.8 respectively for cross validation set suggesting robustness of the model. The RF classified 4-substituted piperidine-4-ol analogues as inactive and active with an accuracy of 96.4% with respect to MC4 R agonist activity. The sensitivity, specificity and MCC value of RF based model was found to be 97.3%, 94.4% and 0.9 respectively.
In order to develop MAA based models, relationship between various MDs with MC4 R agonist activity of piperidine-4-ol was also studied and four independent models were developed for prediction of biological activity. The average IC$_{50}$ (nm) values for each range was also calculated. The accuracy of prediction for MAA based models varied from 78.5% to 97.5%. The QSAR model developed using MLR has a correlation coefficient ($r^2$) of 0.83, significant cross validated correlation coefficient ($q^2$) of 0.76, F test of 10.09 and degree of freedom 32.

High predictability of proposed models of diverse nature offers vast potential for providing lead structures for the development of potent therapeutic agents as MC4 R agonist for the treatment of male sexual dysfunction.


**Case Study-VI**

**Diverse models for prediction of dual mTOR and PI3Ka inhibitory activities of substituted 4-morpholinopyrrolopyrimidines**

The mammalian target of rapamycin (mTOR) family of proteins plays vital role in the regulation of the initiation of mRNA transcription and protein translation. The mTOR is the down regulator of PI3K/AKT/mTOR pathway. The mTOR intracellular pathway is activated in the conditions of proliferative deregulation and many types of cancer. Hence, this pathway constitutes vital therapeutic target in case of cancer as it serves as a convergence point for many growth stimuli and through its downstream substrates control cellular processes that contribute towards the initiation and maintenance of cancer.

In the present study, DT, RF, MAA and MLR has been applied for the development of models for prediction of dual mTOR and PI3Kα inhibitory activities of substituted 4-morpholinopyrrolopyrimidines. A wide variety of 2D and 3D descriptors utilized in the present study were calculated using on-line available *E-Dragon software* (version 1.0) and an in-house computer program for characterization of each analogue of the dataset.
The DT classified the 4-morpholinopyrrolopyrimidine analogues in the training set with an accuracy of 97.4% and the cross validated set with an accuracy of 89.7% with regard to mTOR inhibitory activity. The value of sensitivity, specificity and Matthews correlation coefficient (MCC) for DT based model was found to be 100%, 91.7%, and 0.94 respectively for training set and 92.6% and 83.3% and 0.76 respectively for cross validation set suggesting robustness of the model. The DT classified 4-morpholinopyrrolopyrimidine analogues as inactive and active with respect to PIK3α inhibitory activity in the training set with an accuracy of 97.4% and the cross validated set with an accuracy of 89.7% with regard to mTOR inhibitory activity. The value of sensitivity, specificity and MCC value of RF based model was found to be 100%, 91.7%, and 0.94 respectively for training set and 96.3%, 75% and 0.75 respectively for cross validation set. The RF classified 4-morpholinopyrrolopyrimidine analogues with regard to mTOR inhibitory activity with an accuracy of 94.9%. The sensitivity, specificity and MCC value of RF based model was found to be 96.3%, 91.7% and 0.88 respectively. The RF classified 4-morpholinopyrrolopyrimidine analogues with regard to PIK3α inhibitory activity with an accuracy of 92.3%. The specificity, sensitivity and MCC were of the order of 96.3%, 83.3% and 0.82 respectively.

In order to develop MAA based models, relationship between various MDs with dual mTOR and PIK3α inhibitory activity was also studied and five independent models were developed for prediction of biological activity. The accuracy of prediction for MAA based models for mTOR inhibitory activities varied from 87% to >99%. The average IC$_{50}$ (nm) values for each range and activity were also calculated. The QSAR model developed for prediction of dual mTOR and PIK3α inhibitory activity using MLR has a correlation coefficient ($r^2$) of 0.83, significant cross validated correlation coefficient ($q^2$) of 0.61, F test of 11.74, degree of freedom 32.

The high prediction accuracy of proposed models of diverse nature offers immense potential for providing lead molecules for the development of potent medicinal agents for dual mTOR and PI3Kα inhibition.

*The said models have been accepted for publication in ‘Letters in Drug Design and Discovery’*
Summary and Conclusion

Case study-VII
Application of proposed MDs for the development of models for prediction of hQC inhibitory activity of substituted 3-(1H-imidazol-1-yl) propyl thiourea derivatives

According to the World Health Organization about 18 million people worldwide are currently affected by Alzheimer’s disease (AD). Its incidence increases dramatically with age and by 2050, it is projected that the number of patients could be as high as 25 million. It is the 4th leading cause of death, which occurs within ~7-10 years following the onset of symptoms. Aβ deposition is the causative event of AD pathology and the neurofibrillary tangles, Aβ is N-terminally truncated rendering a glutamine that can subsequently be cyclized into pyroglutamate (pE). The enzyme glutaminyl cyclase (QC) catalyzes the conversion of glutamine to pE. In brains of AD patients, the expression of QC is increased in the earliest stages of pathology, which may be an important event in the pathogenesis. The application of inhibitors of QC can be a new strategy for the treatment of AD.

In the present study, DT, RF and MAA have been utilized to develop suitable models for the prediction of human glutaminyl cyclase (hQC) inhibitory activity of substituted 3-(1H-imidazol-1-yl) propyl thiourea derivatives. The DT classified analogues with an accuracy of 95.5%. The specificity and sensitivity of the training were of the order of 96.4% and 94.1%. In 10 fold cross-validation, 84.4% of substituted 3-imidazolyl propyl thiourea analogues were correctly classified with regard to said biological activity. The specificity and sensitivity of crossvalidated set were found to be 85.7% and 82.3% respectively and value of MCC for training and cross validated set was found to be 0.91 and 0.7 respectively. The RF classified substituted 3-imidazolyl propyl thiourea derivatives as inactive and active with an accuracy of 93.3% with respect to hQC inhibitory activity. The out-of-bag (OOB) estimate of error was found to be only 6.7%. The specificity and sensitivity was of the order of 92.8% and 92% respectively and the value of MCC was found to be 0.86.

In order to develop MAA based models, relationship between various MDs with hQC inhibitory activity of substituted 3-(1H-imidazol-1-yl) propyl thiourea derivatives was also studied and five independent models were developed for prediction of biological activity. The accuracy of prediction for MAA based models varied from 86.8% to >95%.

High predictability indicates enhanced capability of proposed models for prediction of hQC inhibitors so as to facilitate development of potent therapeutic agents for the treatment of Alzheimer’s disease.

The said models have been communicated. (The revised manuscript has been submitted to ACS and Combinatorial Science)
Case Study-VIII
Development of models for prediction of c-Met kinase inhibitory activity of substituted quinolines

The success of targeting tyrosine kinase receptors (RTKs) in human cancer has triggered an entirely new therapeutic strategy. Among those investigated RTK targets, the c-Met kinase has drawn significant attention for the ability of its oncogenic forms to confer growth advantage, protection from apoptosis, invasive properties and resistance to chemotherapies. The aberrant c-Met/HGF signaling plays a major role in tumorigenesis invasion and metastasis in many human tumors. The mutation and over-expression of c-Met proto-oncogene and/or HGF have been detected in different types of malignant solid tumors. Hence, c-Met kinase inhibition has emerged as a most promising approach for the development of anti cancer agents.

In the present study, DT, RF, MAA and MLR have been utilized to develop suitable models for the prediction of c-Met kinase inhibitory activity of substituted quinolines. DT was built by recursively segregating the dataset comprising of 77 analogues into predefined classes (A or B) in order to develop informative classification model. The DT classified analogues of substituted quinolines in the training set with an accuracy of 97.4% and the cross validated set with an accuracy of 94.8% with regard to c-Met kinase inhibitory activity. The value of sensitivity, specificity and Matthews correlation coefficient (MCC) for DT based model was found to be 100%, 90% and 0.93 respectively for training set and 96.5%, 90% and 0.87 respectively for cross validation set, suggesting robustness of the model. The RF classified substituted quinolines with regard to c-Met kinase inhibitory activity with an accuracy of 96.1%. The sensitivity, specificity and MCC value of RF based model was found to be 98.2%, 90% and 0.89 respectively.

In order to develop MAA based models, relationship between various MDs with c-Met kinase inhibitory activity of substituted quinolines was also studied and four independent models were developed for prediction of biological activity. The accuracy of prediction for MAA based models varied from 83% to 95%. The specificity and sensitivity for all the MAA based models was found to be >92% and >90% respectively and the value of MCC was found to be >0.74, suggesting robustness of proposed models with regard to c-Met kinase inhibitory activity. The QSAR model developed through MLR has a correlation coefficient ($r^2$) of 0.78, significant cross validated correlation coefficient ($q^2$) of 0.69, F test of 11.74 and degree of freedom 56.

High predictability of the proposed models offers vast potential for providing lead structures for the development of potent therapeutic agents as c-Met kinsae inhibitor for the treatment of various types of cancers. The said models have been communicated.
Case study-IX

Application of proposed MDs for the development of models for prediction of BACE 1 inhibitory activity of substituted amino hydantoins

Alzheimer’s disease (AD) is the most common form of dementia associated with loss of memory (in particular episodic memory), cognitive decline, behavioural and physical disability ultimately leading to death. The disease is a progressive neurodegenerative disorder for which there is no cure. Two key proteases, γ-secretase and β-secretase (BACE1), are considered to have an important role in the abnormal Aβ production. A few BACE 1 inhibitors like E2609 (Eisai Inc.), LY2886721 (Eli Lilly and Company) and MK-8931 (Merck) are undergoing through the phase I and Phase III clinical trials respectively. Still, BACE1 inhibitors have a long road ahead, as the field data on clinical efficacy and possible side effects are awaited.

In the present study an in silico approach has been successfully employed for the development of diverse models for the prediction of BACE1 inhibitory activity using a dataset comprising of 42 analogues of substituted amino hydantoin derivatives. Decision tree (DT), random forest (RF) and moving average analysis (MAA) were utilized for development of the said models. A total of 46 2D and 3D molecular descriptors (MDs) of diverse nature including the newly proposed descriptors were employed for decision tree and random forest analysis. The values of majority of these descriptors for each analogue involved in the dataset were computed using E-Dragon software (version 1.0) as well as an in-house computer programme. The decision tree classified the analogues with an accuracy of 97.6%. The specificity and sensitivity of the training were of the order of 96% and 100%. In 10 fold cross-validation, 85.7% of substituted amino hydantoin analogues were correctly classified with regard to said biological activity. The specificity and sensitivity of crossvalidated set were found to be 88% and 83% respectively and value of MCC for training and cross validated set was found to be 0.95 and 0.7 respectively. The RF classified substituted amino hydantoins analogues as inactive and active with an accuracy of 92.8% with respect to BACE1 inhibitory activity. The out-of-bag (OOB) estimate of error was found to be only 7.2%. The specificity and sensitivity was of the order of 94.1% and 92% respectively and the value of MCC was found to be 0.85.

Using a single descriptor at a time, five independent MAA based models were developed for predicting the BACE1 inhibitory activities. Extremely low values of average IC$_{50}$ indicate high potency of the active ranges in the proposed models. Similar active ranges for BACE2 inhibitory activity were also exhibited by all the proposed models.
The DT identified the proposed relative distance product index-2 as most important. The relative distance product index-2 and relative distance sum index-3 were also utilized to develop MAA based models along with other MDs. Moreover, the active ranges of the proposed models also exhibited significant inhibitory activity against the BACE2. High accuracy of prediction of proposed models offers vast potential for providing lead structures for the development of potent therapeutic agents as BACE1 inhibitors for the treatment of Alzheimer’s disease. The said models have been communicated.

In order to accelerate drug discovery process numerous fourth and fifth generation detour/distances matrix based molecular descriptors possessing high discriminating power, negligible degeneracy and devoid of correlation with existing numerical descriptors have been successfully developed in the present study. The proposed molecular descriptors offer a vast potential for use in (quantitative) structure-activity/property relationships, similarity/dissimilarity studies, isomer discrimination, lead identification, lead optimization, combinatorial library design so as to facilitate drug design. Subsequently, proposed molecular descriptors were successfully utilized for development of models (through both classification and correlation techniques) for the prediction of diverse biological activities such as Chk2 inhibitory, tyrosine kinase inhibitory, mTOR, PI3Ka inhibitory, cMet kinase inhibitory, anti HIV activity, hQC inhibitory activity, BACE1 inhibitory activity and MC4 agonist activity of molecules. Active ranges proposed by the MAA based models can be particularly useful because these can be easily exploited for providing lead structures either through screening or reverse engineering.

Future trend will be towards development of models for simultaneous prediction of efficacy, safety and bioavailability of lead molecules so as to facilitate development of potent, safe and bioavailable therapeutic agents.