Reviews

From molecular graphs to drugs. A review on the use of topological indices in drug design and discovery

Ernesto Estrada1, Grace Patlewicz1 & Eugenio Uriarte2

1Safety and Environmental Assurance Centre (SEAC), Unilever Colworth, Colworth House, Sharnbrook, Bedford MK44 1LQ, UK
2Faculty of Pharmacy, Department of Organic Chemistry, University of Santiago de Compostela, 15706 Santiago de Compostela, Spain

Received 28 January 2003

A review on the use of topological indices (TIs) in the process of drug design and development is presented on the basis of the most recent advances in this field. The first part of this work is focussed on introducing the definitions of terms used in drug design and discovery, graph-theoretical chemistry as well as topological indices. The second part of the review discusses the use of topological indices in the lead discovery and optimization processes, including similarity/dissimilarity studies, rational combinatorial library design, QSAR, the issue of 2D versus 3D QSAR, as well as the use of novel chemometric techniques, such as artificial neural networks, genetic algorithms and partial least squares in combination with TIs. Recent advances on the definition of TIs accounting for three dimensional molecular features (topographic indices) and molecular chirality are also presented.

Introduction

The main paradigm of medicinal chemistry is that biological activity, as well as physical, physicochemical and chemical properties of organic compounds is inherent on molecular structure1,2. Crum-Brown and Fraser published the first quantitative structure-activity relationship3 in 1868 on the basis of this principle. Despite many advances in the field of theoretical drug design, this principle still serves as a useful starting point for guiding the discovery of new lead compounds4-6.

Presently the number of known organic and inorganic compounds exceeds 20 million (http://www.cas.org/cgi-bin/reagetext.pl). A significant proportion of these compounds (>1,000,000) are available from different chemical suppliers. Most pharmaceutical companies have in-house compound collections, which represent a very valuable source for lead discovery and/or optimization. With the explosion of high throughput screening (HTS), large compound libraries can now be screened with ever increasing biological specificity. As a result of these screens, leads are selected and synthesized for further development through pre-clinical and toxicity testing. However, as this development increases so does the cost of lead failures.

The critical step in drug discovery remains the identification and optimization of lead compounds in a rapid and cost effective way. Computational techniques have advanced rapidly over the past decade and accordingly have played a major role in the development of a number of drugs now on the market or going through clinical trials. These computational techniques are based on the use of a “virtual” world of computer-generated hypotheses, which are tested for practicality. In silico studies minimize the need for resource intensive synthesis and bioassays. It has been stated that the “Early use of computational approaches followed by high throughput experimental determinations may become the norm as best practices are adopted in order to improve the chances of successful preclinical and clinical development”7.

Developing structure-activity relationships for drug compounds using computational or theoretical methods, relies on appropriate representations of molecular structure. This review will focus on the role of topological indices (TIs) from a drug design and discovery perspective. The traditional application of TIs in QSAR will not be covered in this review. Readers are encouraged to consult previous compilations, which can provide an introduction to these types of descriptors and examples of their applications. Here we will firstly introduce a number of basic concepts and definitions from a variety of areas, including drug design and discovery, graph-theory and topological indices. These definitions are presented alphabetically. Sec-
ondly we will discuss the role of TIs in drug discovery, highlighting some of the most recent advances in lead discovery strategies, such as drug design, virtual screening, combinatorial library design, similarity-dissimilarity database searching and lead optimization.

Definitions of drug design terms

Artificial neural networks (ANNs): Algorithms simulating the functioning of human neurons which may be used for pattern recognition problems e.g. to establish quantitative structure-activity relationships.8

Bioisostere: A compound resulting from the exchange of an atom or group of atoms with another, broadly similar, atom or group of atoms. The bioisosteric replacement may be physicochemical or topologically based.9

Combinatorial chemistry: The systematic and repetitive, covalent connection of a set of different “building blocks” of varying structures to each other to yield a large array of diverse molecular entities.10

Combinatorial synthesis: The process to prepare large sets of organic compounds by combining sets of building blocks.9

Combinatorial library: The set of compounds prepared by combinatorial synthesis.9

Comparative Molecular Field Analysis (CoMFA): The 3D-QSAR method that uses statistical correlation techniques for the analysis of the quantitative relationship between the biological activity of a set of compounds with a specified alignment, and their three-dimensional electronic and steric properties.8

Computer-aided drug discovery (CADD): This involves all computer-assisted techniques used to discover, design and optimize compounds with desired structure and properties.8,11

Discriminant analysis: This is a statistical technique to find a set of descriptors which can be used to detect and rationalize separation between activity classes.8

Genetic algorithm (GA): This is an optimization algorithm based on the mechanisms of Darwinian evolution which uses random mutation, crossover and selection procedures to breed better models or solutions from an originally random starting population or sample.8,12

High-throughput screening (HTS): The use of robotics to screen thousands to millions of compounds in an automated form.13-15

Lead compound: Any chemical compound that displays the biological activity of interest. It is not a drug but the first approach to search it.

Lead discovery: The process of identifying active new chemical entities, which by subsequent modification may be transformed into a clinically useful drug. It includes the search of compounds in nature (plants, animals, soils, etc.), the investigation of side effects of existing, and so forth.16,17

Lead optimization: The synthetic modification of a biologically active compound, to fulfill all stereoelectronic, physicochemical, pharmacokinetic and toxicologic requirements for clinical efficacy.9

Mass screening: The most traditional method used in drug discovery comprising large scale screening of chemicals in a battery of biological assays.

Molecular descriptors: Terms that characterize a specific aspect of a molecule.8 These are numbers containing structural information derived from structural representations of molecules under study.

Molecular similarity/dissimilarity: This is a number to express structural relatedness between pairs of molecules.8

Multivariate linear regression (MLR): The use of statistical methods for modeling a set of dependent variables in terms of combinations of predictors.8

Orthogonalization of molecular descriptors: This is an approach in which molecular descriptors are transformed so they no longer mutually correlate. Both the non-orthogonal descriptors and the derived orthogonal descriptors contain the same information. This results in the same statistical parameters of the QSAR models.18-21 However, the coefficient of the QSAR model based on orthogonal descriptors is stable to the inclusion of novel descriptors, thus permitting interpretability of the regression terms and evaluation of the role of individual descriptors to the QSAR model.

Partial least squares (PLS): A robust multivariate generalized regression method using projections to summarize multitudes of potentially collinear variables.8,22

Pharmacophore: The ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response.9

Quantitative structure-activity relationships (QSAR): These are mathematical relationships linking chemical structure and pharmacological activity in a quantitative manner for series of compounds.8
Structural database: A large collection of molecular structures in an electronic format.

Topological indices (TIs): Numerical values associated with chemical constitution for correlation of chemical structure with various physical properties, chemical reactivity or biological activity. They are derived from a topological representation of molecules and can be considered as structure-explicit descriptors in contrast with those structure cryptic derived from a topological representation of molecular structures with various physical properties, chemical reactivity or biological molecules and can be considered as structure-explicit electronic constants.

Topological Sub-structural MOlecular DEscriptors (TOSS-MODE): These descriptors are the spectral moments of the bond matrix of a molecular graph weighted to account for hydrophobic, steric and electronic molecular effects and can be expressed as linear combinations of molecular fragments. They also permit the computation of contributions of any part of a molecule to the pharmacological property studied.

Virtual screening: The use of computational techniques to select a reduced number of potentially active compounds from structural databases. The main objective of this approach is to discriminate potent candidate molecules from inactive or less potent molecules.

Definitions of molecular topological terms

Adjacency matrix A: This is a square symmetric matrix of order n whose elements a<sub>ij</sub> are ones or zeros if the corresponding vertices i and j are adjacent or not.

Bond matrix E: This is a square symmetric matrix of order m whose elements e<sub>ij</sub> are ones or zeros if the corresponding bonds i and j are adjacent or not.

Complete graph, K<sub>n</sub>: A graph with any two vertices adjacent. The number of edges in K<sub>n</sub> is n(n-1)/2.

Distance matrix D: This is a square and symmetric matrix of order n whose elements d<sub>ij</sub> correspond to the topological distances between atoms i and j.

Edge or bond degree, δ(e<sub>k</sub>): This is the number of bonds adjacent to k. Two bonds are adjacent if they are incident to the same atom. The following relation is maintained, δ(e<sub>k</sub>) = δ(v<sub>i</sub>) + δ(v<sub>j</sub>) - 2, if v<sub>i</sub> and v<sub>j</sub> are incident to e<sub>k</sub> (ref. 26).

Graph theoretical invariant: Any mathematical expression based on the elements of a graph that does not depend on numbering of graph elements is an invariant. A topological index (TI) is the numerical result of any graph invariant. For instance, adjacency or distance matrices are not graph invariants as they depend on the numbering of graph vertices. However, a simple invariant can be the order of the adjacency matrix, which is the number of vertices in the graph but which does not depend on the graph elements numbering.

Graph: A mathematical object represented as G = (V, E), where V is the set of vertices and E is the set of edges. The number of vertices in the graph is designated by n and the number of edges by m.

Hydrogen-depleted graph: A molecular graph in which hydrogen atoms are not considered.

Molecular graph: A topological representation of a molecule consisting of a collection of points representing the atoms in the molecule and a set of lines representing the covalent bonds. These points are named vertices and the lines are named edges in the graph theory language.

Path: A sequence of vertices v<sub>1</sub>, v<sub>2</sub>, v<sub>3</sub>, ..., v<sub>i</sub> of a graph, such that v<sub>i</sub> and v<sub>j</sub> are adjacent for all i = 1, 2, ..., m. The length of this path is equal to m.

Spectral moments of a matrix: These are the traces, i.e. the sum of the main diagonal, of the different powers of the corresponding matrix.

Subgraph of a graph G: A graph G* = (V*, E*) where V* is a subset of V and E* is a subset of E.

Topological distance between atoms i and j: This is the shortest path between both atoms.

Topological distance: d<sub>ij</sub> is the length of the shortest path between vertices i and j.

Topological representation: A representation of an object providing information only about the number of elements composing it and their connectivities.

Vertex degree or valency of atom i, δ(v<sub>i</sub>): The number of bonds incident in i.

Weighted graph: A graph of the form G = (V, E, ϕ, V), where V is a set containing all the chemical symbols of the elements and ϕ is a subjective mapping of the elements of V onto the set V. It may be considered as a more realistic representation of the topology of a molecule because it identifies the different atoms in the molecule by labeling them with their chemical symbols.

Definitions of topological indices

First generation Tls: These are integers based on integer graph properties, such as topological distances. The most representative indices of this class are...

\(^1\)In a more exact mathematical way the graph invariant can be taken here as the trace of the 0th power of the adjacency matrix. The trace being the sum of the diagonal elements of such matrix...
are Wiener index\textsuperscript{32} \(W\), Hosoya index\textsuperscript{33} \(Z\) and the centric indices of Balaban\textsuperscript{34} \(B\) and \(C\). The only one that has been used in drug discovery research is the Wiener index.

**Wiener index:** This is the half-sum of all entries in the distance matrix\textsuperscript{32}:

\[
W = \frac{1}{2} \sum_i \sum_j d_{ij} \quad \ldots \quad (1)
\]

**Zagreb indices:** These were developed around 1975 on the basis of atom valencies\textsuperscript{35}:

\[
M_1 = \sum_{i=1}^{\text{atoms}} (\delta(v_i))^2 \quad \quad \ldots \quad (2)
\]

\[
M_2 = \delta(v_i) \delta(v_j) \quad \quad \ldots \quad (3)
\]

**Second generation TIs:** These are real numbers based on integer graph properties. Most of the TIs used in drug discovery today are of this class. Some examples are defined below:

**Molecular connectivity indices:** These indices are based on a graph-theoretical invariant introduced by Randić\textsuperscript{25} years ago in order to compute a “branching” index for alkanes\textsuperscript{36-39}:

\[
\chi^\ast = \chi = \sum_{r=1}^{\text{bonds}} \left[ \delta(v_i) \delta(v_j) \right]^{0.5} \quad \ldots \quad (4)
\]

where the sum is carried out over all adjacent atoms in the molecule, i.e., over all bonds.

**Valence molecular connectivity:** This is an extension of the molecular connectivity indices of Randić carried out by Kier and Hall to account for heteroatom differentiation as well as for different subgraphs in the molecule and is expressed\textsuperscript{36-38} as:

\[
\chi^¥ = \chi^¥ = \sum_{r=1}^{\text{bonds}} \left[ \delta^¥(v_i) \delta^¥(v_j) \right]^{0.5} \quad \ldots \quad (5)
\]

where \(\delta^¥(v_i)\) is the valence degree of the atom \(i\), which is equal to the count of valence electrons in the atoms minus the number of hydrogen bonded to it.

The remaining molecular connectivity indices are defined as follows\textsuperscript{36,37}:

\[
\chi_{ij} = \sum_{\text{all edges}} \left[ \delta(v_i) \delta(v_j) \delta(v_k) \right]^{-1/2} \quad \ldots \quad (6)
\]

where \(ij\) means the type of the fragment for which the connectivity index is defined. For instance, connectivity indices of path, clusters, path-clusters and chains have been defined.

**Balaban J index:** A modification of the molecular connectivity index introduced in 1982 by Balaban as an average distance-based connectivity index\textsuperscript{40}:

\[
J = \frac{m}{\mu + 1} \sum_{\text{bonds}} [D(v_i)D(v_j)]^{1/2} \quad \ldots \quad (7)
\]

where \(D(v_i)\) is the sum of all the topological distances related to atom \(i\). That is, the sum of all the entries of the row or column of the matrix \(D\) corresponding to atom \(i\).

**Charge indices:** This was introduced by Gálvez et al\textsuperscript{41} in order to complement the structural information contained in molecular connectivity indices\textsuperscript{41}:

\[
G_i = \sum_{j=1}^{\text{atoms}} \left| C_{ij} \right| \delta(k, d_{ij}) \quad \ldots \quad (8)
\]

where \(CT_{ij} = m_{ij} - m_{ji}\), where \(m\) are the elements of the \(M\) matrix defined as: \(M = A \times D^\ast\). The matrix \(D^\ast\) is the matrix of the inverse square distances, in which their diagonal entries are assigned as 0 and \(\delta\) is Kronecker’s delta.

**Bond connectivity indices:** This is introduced by Estrada and calculated by using the same graph-theoretical invariant of Randić carrying out the sum by bond degrees instead of atom valencies\textsuperscript{42-46}:

\[
e = \sum_{s=1}^{p_2} \left[ \delta(e_i) \delta(e_j) \right]^{-1/2} \quad \ldots \quad (9)
\]

where the sum is carried out over all pairs of adjacent bonds, \(p_2\).

**Kappa indices of molecular shape and flexibility:** This was introduced by Kier\textsuperscript{47-49} in the 1980’s as the basis for devising a relative index of shape given by the relationship of the number of path of length \(l\) in the molecule \(i\), \(P_n\), to some reference values based on molecules with a given number of atoms, \(n\), in which the values of \(P\) are maximum and minimum, \(P_{\text{max}}\), and \(P_{\text{min}}\). The first order shape attribute, \(\kappa_1\), is given by the following expression\textsuperscript{49}:

\[
\kappa_1 = n(n-1)^2 \left( \frac{1}{P} \right) \quad \ldots \quad (10)
\]

The second and third order kappa indices are defined as follows\textsuperscript{49}:

\[
\kappa_2 = (n-1)(n-2) \left( \frac{1}{P^2} \right) \quad \ldots \quad (11)
\]
In order to account for the variation in size contribution to shape from different atoms the radius of atom X relative to the covalent radius of a carbon \( sp^3 \) hybrid atom is considered. The specific correction in computing \( \kappa \) is made by modifying the count of atoms, \( n \), with a modifier, \( \alpha \), calculated as:

\[
\alpha_X = \left( \frac{r_X}{r_{sp^3}} \right)^{-1}
\]

where \( \alpha \) represents a decrement or increment of \( n \) for a non-carbon \( sp^3 \) element X. The modified kappa shape indices are given by:

\[
\begin{align*}
1_{\kappa}\alpha &= \frac{(n + \alpha)(n + \alpha - 1)}{2} \left( \frac{p_i + \alpha}{p_i + \alpha - 1} \right)^2 \quad \text{when } n \text{ is odd} \quad \ldots \quad (12) \\
2_{\kappa}\alpha &= \frac{(n - \alpha)(n - \alpha - 1)}{2} \left( \frac{p_i + \alpha}{p_i + \alpha - 1} \right)^2 \quad \text{when } n \text{ is even} \quad \ldots \quad (13)
\end{align*}
\]

**Electrotopological state (E-state) indices:** Introduced by Kier and Hall in the 1990's, these are based on graph invariants for each atom in the molecule \(^{50-54}\). These indices can be considered as local molecular descriptors. The E-state index for an atom in a molecule represents the electron accessibility of that atom. It is a combination of electron richness or deficiency together with topological accessibility. The E-state index for atom \( I \) in a molecule is defined as:

\[
S_I = I_I + \sum_{j \neq I} \Delta I_{ij}
\]

The sum is over all other atoms \( j \) within the molecular graph. The term for the perturbation of atom \( I \) by atom \( j \) is defined as:

\[
\Delta I_{ij} = (I_i - I_j) \cdot a_{ij}^2
\]

The intrinsic values are defined as:

\[
I_I = \left[ 2/N \right] \delta \gamma + 1/\delta_I
\]

where \( N \) is the principal quantum number for the valence electrons.

**Third generation TIs:** These are real numbers based on real-number local properties of the molecular graph. These indices are of recent introduction \(^{55-57}\), have very low degeneracy and offer the possibility of a wide selection. Other third generation TIs are based on information theory applied to terms of distance sums or on newly introduced non-symmetrical matrices \(^{58-60}\). This paper will not discuss these kinds of TIs as there have been few applications of such descriptors in drug discovery research.

### Topological Indices in Drug Design/Discovery

One of the first examples of successful design of novel compounds using TIs was described by the Upjohn and Pharmacy team in 1993. They were able to design a new class of heteroarylpiperazine compounds with activity versus HIV integrase \(^{61}\) using Basak's method which was based on the application of a series of principal components derived from a pool of 90 TIs \(^{62-66}\). Another successful design, which included the discovery of new active compounds, was reported by Grassy et al. \(^{67}\). These workers designed immunosuppressive peptides by using TIs \(^{67}\), such as kappa shape indices \(^{46-49}\), molecular connectivity \(^{36,37}\), Balaban index \(^{39}\), Wiener index \(^{32}\), E-state sum \(^{50-54}\), as well as other descriptors for the number of atom types including ellipsoid volume, molecular volume, molar refraction, lipophilicity and molecular mass. The peptides under analysis were derived from the heavy chain of HLA class I, which modulate immune responses \( in vitro \) and \( in vivo \). One of them, derived from HLA-B2702, prolonged skin and heart allograft survival in mice. Using this strategy, these workers were able to design and synthesize a peptide that displayed an immunosuppressive activity approximately 100 times higher than the lead compound tested in an heterotrophic mouse heart allograft model.

A systematic approach for drug design and discovery has been applied from the mid 1990's by the Galvez group at the University of Valencia in Spain. They developed a strategy for drug design based on the combined use of molecular connectivity \(^{36,37}\) and charge indices \(^{40}\) together with other in-house molecular descriptors accounting for molecular shape \(^{68}\). The strategy involved selecting a large data set of drugs containing active as well as inactive compounds, e.g. active meaning analgesic drugs and inactive non-analgesic compounds. A vector containing connectivity, charge and shape indices represented each compound. A classification model was found using linear discriminant analysis (LDA) for the compounds in a training set. This was then tested for predictability by using an external prediction set. The
classification model was used to search for 'active' compounds in a large database of structures which were then tested for activity. In this way the model could evaluate compounds not already synthesized.

In applying this approach, Galvez et al. has been able to discover novel activities for known compounds or design and synthesize new ones. The most significant results obtained by this group include the discovery of compounds with analgesic, hypoglycemic, cytostatic, antibacterial, β-blockers, bronchodilator, antifungal, antiparasitic and antihistaminic activities among others.

The most recent applications of this approach have been reported for the prediction of antifungal activity of a diverse set of compounds in addition to discriminating "drug-like" from "non-drug-like" compounds. Indeed the last study concluded that it was possible to achieve a pattern of general pharmacological activity based on molecular topological indices. The prediction of antihistaminic activity using multilinear regression and linear discriminant analysis were also reported by this group. The topological pattern of activity obtained allowed a reliable prediction of antihistaminic activity in drugs frequently used for other therapeutic purposes. Based on the results, the proposed pattern was apparently only valid for drugs which interacted with histamine through competitive inhibition with H1 receptors. The prediction of some pharmacological properties of hypoglycemic drugs and antibacterial discriminant functions has also been reported using the same approach. In the last report physicochemical parameters, such as heat of formation, HOMO, LUMO, dipole moment, polarizability, hyperpolarizability, PM3 generated IR vibrational frequencies, etc., were calculated together with TIs, such as connectivity and topological charge indices.

Another recent strategy using topological descriptors for drug design and QSAR is the so-called TOPS-MODE approach. The TOPS-MODE approach has been successful in the identification of drug candidates in structural databases. Several compounds were identified as sedative/hypnotic by searching the Merck index database using a TOPS-MODE discriminant model. This model was also able to identify more than 80% of Ras farnesyl transferase (FTase) inhibitors in a simulated virtual screening experiment for anti-cancer compounds. In another experiment TOPS-MODE was able to identify more than 88% of anti-convulsant lead compounds in a data set using the discriminant model previously obtained. More recently, this approach has been used for discriminating nucleosides with anti-HIV activity from those being inactive or having low activity. The main advance of this approach is that it enables a calculation of structural contributions to the pharmacological activity studied. Through calculating these fragment contributions, TOPS-MODE has been successfully applied to the identification of possible pharmacophores and bioisosteres. For instance, the role of the piperazinyl fragment in sedative/hypnotic compounds was analyzed with this approach. The role of di-substitution of nitrogen atoms and the possibilities of making isosteric substitutions of this ring with acyclic fragments was determined. A series of active and inactive fragments were also identified in the case of anticancer compounds showing good agreement with active ones previously reported through using a completely different approach. A 2D pharmacophore and a possible isosteric substitution were identified by studying the anticonvulsant activity of a large data set of drugs. The superposition of the 3D structures of such a fragment in a series of structurally different molecules revealed a good degree of mapping between this 2D fragment and a possible 3D pharmacophore. More recently, the TOPS-MODE approach has allowed the elucidation of a series of structural rules for the development of anti-HIV nucleoside compounds. These rules include, the role of different substituents at different positions of the pyrimidyl ring of nucleosides as well as the role of different types of rings in substituting the frequently used furanosyl ring.

Another area of application for topological indices is in similarity/dissimilarity studies. There is recent interest in this area on account of advances in the synthesis and screening of combinatorial libraries as well as lead identification searching from large structural databases. Consequently, topological similarity has been used both in mining structural databases for new analogues based on a lead structure and for improving the efficacy of lead identification screening programs by using HTS. These types of similarity/dissimilarity studies have also been important in combinatorial synthesis programs for designing diverse or biased libraries for synthesis. In fact, molecular diversity is intimately linked to topological similarity/dissimilarity. The principle of similarity/dissimilarity studies is that, given a compound with a desired biological activity, compounds that are
“topologically similar” are likely to have similar/dissimilar activity. Consequently, the main issue is in identifying these lead compounds from a large pool of chemical structures contained in a database. The application of similarity/dissimilarity searching procedures provides a solution\textsuperscript{85-100}.

There are two approaches to similarity studies of chemical structures. According to the direct-comparison method the similarity between two molecules is calculated after superposition of the structures of both molecules\textsuperscript{101}. These methods are computer-intensive and with some exceptions, are not typically applied in database searching. The other approach is based on the use of molecular descriptors that are used to represent the structures\textsuperscript{102}. The similarity of a probe molecule and each database entry is calculated by comparing, (using an appropriate measurement), the list of descriptors in the probe with the list for the entry. There are many ways of encoding the structural information in the form of molecular descriptors. However, the aim of descriptor selection is to reduce the infinite number of potential descriptors to those most appropriate for a given application. Therefore, the use of TIs for similarity/dissimilarity studies have been viewed as the ultimate form of data reduction, since each topological index represents a single number which characterizes the structure as a whole, with the expectation that similar molecular structures will have similar values of such an index.

One of the TIs similarity/dissimilarity approaches was introduced by Basak \textit{et al.}\textsuperscript{62-66}. These workers used a subset of 3,692 structures from the 25,000 contained in the U.S. Environmental Protection Agency’s database TSCA and computed more than 90 TIs. Ten principal components were selected using principal component analysis. These principal components explained more than 92% of the variance and were used as descriptors in a similarity search of the whole database. The results obtained by these authors for the top-five hits for a selection of randomly chosen queries were as good as those expected from the traditional structural-fragment similarity search. The distinction of functional groups, however, was poorer than expected from the second more traditional method.

Another approach to similarity searching in structural databases is based on the use of atom-type electrotopological state indices\textsuperscript{53,54,103}. An atom-type E-state value is defined for each type of atom in a molecule\textsuperscript{104}, e.g., -Cl, -OH, -CH\textsubscript{2}-, etc. Each atom is classified according to its valence state, number of bonded hydrogens and aromaticity. For an atom type E-state index, the E-state values are summed or averaged for all atoms of the same type in the molecule. These E-state values, summed or averaged for an atom type, may be thought of as numerical components of a hyperspace. Thus a 26 atom-type vector was built for each of the 23,310 structures of a database created by Hall and Kier\textsuperscript{54} (21,842 structures from the Pomona MedChem data set + 663 hydrocarbon structures + 515 polyaromatic hydrocarbons + 290 highly branched alkanes). The following general observations were obtained by applying this approach to the 23,310 structures in the database: (i) nearest neighbours tend to have the same or very similar skeletal structure; (ii) nearest neighbours tend to have the same set of heteroatoms; (iii) nearest neighbours tend to have the same number of atoms; and (iv) there is a pattern in the computed similarity distances (very similar molecules up to 0.10 units; similar molecules having one more (less) carbon atom 0.25-0.30 units; isoskeletal molecules with different heteroatoms lie at greater distances.

The application of this approach for searching similar compounds to drug molecules shows great promise as evidenced by analogues of mefloquin, meperidine, niridazole, and minoxidil.

Another very active area of research in applying TIs to drug discovery is related to the area of combinatorial chemistry. The main objective of creating combinatorial libraries is to provide collections of compounds suitable for drug discovery and optimization\textsuperscript{106}. An important question that should be understood in studying combinatorial synthesis is the importance of designing libraries with great molecular diversity. The main objective of selecting a diverse subset of compounds from a much larger population is to identify a subset that best represents the full range of chemical diversity of the larger population and in doing so saving time and money spent in either synthesising or screening “redundant” compounds\textsuperscript{106,106}. Another important aspect of combinatorial library design is that such a general method should work with virtually any potential chemical building block; thus easily calculated molecular descriptors are preferred\textsuperscript{108}.

TIs have been excellent candidates for analyzing molecular diversity for designing combinatorial libraries. Martin \textit{et al.} showed that the discovery efficiency and productivity in combinatorial synthesis
can be further improved by using an experimental design to maximize molecular diversity for a given library\[109\]. They described new methods to quantify molecular diversity using TIs in combination with other descriptors characterizing lipophilicity, chemical functionality, and specific binding features\[109\].

Another interesting approach to rational design of targeted combinatorial library design with TIs was introduced by Tropsa et al. In 1998 and is called Focus-2D. This approach is based on randomly assembling the building blocks that are used in a combinatorial synthesis to produce a virtual library. The individual compounds in such a library are then represented by means of TIs and the molecular similarities between them are evaluated by modified pairwise Euclidean distances in a multidimensional TIs space. Stochastic algorithms, such as, simulated annealing, genetic algorithms and neural network methods, are used to search the potentially large structural space of virtual libraries in order to identify compounds similar to lead molecules. The building blocks that can be used in combinatorial synthesis of chemical libraries are identified by frequency analysis of building blocks of selected virtual compounds. The compounds obtained by combinatorial synthesis will thus have great similarities to the lead molecules. This method was able to identify five building blocks with frequencies higher than random using morphine as a probe. The results show that Focus 2D provided results better than random with 87% confidence\[110-112\]. Ten building blocks were selected by this method when met-enkephalin was used as a probe with better than random predictions at 95% confidence. Twelve building blocks were selected when a composite probe of met-enkephalin and morphine was used. The results indicated that Focus 2D performed better than random with >99% confidence.

In contrast, Linusson et al. have used TIs in a strategy for the statistical molecular design of building blocks for combinatorial chemistry\[113\]. These authors found that it was of critical importance to investigate the building blocks space carefully and to select an appropriate number of blocks to result in an adequate diversity.

In spite of recent progress in the application of TIs to different areas of drug discovery such as similarity/dissimilarity studies or combinatorial library design, their application to lead optimization continues to be the most active area of research. Consequently, many of the applications of TIs in biological research have been dedicated to QSAR studies. This trend is influenced by historical facts. In fact, the number of QSAR studies using TIs showed a dramatic increase after the introduction of the molecular connectivity indices in the 1970's. The explosion of QSAR using TIs is probably due to the fact that novel descriptors do account for a great deal of structural information and because the computation of such descriptors is now readily accessible using a number of software products. Katritzky et al.\[114\] conducted a study which analyzed five sets of molecular descriptors for the development of quantitative structure-property relationships (QSPRs) and QSARs. TIs were found to account for a significant proportion of the variance describing both physicochemical and biological properties. For the development of QSARs, TIs were best combined with additional parameters such as hydrophobicity or quantum chemical ones, possibly due to the influence of transport across biological membranes\[114\]. Several good reviews\[36,37,112-117\] have summarized many of the applications of TIs in QSAR in the 1970's and 1980's. Several QSAR applications of the E-state index have been compiled in an excellent book by Kier and Hall\[114\]. In light of this we will only highlight the most significant advances of TIs in QSAR that have been reported in the recent literature.

There have been two main sources of criticism regarding the application of TIs in QSAR. The first is the non-inclusion of three-dimensional features of molecules into their definition, which avoids the description of some important structural factors that could be directly related to the biological activity. The second is related to the “physical interpretability” of such molecular descriptors. The first of these difficulties will be discussed below, the second one will be addressed later in the manuscript.

A new era of QSAR research was born following the introduction of Comparative Molecular Field Analysis (CoMFA)\[118\]. Reasonably simplistic, high degree of autonomy, good visualization and clear physicochemical interpretation of the descriptors used (steric and electrostatic) made CoMFA one of the most popular methods for QSAR studies\[119\]. The number of applications of this approach to QSAR studies registered an explosion in the 1990's with more than 383 papers published from 1993 to 1996\[120\]. In spite of the great number of successful applications of CoMFA in QSAR, several problems have persisted with this method. In two different papers by Tropsa et al.\[121,122\] it was shown that the results of a conven-
tional CoMFA could often be non-reproducible due to the sometime strong dependence of the CoMFA cross-validated correlation coefficient, $q^2$, with the orientation of rigidly aligned molecules on a user's terminal. These alignment problems are typical for 3D QSAR methods and despite the proposal of possible solutions, the unambiguous 3D alignment of structurally diverse molecules still remains a difficult task.

A novel approach to 2D QSAR was proposed by Tropsha et al. by combining the simplicity of TIs with the powerful statistical tools employed in 3D QSAR, namely PLS. These authors also incorporated GA to optimize the resulting PLS regression equations by removing irrelevant TIs in a fully automated approach. The use of this elegant 2D QSAR approach showed that the combination of a GA-PLS method with TIs gave comparable or better results than CoMFA. This novel approach had the advantage of lacking 3D-alignment problems characteristic of 3D QSAR. Due to its simplicity, efficacy, and high degree of automation this method provides a powerful alternative to 3D QSAR. More recently, this approach was used to model dopamine D$_1$ antagonists and compare outcomes obtained by a CoMFA method. The work showed that this method was highly competitive for QSAR studies, lacked the problems of alignment and eliminated the possibility of inappropriate conformation selection.

Despite this real advance in QSAR, namely, the use of TIs in combination with powerful statistical techniques, non-consideration of 3D structural features remains a problem. An important step forward in the development of graph-based molecular descriptors has resulted in the definition of topographic descriptors. This field of research will not be discussed in detail here. However, some of the key features of these descriptors will be highlighted here. These kinds of molecular descriptors are based on molecular graphs with appropriate weights (weighted graphs) to account for 3D molecular features. Pioneering work in this area was performed by Milan Randić at the end of the 1980's. Randić proposed the use of topographic distance matrices, firstly based on a graph embedded into a hexagonal lattice and then into a 3D diamond lattice. Further work by Bogdanov et al. proposed a 3D distance matrix to compute a 3D Wiener index. The different direction in the search of topographic indices was followed by Estrada who proposed the use of quantum chemical parameters as vertex and edge weights for building topographic adjacency matrices. In 1993 this author proposed the definition of topographic connectivity indices based on bond order weighted molecular graphs. Furthermore, this type of weighted graphs was used to define a topographic edge adjacency matrix (3D bond matrix) as well as a 3D bond connectivity index. In a parallel approach Estrada introduced molecular graphs with vertices weighted by electron charge density on atoms to define a 3D adjacency matrix and a set of topographic connectivity descriptors. In more recent years different authors have defined other types of topographic descriptors.

There have not been many applications of these descriptors to drug discovery and design. Recently topographic (3D) molecular connectivity indices based on molecular graphs weighted with quantum chemical parameters were used in QSPR and QSAR studies. These descriptors were compared to 2D connectivity indices (vertex and edge ones) and to quantum chemical descriptors in modelling partition coefficient (log $P$) and antibacterial activity of 2-furylethylene derivatives. In describing log $P$, the 3D connectivity indices produced a significant improvement (more than 29%) in the predictive capacity of the model compared to those derived with topological and quantum chemical descriptors. The best linear discriminant model for classifying antibacterial activity of these compounds was also obtained with the use of 3D connectivity indices. The global percentage of good classification obtained with 3D and 2D connectivity as well as quantum chemical descriptors was 94.1, 91.2 and 88.2, respectively. In general, all these models predicted the antibacterial activity of a set of 9 new 2-furylethylene derivatives correctly. The best result was obtained with 3D connectivity indices which correctly classified 100% of these compounds versus 88.9% obtained with 2D connectivity or quantum chemical descriptors.

A specific class of these descriptors, namely, the chirality topological indices (CTIs) are briefly considered here. The main purpose of developing these descriptors is to be able to account for chiral molecules, which are well known to have an important role in medicinal chemistry. Until now there have been few of these descriptors reported in the literature but a serious effort in this direction has been recognized by researchers. Among the CTIs published in the literature we can mention those derived by Pyka for describing enantiomers in thin layer chromatography. Some of these indices have been ration-
alized from the mathematical viewpoint of Gutman and Pyka\textsuperscript{140}, in particular the optical topological index which distinguishes between enantiomers and the stereoisomeric topological index which distinguishes between stereoisomers. The relationships between these indices and the Wiener index have been established. Schultz et al.\textsuperscript{141} have also modified a series of TIs in order to introduce information on chirality of stereocentres in molecules.

The first attempt of using CTIs in describing biological activity was carried out by De Julián-Ortís et al.\textsuperscript{142} in developing descriptors that could differentiate between the pharmacological activity of pairs of enantiomers. The IC\textsubscript{50} values of inhibitors of dopamine D\textsubscript{2} receptor and the \(\sigma\) receptor of a group of 3-hydroxyphenyl piperidines were studied with this approach.

The most recent approach to CTIs was that introduced by Golbraik, Bonchev and Tropsha (GBT)\textsuperscript{143}. In the GBT approach several series of CTIs were introduced on the basis of known TIs such as connectivity indices, Zagreb group indices, charge indices, and others. They make use of a term called the chirality correction, which is added to the vertex degrees of asymmetric atoms in a molecular graph. A QSAR study was performed for a set of ecdysteroids with a high proportion of chiral and enantiomeric compounds. The results obtained using the GBT approach compared favorably with those derived from CoMFA. In all these examples it is still important to carry out further studies in order to demonstrate the advantages and disadvantages of using these molecular descriptors in drug discovery; however, the preliminary results are encouraging.

One of the difficulties arising from applying TIs in drug design and discovery is due to the fact that many TIs are interrelated to some extent. Consequently, their inclusion in a QSAR equation does not meet the necessary requirement of variable independence. To solve this problem Randić proposed a procedure of orthogonalization of molecular descriptors that has been applied with much success to QSPR and QSAR studies\textsuperscript{145-161}.

This approach has been used to predict the activity of 19 flavone derivatives to inhibit cAMP phosphodiesterase\textsuperscript{144}. The best QSAR model with orthogonalized descriptors showed the following statistical parameters: \(R = 0.9365\), \(s = 6.99\), and \(F = 35.70\). However, a significant improvement was observed with three ordered orthogonalized descriptors from which the following statistical parameters resulted\textsuperscript{144}: \(R = 0.9841\), \(s = 3.54\), and \(F = 153.38\). Similar improvements have been observed in other QSAR studies\textsuperscript{145-149}. These results indicate that the use of this approach not only overcame the main drawback of using TIs in QSAR, the great intercorrelation between TIs, but it also is able to improve the quality (both statistical and interpretation) of the QSAR models. Advances in orthogonalization strategies of TIs have been reported in recent years by several groups\textsuperscript{149,150}.

In order to improve the quality of TIs used in drug discovery programs, several strategies have been proposed and applied. One of them is the construction of linear combinations of connectivity indices (LCCI), composite connectivity variables, which are found to be very useful in the development of QSAR models of various molecular properties. This procedure of linear combination and semi-empirical TIs introduced by Pogliani is based on a minimal and expanded set of connectivity indices\textsuperscript{150-155}. This strategy has been focused on the study of physicochemical properties of biochemicals, such as amino acids and DNA bases, as well as the study of such properties of other organic compounds. According to Pogliani some of the main features of LCCI are\textsuperscript{150-155}:

1. The combination with the maximal number of molecular connectivity indices does not always show the best quality; (ii) successive inclusion of the next good index to the best combination of indices does not always produce the next best combination of indices; and (iii) even highly intercorrelated indices can contribute to the best combination of indices.

Another approach to improve the quality of TIs has been introduced by Randić et al.\textsuperscript{156-164} and this has been named the “variable molecular descriptors”. This approach has been proposed as an alternative route for the characterization of heteroatoms in molecules. It consists of replacing the zeros along the main diagonal of the adjacency matrix by variables \(x, y, z,\ldots\) for each type of (hetero)atom and then uses the same invariants for calculating the TIs. In using this approach the authors have been able to modify known TIs, such as connectivity ones\textsuperscript{158-161} to include heteroatoms as well as improve the quality of simple TIs in QSPR studies\textsuperscript{162-164}. This approach opens up new possibilities in the development of TIs applied in the drug discovery process, as the number and quality of the indices derived may be increased comparatively to the traditional ways of obtaining such descriptors. A generalization of the “variable molecular descriptors”
approach which enables the calculation of several known TIs using the same graph invariant has been recently introduced by Estrada.\textsuperscript{165} According to this generalized algorithm, Wiener index, Zagreb group indices, Balaban J index, Randić \( \chi \) index and some others are particular cases of an infinite set of TIs.

Another direction for improving the quality of the models in which TIs are used is based on the application of more powerful statistical techniques as in other areas of drug design and discovery. One of these powerful techniques now widely used in QSAR research is the artificial neural network. The theory and general practice of ANN applications is already well documented as well as its application to the solution of different chemical problems.\textsuperscript{166-168} The applications of ANN to TIs QSAR have shown important advances in recent years. In a recent paper Huuskonen \textit{et al.} proposed a method for predicting the aqueous solubility of drug compounds by using TIs and ANN.\textsuperscript{169} In this work a total of 101 connectivity, shape, and atom-type E-state indices were calculated together with three indicator variables for aromaticity, number of hydrogen bond donors and number of hydrogen bond acceptors, respectively. The topological indices used included: \( 2 \chi, 3 \chi_c, 6 \chi_{ch}, 7 \chi_{ch}^{v}, 9 \chi_{ch}^{v}, 10 \chi_{ch}^{v}, 3 \kappa_{a} \) as well as the E-state for 22 atom types. The data set was composed of 211 drugs and related compounds for which the solubility was expressed as log units varying from 0.55 to −5.60. The statistical data for the training set of 160 compounds showed a reasonable degree of accuracy: \( R^2 = 0.90 \) and \( s = 0.46 \). The predictive power of such model was also acceptable having \( R^2_{pred} = 0.86 \) and \( s = 0.53 \) for the 51 compounds in the prediction set.\textsuperscript{169}

The atom-type electrophyschemical state indices were also applied to predict the partition coefficient of a data set of 345 drug compounds or related complex chemical structures by using multilinear regression analysis (MRA) and ANN. Both MRA and ANN provided reliable log \( P \) estimations, but the application of the ANN provided better prediction ability for training and test series using the same series of parameters.\textsuperscript{170}

Back-propagation ANNs were also used to model the structure-activity relationship of a large data set of 103 capsaicin analogues using TIs including molecular connectivity and charge indices, as well as connection table theory, simple atomic decomposition of molecules into atom type and novel physicochemical descriptors.\textsuperscript{171} The ANN QSAR model showed a high level of correlation between experimental and predicted data. After optimization the ANN predicted the EC\textsubscript{50} of 101 capsaicin analogues, classifying correctly 97\% of the 60 active compounds and 83\% of the 41 inactive ones. Other applications of ANNs for the classification of drugs as active/inactive by using TIs have also been reported. The antimicrobial activity of a heterogeneous group of compounds was discriminated from a series of other inactive drugs using connectivity indices and linear discriminant analysis (LDA) as well as ANNs.\textsuperscript{172} Although both methods were shown to be adequate in differentiating between active and inactive compounds, the ANN was more appropriate since it had a success of 98\% in the prediction set versus 92\% obtained by LDA. More recently, Tomas-Vert \textit{et al.} used ANNs to discriminate antibacterial activity in a data set of 249 active molecules, (belonging to different groups of antibiotic and chemotherapeutic drugs) from a set of 415 inactive drugs (belonging to different therapeutic groups but which did not have antibacterial activity).\textsuperscript{173} A set of 62 topological descriptors were used including a series of indicator variables accounting for the number of different atom types in the molecule and a series of sums of distances from a given heteroatom. However, the last set of descriptors (from 51 to 62 in the original paper) that are designed as “summarily of distance x” really correspond to the extended Wiener indices as defined and applied to QSAR studies by Estrada \textit{et al.}.\textsuperscript{174} The results obtained by these authors show a good classification of 93.6\% of actives and 95.9\% of inactives in an external test group of molecules.\textsuperscript{169} An excellent review on the use of molecular graphs descriptors in neural networks models has been reported by Ivanciuc.\textsuperscript{175}

Another recent example of the application of powerful statistical tools to TIs in QSAR is the use of PLS to generate quantitative models. In a QSAR study based on the use of E-state indices for describing the activity of flavone HIV-1 integrase inhibitors Bomlamwini \textit{et al.} found highly predictive models.\textsuperscript{176} The correlation coefficients of \( R^2 = 0.98 \) (3 PLS components) and \( R^2 = 0.99 \) (5 PLS components) and the corresponding cross-validated coefficients of 0.51 and 0.73 demonstrated the possibilities of a PLS approach to TIs QSAR. According to this report, specific areas of the flavone framework are important to the prediction of HIV integrase inhibitory activity. Similar re-
gions of activity were found using CoMFA on the same data set again showing the robustness of results obtained by E-state indices and PLS.

Finally, we wish to make several comments regarding the physical meaning of TIs as a limitation. All TIs have been introduced in an ad hoc way based on graph-theoretical representations of molecules. It is clear that a graph picture of a molecule is an oversimplification of the proper nature of the molecular structure. However, any sort of representation of the molecular structure, or even of any physical reality, is a simplification of such reality. If this representation is able to describe the properties related to it well, it must be because this model represents a "good" description of such reality. In this way this representation of the physical world immediately takes on a physical meaning. There are several good examples of such "oversimplifications" in physical sciences. For instance, the use of the "liquid-drop" model for representing very heavy atoms by Bohr and Wheeler\textsuperscript{177} allows the prediction of disintegration products of uranium considering that the usual radioactive phenomena would be analogous to the evaporation of a molecule from a liquid-drop (!). It is important to say that the value of the theory resides in its predictability. At present, there is no doubt that the use of TIs in drug design and discovery shows this desired feature namely that the models are predictive.

There have been some attempts in relating molecular connectivity indices to quantum chemical concepts as reported by Zefirov\textsuperscript{178} and Gálvez\textsuperscript{179}, as well as on the basis of intermolecular encounters by Kier and Hall\textsuperscript{180,181}. More recently, Estrada\textsuperscript{182} identified molecular connectivity indices as components of molecular accessibility. The first and second order connectivity indices represented molecular accessibility areas and volumes, respectively, while higher order indices represented magnitudes in higher dimensional spaces. In identifying accessibility peripherals the atom degrees were identified as a measure of the accessible perimeter of the corresponding atom. The Randić and connectivity indices are identified as the two components of the molecular accessibility area. The accessibility perimeter is computed here from the van der Waals and covalent radii of the atoms and the interpenetration angle between the van der Waals circumferences of bonded atoms. The description of the accessibility area in terms of the first order Randić and connectivity indices accounts for the success of these descriptors in correlating different physico-chemical and biological properties since they are a measure of the extension of intermolecular interactions. A theoretical justification for the selection of the exponent in the Randić invariant is provided by the relation between the valence degree and the accessibility perimeters calculated in this work. This work opens the doors to interpreting other TIs in terms of physical principles currently used in chemistry.

Conclusions

The actual application of topological indices in drug design and discovery covers a large part of the field including lead discovery and lead optimization. These fields have been mainly covered for this current review and include virtual screening, drug design, combinatorial library design, QSAR, structure-pharmacokinetics, structure-toxicity relationships, and so forth.

We have dedicated some sections of this review to what we considered were important areas of TI development in drug design and discovery. These areas include the use of more sophisticated statistical techniques for developing TI QSARs, such as GA, PLS, ANNs, etc. Such techniques will provide researchers with more appropriate tools for developing robust models using these descriptors. We also considered how the application of topographic and chirality indices in drug design and discovery represent an important step forward. We believe that the study of the structural nature of TIs and their derived models will be of great significance in continuing this area of research active over the next few years.

References

24 This matrix was first introduced as the topological bond matrix by Heilbronner E, Helv chim Acta, 60 (1977) 2248.
38 Randic M, J Am chem Soc, 97 (1975) 6609.