Quasi-orthogonal Basis Sets of Molecular Graph Descriptors as a Chemical Diversity Measure

Ovidiu Ivanciuc,*,† Stavros L. Taraviras,‡ and Daniel Cabrol-Bass*,‡

Department of Organic Chemistry, Faculty of Chemical Technology, University "Politehnica" of Bucharest,
Oficiul 12 CP 243, 78100 Bucharest, Romania, and GRECFO-LARTIC, University of Nice-Sophia Antipolis,
28 Avenue Valrose, 06108 Nice Cedex 2, France

Received July 2, 1999

In the pharmaceutical industry, the virtual screening of combinatorial libraries is used to rationally select compounds for biological testing from databases of hundreds of thousands of compounds. In addition to structural descriptors, such as fingerprints and pharmacophores, the application of relatively simple structural descriptors traditionally used in quantitative structure−activity studies offers speed and efficiency for rapidly measuring the molecular diversity of such collections. We explore new topological indices computed from the molecular graph as potential structural descriptors for the characterization of molecular diversity. A database of 2000 compounds randomly selected from the National Cancer Institute AIDS database was used to measure the intercorrelation of the descriptors. The initial collection of 240 structural descriptors was reduced to several quasi-orthogonal sets of up to 9 descriptors, using different thresholds for the maximum intercorrelation coefficient.

INTRODUCTION

High throughput screening and combinatorial chemistry techniques have evolved from a random selection of chemical compounds to the use of various statistical methods to focus the molecular diversity.[1] Instead of testing millions of randomly chosen compounds, medicinal chemists prefer to use computational chemistry methods that offer for biological testing a smaller number of rationally selected compounds. The virtual (*in silico*) screening of combinatorial libraries is widely used to design optimally diverse libraries and focused or targeted libraries. Virtual screening methods can offer the best set of compounds for a combinatorial synthetic scheme to maximize the chances of finding a drug lead, using both efficient statistical techniques (such as clustering algorithms) and a set of numerical descriptors of the chemical structure. As molecular databases increase in size, information content, and complexity, the application of complicated descriptors becomes impractical for efficient data mining. Therefore, using much simpler and not so demanding descriptors for a first fast diversity screening of a very large dataset could facilitate the entire process. This task can be very well fulfilled by structural descriptors traditionally used in quantitative structure−property relationship (QSPR) and quantitative structure−activity relationship (QSAR) models.[2−10]

More than 1000 structural descriptors from five classes are usually used in QSPR and QSAR studies: constitutional, which relate directly to the chemical constitution of the molecule, such as numbers and types of atoms and bonds, number of rings, molecular weight, etc.; graph theoretic and topological indices, which describe the atomic connectivity within the molecule; geometrical; electrostatic; and quantum descriptors. From these descriptors, the topological indices derived from the molecular graph, together with other two-dimensional descriptors, were found to give good results in database screening.

Topological indices (TIs) have several obvious advantages when compared with geometrical, electrostatic, and quantum descriptors: they are computed only from the information contained in the molecular graph, they have a unique value for a particular chemical compound, and their calculation requires small computational resources. All the above characteristics indicate that TIs are well suited for the virtual screening of databases of hundreds of thousands of compounds.

One must note that all geometrical and quantum descriptors are derived from the three-dimensional structure of a chemical compound whose geometry is generated by means of builders using expert rules (e.g., Corina or Concord) or by distance geometry methods (e.g., Rubicon) and, eventually, optimized with a molecular mechanics, semiempirical, or ab initio quantum method. Moreover, the determination of energetic minima for the molecular geometry is usually performed by conformational analysis. For flexible compounds containing more than five torsion angles, the conformational analysis is difficult and requires large computational resources. Molecular dynamics techniques offer detailed information on the population of conformers for organic compounds, but they require expensive computer simulations, do not offer simple structural descriptors that characterize the entire population, and cannot be applied for large sets of compounds, as required in virtual screening. To make things even more difficult, not all computational methods give reliable results in the search of the conformational space, and the geometry of the global minimum depends on the particular molecular mechanics or quantum method employed.

* Corresponding authors. E-mail: o_ivanciuc@chim.upb.ro and cabrol@unice.fr., respectively.
† University "Politehnica" of Bucharest.
‡ University of Nice-Sophia Antipolis.

10.1021/ci990064x CCC: $19.00
Published on Web 12/17/1999

Together with fingerprints and pharmacophores, TIs are well-suited descriptors for a fast and efficient virtual screening of large databases. Once the number of compounds is reduced to a reasonable number, geometrical and quantum descriptors can be used for the lead selection. Considering the advantages of graph descriptors, they represent valuable descriptors that complement (and not substitute) the structural information encoded in other classes of descriptors. Because TIs are global descriptors of the molecular graph,[11−19] they do not contain explicit information regarding the number of functional groups, pharmacophores, volume, surface area, interatomic distances, charge distribution, orbital energy, or electrostatic potential; such information must be provided by other structural descriptors.

From the large number of TIs used in QSPR and QSAR models,[11−36] only a small fraction were tested in the virtual screening of combinatorial libraries. Two main causes can be identified to explain this situation: they are not readily computed by the commercial software, and many TIs were defined only for hydrocarbons and there are no suitable mathematical equations and heteroatom parameters to compute them for any organic compound. Recently, Ivanciuc introduced several parameter sets for the atoms and bond in the molecular graph[24] and defined graph operators as a generalization of TIs;[30] using any molecular graph matrix[18] and the new parameters, graph operators can provide structural descriptors for any organic compound. In this paper we explore new topological indices computed from the molecular graph as potential structural descriptors for the characterization of molecular diversity.

MOLECULAR GRAPH DESCRIPTORS

Structural descriptors derived from the molecular graph are widely used in modeling physical, chemical, or biological properties, in similarity and diversity assessment, database mining, and in the virtual screening of combinatorial libraries. A topological index is a numerical descriptor of the molecular structure derived from the corresponding molecular graph. The topological description of a molecule contains information on the atom−atom connectivity in the molecule, and encodes the size, shape, and branching features that determine the molecular properties. This graph description of molecules neglects information on bond lengths, bond angles, and torsion angles. Numerous reviews[11−19] were published on the theory and applications of topological indices in QSPR and QSAR models. Because the majority of the structural descriptors used in this paper to describe the chemical diversity were recently introduced, in this section we will briefly present their equations.

Weighting Schemes. In the chemical graph theory, an organic compound containing heteroatoms and/or multiple bonds can be represented as a vertex- and edge-weighted molecular graph.[16] A vertex- and edge-weighted (VEW) molecular graph $G = G(V,E,Sy,Bo,Vw,Ew,w)$ consists of a vertex set $V = V(G)$, an edge set $E = E(G)$, a set of chemical symbols of the vertices $Sy = Sy(G)$, a set of topological bond orders of the edges $Bo = Bo(G)$, a vertex weight set $Vw(w) = Vw(w,G)$, and an edge weight set $Ew(w) = Ew(G)$. The elements of the vertex and edge weight sets are computed with the weighting scheme w. Usually, hydrogen atoms are not considered in the molecular graph, and in a VEW graph the weight of a vertex corresponding to a carbon atom is 0, while the weight of an edge corresponding to a carbon−carbon single bond is 1. Also, the topological bond order Bo_{ij} of an edge e_{ij} takes the value 1 for single bonds, 2 for double bonds, 3 for triple bonds, and 1.5 for aromatic bonds. Aromatic systems were identified from the graph-theoretical description of the molecules by an expert system written in Prolog, which recognizes mono- and polycyclic systems, with five-, six-, and seven-membered rings which obey the generalized Hückel's rule.[37] Several procedures for computing vertex and edge weights were proposed in the literature.[16,20−24]

In a weighting scheme w the vertex Vw and edge Ew parameters are computed from a property p_i associated with every vertex v_i from G, $v_i \in V(G)$, and the topological bond order Bo of all edges from the molecular graph. The vertex parameter $Vw(w)_i$ for the vertex v_i is

$$Vw(w)_i = 1 - p_C/p_i \tag{1}$$

and the edge parameter $Ew(w)_{ij}$ for the edge between vertices v_i and v_j is

$$Ew(w)_{ij} = p_C p_C/Bo_{ij} p_i p_j \tag{2}$$

where p_i is the atomic property of vertex v_i, p_j is the atomic property of vertex v_j, and p_C is the atomic property for carbon atom. Several weighting schemes for molecular graphs were defined by applying eqs 1 and 2 to different atomic properties: Z,[20] when p is the atomic number; A,[24] when p is the atomic mass; P,[24] when p is the atomic polarizability; E,[24] when p is the atomic electronegativity; R,[24] when p is the atomic radius. Similar equations were used to define the X and Y weighting schemes, using different sets of values for the atomic radius and electronegativity.[23]

The AH weighting scheme[24] uses the following equation to define the vertex parameter $Vw(AH)_i$ for the non-hydrogen atom i:

$$\begin{aligned} Vw(AH)_i = 1 - A_C/(A_i + NoH_i A_H) = \\ 1 - 12.011/(A_i + 1.0079 NoH_i) \end{aligned} \tag{3}$$

The edge parameter $Ew(AH)_{ij}$ for the bond between atoms i and j is defined with the equation[24]

$$\begin{aligned} Ew(AH)_{ij} = A_C A_C/Bo_{ij}(A_i + NoH_i A_H)(A_j + NoH_j A_H) = \\ (12.011)12.011/Bo_{ij}(A_i + 1.0079 NoH_i) \times \\ (A_j + 1.0079 NoH_j) \end{aligned} \tag{4}$$

where $A_C = 12.011$ is the atomic mass for carbon, $A_H = 1.0079$ is the atomic mass for hydrogen, NoH_i is the number of hydrogen atoms bonded to the heavy atom i, and NoH_j is the number of hydrogen atoms bonded to the heavy atom j.

Molecular Matrices. The large majority of the topological indices proposed in the literature were derived from the adjacency matrix and the distance matrix. Recently, several new molecular matrices were defined and used to compute new structural descriptors.[18] In this paper the graph descriptors are computed from three molecular matrices: adjacency **A**, distance **D**, and reciprocal distance[25−29] **RD**.

The Vertex Valency. The valency of the vertex v_i, $\mathbf{val}(w)_i = \mathbf{val}(w,G)_i$, is defined as the sum of the weights $Ew(w)_{ij}$ of all edges e_{ij} incident with vertex v_i:[30,33,34,36]

$$\mathbf{val}(w)_i = \sum_{e_{ij}\in E(G)} Ew(w)_{ij} \quad (5)$$

where w is the weighting scheme used to compute the Ew parameters.

The Vertex Sum Operator. Consider the vertex v_i from the graph G with N vertices and the symmetric graph matrix $\mathbf{M}(w) = \mathbf{M}(w,G)$ computed with the weighting scheme w. The vertex sum of the vertex v_i, $\mathbf{VS}(\mathbf{M},w,)_i = \mathbf{VS}(\mathbf{M},w,G)_i$, is defined as the sum of the elements in the column i, or row i, of the molecular matrix $\mathbf{M}$:[15-17,23,30]

$$\mathbf{VS}(\mathbf{M},w,G)_i = \sum_{j=1}^{N}[\mathbf{M}(w)]_{ij} = \sum_{j=1}^{N}[\mathbf{M}(w)]_{ji} \quad (6)$$

The **VS** operator is identical with the degree vector **Deg** if **M** is the adjacency matrix **A**, to the distance sum **DS** if **M** is the distance matrix **D**, and to the reciprocal distance sum[28] **RDS** if **M** is the reciprocal distance matrix **RD**.

The Chi Operator. The highly successful Randić connectivity index[31] χ was extended for connected subgraphs by Kier and Hall:[11,12]

$${}^{m}\chi_t^{\mathrm{v}} = \sum_{i=1}^{s}\prod_{j=1}^{n}(\delta_j^{\mathrm{v}})^{-1/2} \quad (7)$$

where s is the number of connected subgraphs of type t with m edges, n is the number of vertices of the subgraph, and δ_j^{v} is the valence atomic connectivity. The **Chi** operator was derived from the Kier and Hall connectivity indices by replacing the local invariant δ^{v} with any other vertex invariant. Consider a vertex structural descriptor $\mathbf{VSD}(\mathbf{M},w) = \mathbf{VSD}(\mathbf{M},w,G)$ that assigns a numerical invariant $\mathbf{VSD}(\mathbf{M},w)_i$ to each vertex v_i from the VEW molecular graph G. The **Chi** operator $\mathbf{Chi}(\mathbf{VSD},\mathbf{M},w) = \mathbf{Chi}(\mathbf{VSD},\mathbf{M},w,G)$ of the graph G is[30]

$${}^{m}\mathbf{Chi}(\mathbf{VSD},\mathbf{M},w)_t = \sum_{i=1}^{s}\prod_{j=1}^{n}(\mathbf{VSD}(\mathbf{M},w)_j)^{-1/2} \quad (8)$$

where s is the number of connected subgraphs of type t with m edges, n is the number of vertices of the subgraph, and w is the weighting scheme. The summation is calculated over all connected subgraphs with m edges of the following types: path ($t = p$), cluster ($t = c$), and path/cluster ($t = pc$).

The Wiener Operator. Consider the VEW molecular graph G with N vertices and its symmetric molecular matrix $\mathbf{M}(w) = \mathbf{M}(w,G)$ computed with the weighting scheme w. The Wiener operator $\mathbf{Wi}(\mathbf{M},w) = \mathbf{Wi}(\mathbf{M},w,G)$ is[15-17,23,30]

$$\mathbf{Wi}(\mathbf{M},w,G) = \sum_{i=1}^{N}\sum_{j=1}^{N}[\mathbf{M}(w)]_{ij} \quad (9)$$

Using the Wiener operator, the distance matrix **D** gives the Wiener index W, and the reciprocal distance matrix **RD** gives the Harary index.[27,28] A large number of Wiener-type topological indices[19] can be computed with this operator.

The Hyper-Wiener Operator. Consider the vertex- and edge-weighted graph G with N vertices and its molecular matrix $\mathbf{M}(w) = \mathbf{M}(w,G)$ computed with the weighting scheme w. The hyper-Wiener operator $\mathbf{HyWi}(\mathbf{M},w) = \mathbf{HyWi}(\mathbf{M},w,G)$ of the VEW graph G is[17,18,23,30]

$$\mathbf{HyWi}(\mathbf{M},w,G) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}([\mathbf{M}(w)]_{ij}^{\,2} + [\mathbf{M}(w)]_{ij}) \quad (10)$$

If **M** is the distance matrix, the **HyWi** operator is identical with the hyper-Wiener index WW.[32] Topological indices computed with the hyper-Wiener operator were used to develop structure−property models for the boiling points of ethers, peroxides, acetals, and their sulfur analogues.[23]

The Characteristic Polynomial Operator. The characteristic polynomial operator $\mathbf{Ch}(\mathbf{M},w,G,x)$ represents the characteristic polynomial of the matrix $\mathbf{M}(w) = \mathbf{M}(w,G)$ computed with the weighting scheme w for a molecular graph G with N vertices:[15-17,30,33,34]

$$\mathbf{Ch}(\mathbf{M},w,G,x) = \mathbf{det}(x\mathbf{I} - \mathbf{M}(w)) = \sum_{n=0}^{N}c_n x^{N-n} \quad (11)$$

where **I** is the unit matrix of order N and c_n is the nth coefficient of the characteristic polynomial.

The Matrix Spectrum Operator. Consider the VEW graph G with N vertices and its molecular matrix $\mathbf{M}(w) = \mathbf{M}(w,G)$ computed with the weighting scheme w. The matrix spectrum operator $\mathbf{Sp}(\mathbf{M},w,G) = \{x_i,\ i = 1, 2, ..., N\}$ represents the eigenvalues of the matrix $\mathbf{M}(w)$ or the roots of the polynomial $\mathbf{Ch}(\mathbf{M},w,G,x)$, $\mathbf{Ch}(\mathbf{M},w,G,x) = 0$. The $\mathbf{MinSp}(\mathbf{M},w,G)$ and $\mathbf{MaxSp}(\mathbf{M},w,G)$ operators are equal to the minimum and maximum values of $\mathbf{Sp}(\mathbf{M},w,G)$, respectively:[15-17,30,35]

$$\mathbf{MinSp}(\mathbf{M},w,G) = \mathbf{min}\{\mathbf{Sp}(\mathbf{M},w,G)\} \quad (12)$$

$$\mathbf{MaxSp}(\mathbf{M},w,G) = \mathbf{max}\{\mathbf{Sp}(\mathbf{M},w,G)\} \quad (13)$$

Molecular graph descriptors computed with these two spectral operators were used with success in QSPR studies to estimate the boiling points of acyclic compounds containing oxygen or sulfur atoms,[23] to compute the boiling points, heat of vaporization, molar refraction, molar volume, critical pressure, critical temperature, and surface tension of alkanes,[29,35] and to model the amine boiling points.[36] One must note that the BCUT descriptors[8-10] are computed with the same formula as the **MinSp** and **MaxSp** operators, using molecular matrices weighted with a different set of parameters.

The Hosoya Operator. Let $\mathbf{M}(w) = \mathbf{M}(w,G)$ be the molecular matrix computed with the weighting scheme w of a VEW graph G with N vertices. The Hosoya operator $\mathbf{Ho}(\mathbf{M},w) = \mathbf{Ho}(\mathbf{M},w,G)$ is defined as the sum of the absolute values for the coefficients c_n of the characteristic polynomial of the matrix **M**:[30,33,34]

$$\mathbf{Ho}(\mathbf{M},w,G) = \sum_{n=0}^{N}|c_n| \quad (14)$$

Structural descriptors computed with the Hosoya operator from distance-valency matrices were used with good results to develop QSPR models for the boiling point, heat capacity, Gibbs energy, formation enthalpy, refractive index, and density of alkanes.[34]

METHODS

Chemical Database. The chemical compounds used in this study were taken from the AIDS database, Developmental Therapeutics Program, National Cancer Institute.[38] The database contains 32 110 compounds with molecular weight ranging between 26 and 2839; a statistical study of the database shows that 95% of the compounds have molecular weight lower than 660. From the entire database we have randomly selected 2000 compounds containing H, C, N, O, P, S, and halogens, with a molecular weight lower than 660. During the selection of the database, we have considered only the molecular files containing one chemical compound.

Structural Descriptors. Using the operators presented in the previous section, 240 structural descriptors were computed with six weighting schemes. If not otherwise specified, the graph operators were applied to the adjacency **A**, distance **D**, and reciprocal distance **RD** molecular matrices. The list of the descriptors used in the present study is (1) the molecular weight, **MW**; (2) the Kier and Hall connectivity indices ${}^0\chi^v$, ${}^1\chi^v$, ${}^2\chi^v$, ${}^3\chi_p v$, ${}^3\chi_c v$;[11,12] (3) the **Chi** indices[30] ^{0}**Chi**(**VSD**,*w*), ^{1}**Chi**(**VSD**,*w*), ^{2}**Chi**(**VSD**,*w*), ^{3}**Chi**(**VSD**,*w*)$_p$, ^{3}**Chi**(**VSD**,*w*)$_c$, where the vertex invariant **VSD** is the valency **Val**,[33,36] the distance sum **DS**,[15,16] and the reciprocal distance sum **RDS**;[28] (4) the Wiener indices computed with the Wiener operator **Wi**(**M**^{p},*w*) from the *p*th power of the molecular matrix **M**, with *p* between 1 and 5;[39] (5) the hyper-Wiener indices computed with the hyper-Wiener operator **HyWi**(**M**,*w*);[17,30] (6) the spectral operators **MinSp**(**M**,*w*) and **MaxSp**(**M**,*w*);[17,30] and (7) the Hosoya indices computed with the Hosoya operator **Ho**(**A**,*w*).[33,34]

Descriptor Selection. The importance of using orthogonalized descriptors for avoiding redundancy and building efficient QSAR–QSPR models is a well-established issue in the literature.[40–42] Starting from a population of *n* structural descriptors, the selection of sets containing *k* quasi-orthogonal descriptors is a particularly difficult and time-consuming problem. A complete selection from *n* descriptors requires an exhaustive search over a number of subsets of *k* descriptors equal to

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} \tag{15}$$

An additional difficulty is the fact that, for a given maximum intercorrelation coefficient $r_{\max}$, the largest value of *k* is not known a priori.

Due to that the exhaustive search is not feasible in the present case, the sets of quasi-orthogonal descriptors were selected using a heuristic algorithm. The selection rationale that we are presenting below was not based on building a model by linearly projecting the original variables into new latent variables as in factor analysis, but to discover sets of individual descriptors in which these variables are as orthogonal as possible with respect to each other. The heuristic algorithm for the selection of sets of quasi-orthogonal sets of structural descriptors comprises two distinct parts. In the first one, the procedure makes an exhaustive search on all pairs of descriptors to identify all pairs of uncorrelated descriptors. The second part is a Greedy approach that takes a set of *n* descriptors and adds to it a new descriptor, to make an $n + 1$ set of quasi-orthogonal descriptors, whenever the new descriptor is not correlated with the descriptors already selected in that set.

We describe below the main steps of the procedure of selecting sets of descriptors with a pairwise correlation coefficient lower than a threshold $r_{\max}$:

(1) Start from a collection of *N* structural descriptors and compute the intercorrelation matrix.

(2) Select a descriptor *i* and put it in the quasi-orthogonal descriptors list QODL; set the descriptors count to 1.

(3) From the remaining $N - 1$ descriptors select a descriptor *j*; if $|r_{ij}| \le r_{\max}$, add descriptor *j* to QODL and set DC to 2.

(4) Take a descriptor *k* not yet tested; if for all DC descriptors *l* from the QODL set $|r_{kl}| \le r_{\max}$, add descriptor *k* to QODL and increase DC with 1.

(5) Repeat step 4 until all descriptors are tested.

(6) Print the list of quasi-orthogonal descriptors QODL.

(7) Repeat step 3 until all descriptors are tested.

(8) Repeat step 2 for all *N* descriptors.

(9) Inspect all solutions printed in step 6 and select those with the maximal number of descriptors DC.

The heuristic algorithm presented above was extensively tested and offered good solutions with modest computational resources. Although it does not necessarily find the optimal set of quasi-orthogonal descriptors, it is a good compromise when dealing with large sets of descriptors. In the present investigation we have used four values for $r_{\max}$, namely 0.05, 0.10, 0.15, and 0.20.

Classification of the Lists of Quasi-orthogonal Descriptors. For a threshold $r_{\max}$ the descriptor selection algorithm produces a list containing sets of quasi-orthogonal descriptors. Given that this list may be rather long, we introduce here a simple statistical index that measures the pairwise intercorrelation between all descriptors in a set for a rapid ranking of all the sets in a list. For a set of quasi-orthogonal descriptors we define the total squared intercorrelation coefficient:

$$\mathrm{TIC} = \sum_{i=1}^{\mathrm{DC}-1} \sum_{j=i+1}^{\mathrm{DC}} r_{ij}^{\ 2} \tag{16}$$

where DC is the number of descriptors from the set and r_{ij} is the intercorrelation coefficient between descriptors *i* and *j*. This statistical index can be used to compare the sets containing the same number of descriptors obtained with a given $r_{\max}$: the lower the TIC value, the more pairwise orthogonal is the set of descriptors. In the present study, for all thresholds $r_{\max}$ we report only the sets of descriptors with the maximum DC, sorted in increasing order of TIC.

However, the simplicity of this index lies mainly in that it does not incorporate explicit information about multicollinearity. Therefore, for the sets reported below which have been ranked higher according to TIC, we have also calculated the well-established metric of the variance inflation factors (VIF)[43] to additionally test for multicollinearity effects within each set. The VIFs represent the diagonal elements of the inverse correlation matrix which, for variable *i*, is computed as

$$\mathrm{VIF}_i = 1/(1 - R_i^2) \tag{17}$$

where R$_i$ is the global correlation coefficient of variable *i* in terms of the other variables in a set. The closer the VIF$_i$ is to unity the less cross-correlated i is to the other variables in a set.

RESULTS AND DISCUSSION

The statistical analysis of the whole set of 240 structural descriptors indicates that a large number of descriptors are highly intercorrelated. In the virtual screening of combinatorial libraries it is desirable to avoid the use of highly intercorrelated descriptors because they contain related structural information and the computational effort increases without any advantage for the molecular diversity of the selected compounds. What is more, irrelevant dimensions inflate unnecessarily the chemical space, and the selection problem increases exponentially with the number of dimensions. Ideally, in virtual screening one must use a set of orthogonal descriptors that cover the dimensionality of the chemical space. In the real world, the descriptors are more or less intercorrelated and the dimensionality of the chemical space is not known beforehand. The scope of our investigation is to reduce the set of 240 structural descriptors to several sets containing a much lower number of quasi-orthogonal descriptors. To evaluate the dimensionality of the chemical space, we performed principal component analysis (PCA) with standardized variables and used it as a benchmark for the size of the reduced sets of descriptors. Different runs of cross-validated PCA with randomly and systematically selected segments at 10% level of the data set were performed and showed that the suggested number of PCs is about 9. In particular, 8 PCs explained about 89.5% of the variance, 9 PCs about 90.9% of the variance, and 10 PCs 91.8% of the variance.

For the statistical analysis of the 240 structural descriptors, we have used the set of 2000 compounds selected from the NCI AIDS database. We present below the sets of quasi-orthogonal structural descriptors generated with the heuristic approach described in the previous section. For each threshold $r_{\max}$ we report only the sets with a maximal number of orthogonal descriptors.

Basis Sets of Descriptors with $r_{\max}$ equal to 0.05. With this threshold the descriptor selection algorithm found 27 sets consisting of 6 descriptors each. The whole collection, sorted in increasing order of TIC, is given in the Supporting Information; the first five sets of orthogonal descriptors from this collection are presented below. In brackets following the name of each descriptor is its VIF value with respect to the other descriptors in the same set:

(1) **Ho**(**A**,*E*) [1.000 010], **Ho**(**A**,*P*) [1.000 006], **MinSp**(**RD**,*E*) [1.000 561], **MinSp**(**RD**,*P*) [1.000 257], ^{3}**Chi**(**DS**,*AH*)$_{\mathrm{c}}$ [1.000 574], ^{3}**Chi**(**RDS**,*P*)$_{\mathrm{c}}$ [1.000 171]

(2) **Ho**(**A**,*E*) [1.000 013], **Ho**(**A**,*P*) [1.000 006], **MinSp**(**RD**,*E*) [1.000 596], **MinSp**(**RD**,*P*) [1.000 313], ^{3}**Chi**(**DS**,*A*)$_{\mathrm{c}}$ [1.000 586], ^{3}**Chi**(**RDS**,*P*)$_{\mathrm{c}}$ [1.000 148]

(3) **Ho**(**A**,*P*) [1.000 004], **Ho**(**A**,*Z*) [1.000 001], **MinSp**(**RD**,*E*) [1.000 567], **MinSp**(**RD**,*P*) [1.000 284], ^{3}**Chi**(**DS**,*AH*)$_{\mathrm{c}}$ [1.000 583], ^{3}**Chi**(**RDS**,*P*)$_{\mathrm{c}}$ [1.000 173]

(4) **Ho**(**A**,*A*) [1.000 001], **Ho**(**A**,*P*) [1.000 004], **MinSp**(**RD**,*E*) [1.000 567], **MinSp**(**RD**,*P*) [1.000 285], ^{3}**Chi**(**DS**,*AH*)$_{\mathrm{c}}$ [1.000 584], ^{3}**Chi**(**RDS**,*P*)$_{\mathrm{c}}$ [1.000 173]

(5) **Ho**(**A**,*P*) [1.000 004], **Ho**(**A**,*Z*) [1.000 001], **MinSp**(**RD**,*E*) [1.000 603], **MinSp**(**RD**,*P*) [1.000 343], ^{3}**Chi**(**DS**,*A*)$_{\mathrm{c}}$ [1.000 596], ^{3}**Chi**(**RDS**,*P*)$_{\mathrm{c}}$ [1.000 150]

An inspection of the 27 sets of descriptors shows that the majority of the descriptors are computed with the ^{3}**Chi**(**VSD**,*w*)$_{\mathrm{c}}$ operator, namely 54; other operators that have a fairly high presence are **MinSp**, 42 times; **Ho**, 40 times; and **Wi**, 14 times. The descriptors computed with the **MaxSp**, **HyWi**, ^{0}**Chi**, ^{1}**Chi**, and ^{2}**Chi** operators were not included in this collection of quasi-orthogonal descriptors. As pointed out in the previous section, Pearlman proposed the BCUT descriptors as a metric for determining the molecular diversity;[8−10] our analysis shows the importance of the related descriptors **MinSp**, but points also to the high relevance of the ^{3}**Chi**(**VSD**,*w*)$_{\mathrm{c}}$ indices that represent the contribution of isobutane-like subgraphs, weighted with different atomic invariants. Another important source of structural descriptors is the Hosoya operator **Ho**.

The **MinSp**, **Ho**, and **Wi** indices are mainly derived from the reciprocal distance matrix: **RD**, 42 times; **A**, 40 times; **D**, 14 times. Traditionally, the topological indices were derived mainly from the adjacency and distance matrices; our finding indicates that other molecular matrices, the reciprocal distance in our case, can generate useful structural descriptors. Because a large number of molecular matrices were recently introduced,[18,33,36] it will be of interest to extend the search of quasi-orthogonal descriptors for the new matrices. Among the **Chi** indices, 33 are derived from **DS**, 27 from **RDS**, and 6 from **Val**. This finding indicates that **Chi** descriptors derived from the distance or reciprocal distance matrices provide useful indices for measuring the molecular diversity. The analysis of the weighting schemes shows that the *P* weights derived from the atomic polarizability, and based on the experimental values reported by Nagle,[44] are used with the greatest frequency: *P*, 81 times; *E*, 35 times; *A*, 22 times; *AH*, 13 times; *Z*, 8 times; *R*, 3 times. Two descriptors are present in all 27 sets, namely **Ho**(**A**,*P*) and ^{3}**Chi**(**RDS**,*P*)$_{\mathrm{c}}$, while two other descriptors are present in 21 sets, namely **MinSp**(**RD**,*E*) and **MinSp**(**RD**,*P*). The Wiener indices derived from the adjacency and reciprocal distance matrices, **Wi**(**A**,*w*) and **Wi**(**RD**,*w*), were not selected, while **Wi**(**D**^{4},*w*) was selected 6 times and **Wi**(**D**^{5},*w*) 8 times. Although the Wiener indices are useful in developing QSPR and QSAR models, they are not of great use in selecting quasi-orthogonal descriptors; only those derived from the higher powers of the distance matrix were retained by the selection algorithm.

Basis Sets of Descriptors with $r_{\max}$ equal to 0.10. A total of 96 sets containing 7 descriptors each were obtained for this threshold. The complete list of descriptors in each set is presented in the Supporting Information. From this collection of quasi-orthogonal descriptors we present the first five sets below. Again in brackets following each descriptor is its VIF with respect to the other descriptors in the same set:

(6) **Ho**(**A**,*E*) [1.000 917], **Ho**(**A**,*P*) [1.000 114], **MinSp**(**RD**,*E*) [1.000 006], **MinSp**(**RD**,*P*) [1.004 272], ^{3}**Chi**(**DS**,*A*)$_{\mathrm{c}}$ [1.004 529], ^{0}**Chi**(**DS**,*Z*) [1.001 988], ^{3}**Chi**(**RDS**,*P*)$_{\mathrm{c}}$ [1.000 043]

(7) **Ho**(**A**,*P*) [1.000 118], **Ho**(**A**,*Z*) [1.000 294], **MinSp**(**RD**,*E*) [1.000 006], **MinSp**(**RD**,*P*) [1.004 365], ^{3}**Chi**(**DS**,*A*)$_{\mathrm{c}}$ [1.004 528], ^{0}**Chi**(**DS**,*Z*) [1.002 336], ^{3}**Chi**(**RDS**,*P*)$_{\mathrm{c}}$ [1.000 044]

(8) **Ho**(**A**,*A*) [1.000 286], **Ho**(**A**,*P*) [1.000 118], **MinSp**(**RD**,*E*) [1.000 006], **MinSp**(**RD**,*P*) [1.004 366], ^{3}**Chi**(**DS**,*A*)$_{\mathrm{c}}$ [1.004 528], ^{0}**Chi**(**DS**,*Z*) [1.002 330], ^{3}**Chi**(**RDS**,*P*)$_{\mathrm{c}}$ [1.000 044]

(9) **Ho**(**A**,*A*) [1.000 280], **Ho**(**A**,*P*) [1.000 117], **MinSp**(**RD**,*E*) [1.000 002], **MinSp**(**RD**,*P*) [1.004 266], ^{3}**Chi**(**D**-

Table 1. Intercorrelation Matrix for the 11 Descriptors that Belong to the Three Sets of Descriptors with a Correlation Coefficient Lower than 0.15: **1**, **Ho**(**A**,*P*); **2**, **Ho**(**A**,*R*); **3**, **MinSp**(**A**,*E*); **4**, **MinSp**(**RD**,*E*); **5**, **MinSp**(**RD**,*P*); **6**, **Wi**(**D**^{5},*A*); **7**, **Wi**(**D**^{5},*E*); **8**, **Wi**(**D**^{5},*Z*); **9**, ^{3}**Chi**(**DS**,*A*)$_{c}$; **10**, ^{0}**Chi**(**DS**,*E*); **11**, ^{2}**Chi**(**RDS**,*P*)

descriptor	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1	1.000	0.054	−0.005	−0.024	−0.002	0.024	0.022	0.024	−0.011	0.035	0.057
2	0.054	1.000	−0.078	0.006	−0.006	0.126	0.122	0.126	−0.015	0.082	0.068
3	−0.005	−0.078	1.000	−0.100	−0.088	−0.015	−0.015	−0.015	0.076	−0.074	−0.099
4	−0.024	0.006	−0.100	1.000	0.037	−0.004	−0.003	−0.004	0.014	−0.138	−0.052
5	−0.002	−0.006	−0.088	0.037	1.000	0.014	0.012	0.014	−0.045	0.065	0.145
6	0.024	0.126	−0.015	−0.004	0.014	1.000	*0.999*	1.000	−0.020	−0.084	0.058
7	0.022	0.122	−0.015	−0.003	0.012	*0.999*	1.000	*0.999*	−0.020	−0.087	0.056
8	0.024	0.126	−0.015	−0.004	0.014	*1.000*	0.999	1.000	−0.020	−0.084	0.058
9	−0.011	−0.015	0.076	0.014	−0.045	−0.020	−0.020	−0.020	1.000	−0.047	−0.083
10	0.035	0.082	−0.074	−0.138	0.065	−0.084	−0.087	−0.084	−0.047	1.000	−0.013
11	0.057	0.068	−0.099	−0.052	0.145	0.058	0.056	0.058	−0.083	−0.013	1.000

S,*AH*)$_{c}$ [1.005 602], ^{0}**Chi**(**DS**,*Z*) [1.002 897], ^{3}**Chi**(**RDS**,*P*)$_{c}$ [1.000 062]

(10) **Ho**(**A**,*P*) [1.000 118], **MinSp**(**RD**,*E*) [1.000 010], **MinSp**(**RD**,*P*) [1.004 577], **Wi**(**D**^{5},*A*) [1.000 584], ^{3}**Chi**(**DS**,*A*)$_{c}$ [1.004 317], ^{0}**Chi**(**DS**,*Z*) [1.016 477], ^{3}**Chi**(**RDS**,*P*)$_{c}$ [1.000 070]

The analysis of the 96 sets of descriptors indicates that the ^{3}**Chi**(**VSD**,*w*)$_{c}$ operator is the source of a large number of indices retained by the selection algorithm: ^{3}**Chi**(**VSD**,*w*)$_{c}$, 177 times; **MinSp**, 156 times; **Ho**, 137 times; ^{0}**Chi**(**VSD**,*w*), 96 times; **Wi**, 55 times; ^{2}**Chi**(**VSD**,*w*), 23 times; ^{3}**Chi**(**VSD**,*w*)$_{p}$, 23 times; ^{1}**Chi**(**VSD**,*w*), 5 times. The descriptors computed with the **MaxSp** and **HyWi** operators were not included in this collection of quasi-orthogonal descriptors. The analysis of the collection containing quasi-orthogonal descriptors at the level $r_{max} = 0.10$ points again to the importance of the indices derived from the ^{3}**Chi**(**VSD**,*w*)$_{c}$, **MinSp**, and **Ho** operators.

Among the **Chi** indices, there is a large difference between the presence of ^{3}**Chi**(**VSD**,*w*)$_{c}$ and ^{3}**Chi**(**VSD**,*w*)$_{p}$ (an index that takes into account the contribution of butane-like subgraphs). The importance of the distance sum and reciprocal distance sum in generating relevant **Chi** indices is observed again: 192 **Chi** indices are derived from **DS**, 97 from **RDS**, and 35 from **Val**. It appears that the valency is less fit than **DS** and **RDS** in generating quasi-orthogonal descriptors.

The inspection of the weighting schemes used to compute the indices shows again the importance of the *P* weights: *P*, 293 times; *E*, 187 times; *A*, 90 times; *Z*, 56 times; *AH*, 30 times; *R*, 16 times. The above order is almost identical with that observed in the $r_{max} = 0.05$ case, with only one inversion, between *Z* and *AH*. The **MinSp**, **Ho**, and **Wi** indices are mainly derived from the adjacency matrix: **A**, 202 times; **RD**, 91 times; **D**, 55 times.

One descriptor appears in all 96 sets, namely **Ho**(**A**,*P*), while several other indices have a high frequency: ^{3}**Chi**(**RDS**,*P*)$_{c}$, 85 times; ^{0}**Chi**(**DS**,*E*), 66 times; **MinSp**(**A**,*E*), 65 times; **MinSp**(**RD**,*P*), 61 times. Identically with the situation in the $r_{max} = 0.05$ case, the Wiener indices derived from the adjacency and reciprocal distance matrices, **Wi**(**A**,*w*) and **Wi**(**RD**,*w*), were not selected, while **Wi**(**D**^{3},*w*) was selected 9 times, **Wi**(**D**,^{4}*w*) 20 times, and **Wi**(**D**,^{5}*w*) 26 times.

Basis Sets of Descriptors with r_{max} equal to 0.15. Using a threshold $r_{max} = 0.15$, we have obtained the following three sets of nine orthogonal descriptors, along with the indicated corresponding VIF values with respect to the other descriptors in the same set:

(11) **Ho**(**A**,*P*) [1.001 435], **Ho**(**A**,*R*) [1.009 551], **MinSp**(**A**,*E*) [1.029 633], **MinSp**(**RD**,*E*) [1.026 894], **MinSp**(**RD**,*P*) [1.021 002], **Wi**(**D**^{5},*A*) [1.007 943], ^{3}**Chi**(**DS**,*A*)$_{c}$ [1.005 154], ^{0}**Chi**(**DS**,*E*) [1.040 236], ^{2}**Chi**(**RDS**,*P*) [1.034 436]

(12) **Ho**(**A**,*P*) [1.001 497], **Ho**(**A**,*R*) [1.009 692], **MinSp**(**A**,*E*) [1.029 665], **MinSp**(**RD**,*E*) [1.026 915], **MinSp**(**RD**,*P*) [1.020 908], **Wi**(**D**^{5},*E*) [1.007 929], ^{3}**Chi**(**DS**,*A*)$_{c}$ [1.005 151], ^{0}**Chi**(**DS**,*E*) [1.040 355], ^{2}**Chi**(**RDS**,*P*) [1.034 758]

(13) **Ho**(**A**,*P*) [1.001 433], **Ho**(**A**,*R*) [1.009 552], **MinSp**(**A**,*E*) [1.029 633], **MinSp**(**RD**,*E*) [1.026 894], **MinSp**(**RD**,*P*) [1.021 002], **Wi**(**D**^{5},*Z*) [1.007 939], ^{3}**Chi**(**DS**,*A*)$_{c}$ [1.005 154], ^{0}**Chi**(**DS**,*E*) [1.040 236], ^{2}**Chi**(**RDS**,*P*) [1.034 447]

An inspection of the above lists reveals that eight out of nine descriptors are common in the three sets; moreover, although the sixth descriptor is different in each list, its origin is common, being derived from the Wiener operator applied to the fifth power of the distance matrix computed with different weighting schemes: **Wi**(**D**^{5},*A*), **Wi**(**D**^{5},*E*), **Wi**(**D**^{5},*Z*). In Table 1 we present the intercorrelation matrix for all 11 descriptors that appear in these three lists; from this matrix one can see that the three descriptors derived from **Wi**(**D**^{5}) are highly intercorrelated (their intercorrelation coefficients are highlighted in italics in Table 1), while all the remaining correlation coefficients are lower than the threshold.

Each set obtained at the level $r_{max} = 0.15$ contains two Hosoya indices, three **MinSp** indices, three **Chi** indices, and one **Wi** index; this situation is in line with our observations for the previous two collections' quasi-orthogonal descriptors. While the **Ho**, **MinSp**, and **Wi** indices represent a global measure of the molecular shape and size, the **Chi** indices represent weighted contributions of various subgraphs from the molecular graph. The **Chi** index ^{0}**Chi**(**DS**,*E*) collects the contribution of the non-hydrogen atoms in the molecular graph; this contribution, weighted by the distance sum **DS**, is a measure of molecular size. The index ^{2}**Chi**(**RDS**,*P*) represents the contribution of propane-like subgraphs, weighted by the reciprocal distance sum **RDS**, while ^{3}**Chi**(**DS**,*A*)$_{c}$ represents the contribution of isobutane-like subgraphs, weighted by the distance sum **RD**; both these indices represent a measure of molecular branching.

Basis Sets of Descriptors with r_{max} equal to 0.20. The descriptor selection algorithm found in this case 50 sets consisting of 9 descriptors each. The whole collection of orthogonal descriptors at the 0.20 threshold is given in the Supporting Information, sorted in increasing order of TIC; the first five sets from this collection are presented below with their VIF values with respect to the other descriptors in the same set:

(14) **Ho**(**A**,*P*) [1.000 885], **Ho**(**A**,*R*) [1.024 842], **MinSp**(**RD**,*E*) [1.000 246], **MinSp**(**RD**,*P*) [1.046 463], **Wi**(**A**^{4},*A*) [1.113 586], **Wi**(**D**^{4},*A*) [1.034 447], ^{3}**Chi**(**DS**,*A*)$_c$ [1.024 800], ^{0}**Chi**(**DS**,*P*) [1.055 220], ^{3}**Chi**(**RDS**,*P*)$_c$ [1.022 966]

(15) **Ho**(**A**,*P*) [1.000 903], **Ho**(**A**,*R*) [1.024 800], **MinSp**(**RD**,*E*) [1.000 239], **MinSp**(**RD**,*P*) [1.046 211], **Wi**(**A**,^{4}*Z*) [1.113 586], **Wi**(**D**^{4},*A*) [1.034 447], ^{3}**Chi**(**DS**,*A*)$_c$ [1.024 779], ^{0}**Chi**(**DS**,*P*) [1.055 334], ^{3}**Chi**(**RDS**,*P*)$_c$ [1.023 238]

(16) **Ho**(**A**,*P*) [1.000 921], **Ho**(**A**,*R*) [1.010 090], **MaxSp**(**A**,*A*) [1.128 668], **MinSp**(**RD**,*E*) [1.030 418], **Wi**(**D**^{5},*E*) [1.008 798], **Wi**(**RD**^{5},*A*) [1.117 318], ^{0}**Chi**(**DS**,*E*) [1.107 665], ^{3}**Chi**(**DS**,*E*)$_p$ [1.075 581], ^{3}**Chi**(**RDS**,*P*)$_c$ [1.008 771]

(17) **Ho**(**A**,*P*) [1.000 880], **Ho**(**A**,*R*) [1.009 939], **MaxSp**(**A**,*A*) [1.128 668], **MinSp**(**RD**,*E*) [1.030 386], **Wi**(**D**^{5},*Z*) [1.008 504], **Wi**(**RD**^{5},*A*) [1.119 821], ^{0}**Chi**(**DS**,*E*) [1.107 800], ^{3}**Chi**(**DS**,*E*)$_p$ [1.075 512], ^{3}**Chi**(**RDS**,*P*)$_c$ [1.008 769]

(18) **Ho**(**A**,*P*) [1.000 881], **Ho**(**A**,*R*) [1.009 938], **MaxSp**(**A**,*A*) [1.128 668], **MinSp**(**RD**,*E*) [1.030 386], **Wi**(**D**^{5},*A*) [1.008 509], **Wi**(**RD**^{5},*A*) [1.119 821], ^{0}**Chi**(**DS**,*E*) [1.107 800], ^{3}**Chi**(**DS**,*E*)$_p$ [1.075 512], ^{3}**Chi**(**RDS**,*P*)$_c$ [1.008 769]

The examination of the 50 sets of descriptors indicates that the majority of the descriptors are computed with the **MinSp** operator, namely 118; other operators that have a fairly high presence are **Ho**, 100 times; ^{3}**Chi**(**VSD**,*w*)$_c$, 82 times; **Wi**, 82 times; and ^{0}**Chi**(**VSD**,*w*), 50 times. The descriptors computed with the **HyWi**, ^{1}**Chi**, and ^{2}**Chi** operators were not included in this collection of quasi-orthogonal descriptors.

The **MinSp**, **Ho**, **Wi**, and **MaxSp** indices are mainly derived from the adjacency matrix: **A**, 177 times; **RD**, 82 times; **D**, 50 times. Among the **Chi** indices, 100 are derived from **DS**, 23 from **RDS**, and 18 from **Val**. This finding confirms the previous observations indicating that the distance sum invariant generates **Chi** descriptors quasi-orthogonal with the **Ho** and **MinSp** indices. The analysis of the weighting schemes shows that the atomic polarizability weights *P* are used with the greatest frequency: *P*, 116 times; *R*, 100 times; *E*, 70 times; *A*, 94 times; *Z*, 51 times; *AH*, 19 times. Although in the previous collections of descriptors the atomic radius weights *R* were not used with a great frequency, for $r_{\max}$ equal to 0.20 this weighting scheme ocupies the second position, and surpasses the electronegativity weighting scheme *E*.

Two descriptors are present in all 50 sets, namely **Ho**(**A**,*P*) and **Ho**(**A**,*R*), while several other indices have a high frequency: ^{3}**Chi**(**DS**,*Z*)$_c$, 39 times; **MinSp**(**RD**,*R*), 36 times; **MinSp**(**A**,*P*), 27 times; ^{0}**Chi**(**DS**,*E*), 27 times; ^{3}**Chi**(**RDS**,*P*)$_c$, 23 times. From the Wiener indices derived from the adjacency matrix, **Wi**(**A**,*w*), **Wi**(**RD**^{5},*w*) was selected 27 times, **Wi**(**D**^{4},*w*) 25 times, and **Wi**(**D**^{5},*w*) 25 times. This observation indicates that Wiener indices derived from the higher powers of the distance and reciprocal distance matrices generate **Wi** indices that are quasi-orthogonal with the **MinSp**, **Ho**, and **Chi** indices retained by the selection algorithm.

By inspection of all 18 basis sets presented above, it is evident that in all cases all VIF values follow the same pattern; i.e., they are invariable very close to unity, a clear demonstration of total absence of multicollinearity in addition to pairwise correlations among the descriptors in all of these sets.

After such an important reduction of the number of descriptors, from 240 to 6 (for $r_{\max}$ equal to 0.05), 7 (for $r_{\max}$ equal to 0.10), and 9 (for $r_{\max}$ equal to 0.15 and 0.20), it is important to determine the similarity between the chemical distance measured with the whole set of 240 descriptors and that measured with one of the sets of quasi-orthogonal structural descriptors presented in this work. Because in structural diversity screening of a database one selects the most distant compounds, we have investigated the way in which this selection is influenced by the decrease in the number of descriptors.

From the results reported above, we have selected for further investigation the sets 11, 12, and 13, due to that their "dimensionality" is 9, that is, equal to the apparent dimensionality of the chemical space as it was defined by the PCA and also because they resulted from a low value of the maximum intercorrelation coefficient ($r_{\max} = 0.15$).

Validation of the Basis Sets of Quasi-orthogonal Descriptors. The validity of these three sets has been checked by performing a clustering of the complete set of 240 structural descriptors. The complete linkage, single linkage, and Ward's hierarchical clustering methods have been applied using the squared intercorrelation coefficient as metric. If one wants to use a common cluster level of all methods tested, this level must be fixed to 28, a number which may be considered too large. However, close examination of the agglomeration process shows that some methods, such as Ward's from the hierarchical ones, treat "outlier" objects as a group until much later than the other methods, with the rationale that they are "rather different" as compared to the rest of the objects. If we eliminate such methods from consideration, the other hierarchical methods separated well the members of different basis sets in distinct clusters much earlier in the agglomeration process: at around 15-cluster level for complete linkage, at 6-cluster level (for the six-member basis sets) and 9-cluster level (for the nine-member sets) for single linkage. Despite the fact that these clustering methods give slightly different clusters, in all cases the descriptors from a basis set fall in distinct clusters. This observation supports the assumption that they all convey individually different information concerning the molecular diversity.

Evaluation of the Chemical Distance Conservation. By imposing the condition of quasi-orthogonality, the initial set of 240 structural descriptors was reduced to several basis sets of up to 9 descriptors. After such an important reduction in the number of descriptors, it is important to verify that the chemical distance measured with the proposed basis sets is not greatly distorted when compared with that determined with the initial set of descriptors. Because we are particularly interested in the structural diversity characterization, we have investigated the extent to which for any given molecule the same most dissimilar molecules are selected by the initial set of descriptors and the basis sets of quasi-orthogonal descriptors. However, and in order to take into account possible distortions of the initial 240-dimensional "large" chemical space due to highly correlated dimensions (descriptors), we compared for each molecule its list of most dissimilar molecules not on the space defined by the entire collection of the 240 individual descriptors, but on the PCA space, against the same list on the space defined by the basis sets of quasi-orthogonal descriptors.

All descriptors were autoscaled, to remove the bias that favors descriptors with large values when computing the

Table 2. Occurrence of the *k* Furthest Molecules (kFN_S) Determined with the Small Set of Descriptors (11) in the List of the 10 Furthest Molecules Determined with the 9 First Principal Components ($10FN_{PC}$)

	$10FN_{PC}$									
kFN_S	1	2	3	4	5	6	7	8	9	10
1	1997									
2	1990	10								
3	54	1946	0							
4	41	1922	37	0						
5	40	1785	170	5	0					
6	1	525	1436	38	0	0				
7	1	370	1445	180	4	0	0			
8	1	42	618	1314	25	0	0	0		
9	0	7	529	1419	45	0	0	0	0	
10	0	2	160	1270	561	7	0	0	0	0

Table 3. Occurrence of the *k* Furthest Molecules (kFN_S) Determined with the Small Set of Descriptors (12) in the List of the 10 Furthest Molecules Determined with the 9 First Principal Components ($10FN_{PC}$)

	$10FN_{PC}$									
kFN_S	1	2	3	4	5	6	7	8	9	10
1	1997									
2	1990	10								
3	54	1946	0							
4	41	1922	37	0						
5	40	1785	170	5	0					
6	1	525	1436	38	0	0				
7	1	370	1445	179	5	0	0			
8	0	43	617	1315	25	0	0	0		
9	0	3	524	1426	47	0	0	0	0	
10	0	2	63	607	1294	34	0	0	0	0

Table 4. Occurrence of the *k* Furthest Molecules (kFN_S) Determined with the Small Set of Descriptors (13) in the List of the 10 Furthest Molecules Determined with the 9 First Principal Components ($10FN_{PC}$)

	$10FN_{PC}$									
kFN_S	1	2	3	4	5	6	7	8	9	10
1	1997									
2	1990	10								
3	54	1946	0							
4	41	1922	37	0						
5	40	1785	170	5	0					
6	1	525	1436	38	0	0				
7	1	370	1445	180	4	0	0			
8	1	42	618	1314	25	0	0	0		
9	0	7	529	1419	45	0	0	0	0	
10	0	2	156	1269	566	7	0	0	0	0

chemical distance. In a first experiment we have used the descriptor projections on the 9 first principal components (as the apparent rank of the PCA space is about 9) to construct the $10FN_{PC}$ lists, and for each compound from the database of 2000 molecules we have selected the 10 most distant molecules, denoted for each molecule as $10FN_{PC}$, i.e., its 10 furthest neighbors on the PCA space. For each reduced set we have compared the occurrence of the kFN_S (*k*-furthest neighbors, the list of the *k* most distant molecules determined by the small set of descriptors) into the $10FN_{PC}$ list.

For the quasi-orthogonal sets 11, 12, and 13 we have obtained the results presented in Tables 2, 3, and 4, which differ slightly for the three basis sets used. For all three sets the most distant molecule is present in the $10FN_{PC}$ list in 1997 cases, while the entire $2FN_S$ list is found in the $10FN_{PC}$ list for 10 molecules. However, for the remaining 1990 molecules at least 1 molecule of the two most distant in $2FN_S$ is present in $10FN_{PC}$. The results are also good for the $3FN_S$ list for which at least 2 molecules are present in the $10FN_{PC}$ list for 1946 molecules and at least 1 for the other 54 molecules. For the $4FN_S$ list, 3 out of 4 most distant molecules are present in the $10FN_{PC}$ list for 37 molecules, but 2 out of 4 are present for 1922 molecules and at least 1 out of 4 is present for the remaining 41 molecules. Similar patterns appear for the rest of the kFN_S lists meaning that for the major part of the *k* most distant molecules, although not necessarily exactly all *k* of those, will be retrieved by the small sets, too. These findings add to that despite the great reduction of the number of descriptors, from 240 to 9, the chemical distance is not greatly distorted.

CONCLUDING REMARKS

Virtual screening methods can offer the best set of compounds for a combinatorial synthetic scheme to maximize the chances of finding a drug lead, using both efficient statistical techniques (such as clustering algorithms) and a set of numerical descriptors of the chemical structure. Together with fingerprints and pharmacophores, topological indices represent a good choice of descriptors for a fast and efficient virtual screening of large combinatorial libraries. Once the number of compounds is reduced, geometrical and quantum descriptors can be used for the lead selection. In the present investigation we have explored new topological indices as potential structural descriptors for the characterization of molecular diversity. A database of 2000 compounds randomly selected from the National Cancer Institute AIDS database was used to measure the intercorrelation of the descriptors. The initial set of 240 structural descriptors was reduced, by means of a novel heuristic algorithm, to several quasi-orthogonal sets of up to 9 descriptors, using different thresholds for the maximum intercorrelation coefficient between any pair of descriptors within any given set. The sizes of the basis sets thus obtained were in good agreement with the results of a performed PCA showing that the algorithm can recover the true dimensionality of the chemical space in a satisfactory manner. Moreover, a comparison of the diversity measure offered by the projection of the entire collection of 240 descriptors on the PCA space and that of the quasi-orthogonal sets demonstrates that there is a good degree of similarity between them.

It would be interesting, in a further study, to compare the efficiency of this novel heuristic approach with other, recently proposed, powerful selection methods based on multiregression analysis.[45,46]

The sets of quasi-orthogonal descriptors proposed in this work can be used to design optimally diverse libraries. All topological indices selected in these sets can be easily computed from the molecular graph of any organic compound. The topological indices complement the structural information contained in other descriptors such as fingerprints and pharmacophores to allow for a rapid and efficient virtual screening of large combinatorial libraries.

ACKNOWLEDGMENT

O.I. thanks the Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche of France for a PAST grant and acknowledges the kind hospitality of the LARTIC group during his stay in Nice. We acknowledge the partial financial support of this research by the Romanian Ministry of National Education under Grant 330 84 T96.

Supporting Information Available: Tables listing collection of 27 basis sets of 6 descriptors each, sorted in increasing order of TIC (11 pages). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Warr, W. A. Combinatorial Chemistry. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer III, H. F., Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, 1998; pp 401−417.

(2) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.

(3) Martin, Y. C.; Bures, M. G.; Brown, R. D. Validated Descriptors for Diversity Measurements and Optimization. *Pharm. Pharmacol. Commun.* **1998**, *4*, 147−152.

(4) Grassy, G.; Calas, B.; Yasri, A.; Lahana, R.; Woo, J.; Iyer, S.; Kaczorek, M.; Floc'h, R.; Buelow, R. Computer-Assisted Rational Design of Immunosuppressive Compounds. *Nature Biotechnol.* **1998**, *16*, 748−752.

(5) Martin, E. J.; Critchlow, R. E. Beyond Mere Diversity: Tailoring Combinatorial Libraries for Drug Discovery. *J. Comb. Chem.* **1999**, *1*, 32−45.

(6) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* **1999**, *1*, 55−68.

(7) Bayada, D. M.; Hamersma, H.; van Geerestein, V. J. Molecular Diversity and Representativity in Chemical Databases. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1−10.

(8) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. *Perspect. Drug Discovery Des.* **1998**, *9/10/11*, 339−353. Pearlman, R. S.; Smith, K. M. Software for Chemical Diversity in the Context of Accelerated Drug Discovery. *Drugs Future* **1998**, *23*, 885−895.

(9) Stanton, D. T. Evaluation and Use of BCUT Descriptors in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 11−20.

(10) Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.

(11) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.

(12) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press: Letchworth, 1986.

(13) Ivanciuc, O.; Balaban, A. T. Graph Theory in Chemistry. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer III, H. F., Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, 1998; pp 1169−1190.

(14) Balaban, A. T.; Ivanciuc, O. Historical Development of Topological Indices. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: Amsterdam, 1999; pp 21−57.

(15) Ivanciuc, O.; Balaban, A. T. The Graph Description of Chemical Structures. In: *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: Amsterdam, 1999; pp 59−167.

(16) Ivanciuc, O.; Ivanciuc, T.; Balaban, A. T. Vertex- and Edge-Weighted Molecular Graphs and Derived Structural Descriptors. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: Amsterdam, 1999; pp 169−220.

(17) Ivanciuc, O.; Ivanciuc, T. Matrices and Structural Descriptors Computed from Molecular Graph Distances. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: Amsterdam, 1999; pp 221−277.

(18) Ivanciuc, O.; Ivanciuc, T.; Diudea, M. V. Molecular Graph Matrices and Derived Structural Descriptors. *SAR QSAR Environ. Res.* **1997**, *7*, 63−87.

(19) Diudea, M. V.; Gutman, I. Wiener-Type Topological Indices. *Croat. Chem. Acta* **1998**, *71*, 21−51.

(20) Barysz, M.; Jashari, G.; Lall, R. S.; Srivastava V. K.; Trinajstić, N. On the Distance Matrix of Molecules Containing Heteroatoms. In *Chemical Applications of Topology and Graph Theory*; King, R. B., Ed.; Elsevier: Amsterdam, 1983; pp 222−227.

(21) Balaban, A. T. Chemical Graphs. Part 48. Topological Index J for Heteroatom-Containing Molecules Taking Into Account Periodicities of Element Properties. *MATCH* (*Commun. Math. Chem.*) **1986**, *21*, 115−122.

(22) Balaban, A. T.; Ivanciuc, O. FORTRAN 77 Computer Program for Calculating the Topological Index J for Molecules Containing Heteroatoms. In *MATH/CHEM/COMP 1988*, Proceedings of an International Course and Conference on the Interfaces Between Mathematics, Chemistry and Computer Sciences, Dubrovnik, Yugoslavia, 20−25 June 1988: Graovac, A., Ed.; Elsevier: Amsterdam; *Stud. Phys. Theor. Chem.* **1989**, *63*, 193−212.

(23) Ivanciuc, O.; Ivanciuc, T.; Balaban, A. T. Design of Topological Indices. Part 10. Parameters Based on Electronegativity and Covalent Radius for the Computation of Molecular Graph Descriptors for Heteroatom-Containing Molecules. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 395−401.

(24) Ivanciuc, O. Design of Topological Indices. Part 12. Parameters for Vertex- and Edge-Weighted Molecular Graphs. *Rev. Roum. Chim.* **2000**, *45*.

(25) Ivanciuc, O. Design on Topological Indices. 1. Definition of a Vertex Topological Index in the Case of 4-Trees. *Rev. Roum. Chim.* **1989**, *34*, 1361−1368.

(26) Balaban, T. S.; Filip, P. A.; Ivanciuc, O. Computer Generation of Acyclic Graphs Based on Local Vertex Invariants and Topological Indices. Derived Canonical Labeling and Coding of Trees and Alkanes. *J. Math. Chem.* **1992**, *11*, 79−105.

(27) Plavšić, D.; Nikolić, S.; Trinajstić, N.; Mihalić, Z. On the Harary Index for the Characterization of Chemical Graphs. *J. Math. Chem.* **1993**, *12*, 235−250.

(28) Ivanciuc, O.; Balaban, T.-S.; Balaban, A. T. Design of Topological Indices. Part 4. Reciprocal Distance Matrix, Related Local Vertex Invariants and Topological Indices. *J. Math. Chem.* **1993**, *12*, 309−318.

(29) Diudea, M. V.; Ivanciuc, O.; Nikolić, S.; Trinajstić, N. Matrices of Reciprocal Distance, Polynomials and Derived Numbers. *MATCH* (*Commun. Math. Comput. Chem.*) **1997**, *35*, 41−64.

(30) Ivanciuc, O. Design of Topological Indices. Part 19. Computation of Vertex and Molecular Graph Structural Descriptors with Operators. *Rev. Roum. Chim.* **2000**, *45*.

(31) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* 1975, *97*, 6609−6615.

(32) Klein, D. J.; Lukovits, I.; Gutman, I. On the Definition of the Hyper-Wiener Index for Cycle-Containing Structures. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 50−52.

(33) Ivanciuc, O. Design of Topological Indices. Part 11. Distance-Valency Matrices and Derived Molecular Graph Descriptors. *Rev. Roum. Chim.* **2000**, *45*.

(34) Ivanciuc, O. Design of Topological Indices. Part 18. Modeling the Physical Properties of Alkanes with Molecular Graph Descriptors Derived from the Hosoya Operator. *Rev. Roum. Chim.* **2000**, *45*.

(35) Ivanciuc, O.; Diudea, M. V.; Khadikar, P. V. New Topological Matrices and Their Polynomials. *Ind. J. Chem.* **1998**, *37A*, 574−585.

(36) Ivanciuc, O. Design of Topological Indices. Part 14. Distance-Valency Matrices and Structural Descriptors for Vertex- and Edge-Weighted Molecular Graphs. *Rev. Roum. Chim.* **2000**, *45*.

(37) Laidboeur, T. Représentation et Manipulations Informatiques des Structures et de l'Information Chimique. Doctoral Thesis, Faculty of Sciences, University of Nice-Sophia Antipolis, Nice, France, 1996.

(38) National Cancer Institute, Developmental Therapeutics Program, AIDS database, http://dtp.nci.nih.gov/.

(39) Ivanciuc, O. Design of Topological Indices. Part 16. Matrix Power Operators for Molecular Graphs. *Rev. Roum. Chim.* **2000**, *45*.

(40) Lučić, B.; Nikolić, S.; Trinajstić, N.; Juretić, D. The Structure−Property Models Can Be Improved Using the Orthogonalized Descriptors.*J. Chem. Inf. Comput. Sci.* 1995, *35*, 532−538.

(41) Šoškić, M.; Plavšić, D.; Trinajstić, N. Link Between Orthogonal and Standard Multiple Linear Regression Models. *J. Chem. Inf. Comput. Sci.* 1996, *36*, 829−832.

(42) Klein, D. J.; Randić, M.; Babić, D.; Lučić, B.; Nikolić, S.; Trinajstić, N. Hierarchical Orthogonalization of Descriptors. *Int. J. Quantum Chem.* **1997**, *63*, 215−222.

(43) Beisley, D. A.; Kuh, E.; Welsh, R. E. *Regression Diagnosis*; Wiley: New York, 1980.

(44) Nagle, J. K. Atomic Polarizability and Electronegativity. *J. Am. Chem. Soc.* **1990**, *112*, 4741−4747.

(45) Lučić, B.; Trinajstić, N. Multivariate Regression Outperforms Several Robust Architectures of Neural Networks in QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 121−132.

(46) Lučić, B.; Trinajstić, N.; Sild, S.; Karelson, M.; Katritzky, A. R. A New Efficient Approach for Variable Selection Based on Multiregression: Prediction of Gas Chromatography Retention Times and Response Factors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 610−621.

CI990064X