

Interrelationship of Major Topological Indices Evidenced by Clustering*

Subhash C. Basak,^{a,**} Brian D. Gute,^a and Alexandru T. Balaban^b

^aNatural Resources Research Institute, University of Minnesota Duluth, Duluth, MN 55811, USA

^bTexas A&M University at Galveston, Department of Marine Sciences, 5007 Avenue U, Galveston, TX 77551, USA

RECEIVED FEBRUARY 24, 2003; REVISED JULY 21, 2003; ACCEPTED OCTOBER 20, 2003

This study examines the mutual relatedness of 318 major topological indices (TIs) for three sets of molecules: (i) a set of 139 hydrocarbons, (ii) a diverse set of 1029 compounds and (iii) a diverse set of 2887 compounds. The TIs included in this study are those that have been frequently used in the characterization of structure and QSAR/ QSPR studies. After variable reduction based on the elimination of TIs for which all values were zero and those that were completely correlated with another TI, a variable clustering technique was used to cluster the TIs which resulted in 16, 37 and 56 clusters, respectively, for the three data sets mentioned above. Analysis of the correspondence among the clusters derived from the three groups of chemicals has been carried out in an effort to understand the dimensionality of the structure spaces derived for the three different sets of chemicals and the structural aspects characterized by the various TIs.

Key words
topological indices
cluster analysis
diverse compounds databases

INTRODUCTION

A major trend in mathematical and computational chemistry, drug discovery, predictive toxicology and quantitative structure-activity/property relationship (QSAR/QSPR) studies is the application of topological indices for predicting biomedical, toxicological, physicochemical, and technological properties of chemicals from their structure.^{1–4} Both in drug discovery and in the hazard assessment of environmental chemicals one is faced with a large number of candidate chemicals, the majority of which do not have available property data.^{2–5} Therefore, property-property correlations to estimate complex properties

from simpler experimental properties is not an attractive option in such cases, primarily because experimental properties are not available for the majority of candidate chemicals. The other viable alternative is to estimate necessary properties from parameters that can be calculated directly from molecular structure directly without any input of experimental data. One important class of theoretically derived parameters that are being used more and more frequently in QSAR/QSPR studies are topological indices.

Topological indices are numerical graph invariants derived from different types of weighted molecular graphs. In

* This paper is dedicated to Professor Nenad Trinajstić's 65th birthday with best wishes for continued success in applications of Chemical Graph Theory.

** Author to whom correspondence should be addressed. (E-mail: sbasak@nrri.umn.edu)

the graph theoretical formalism, a molecule is represented by vertices (atoms) and edges (bonds). Mathematically, a graph $G = V, E$ is an ordered pair where the nonempty set V represents the set of atoms and the E symbolizes the set of bonds. A graph invariant is a graph theoretic property that has the same value for isomorphic graphs.^{6–8} A topological index (TI) is a graph invariant that consists of a single numerical value derived from a molecular graph. Therefore, a TI carries out a numerical characterization of molecular topology and is usually sensitive to such chemically important features of molecular structure as size, shape, branching, cyclicality, heterogeneity of atoms or bonds, and neighborhoods of atoms.

In the following, we shall use interchangeably the terms TI, molecular descriptor, parameter, or variable; TIs have found wide application in QSAR/ QSPR studies.^{1,2,9–27} Different groups have developed novel TIs based on various theoretical reasonings.^{28–44} A fairly complete list of older and newer TIs may be found in the introductory chapter of the book cited under Refs. 1 and 6, but among these indices we limited ourselves to those incorporated in three computer programs (POLLY, Triplet, and Molconn-Z). Numerous QSARs/ QSPRs have been developed on mostly congeneric sets of structures with good results.^{40,45–47} Our experience with developing QSARs/ QSPRs on large and heterogeneous data sets indicated that we need a broad range of TIs belonging to the major classes rather than using one TI or one class of TIs at a time in a piecemeal manner.^{9,10} Basak *et al.* calculated large numbers of TIs belonging to different classes for sets of chemicals ranging from small congeneric sets to heterogeneous subsets of the Toxic Substances Control Act (TSCA) inventory consisting of between one to three thousand chemicals,⁴⁷ as well as for a set of over 248 000 psoralen derivatives.⁵ They studied the interrelatedness of such indices in an effort to extract orthogonal information using methods such as principal components analysis (PCA) and variable clustering (VC). One goal of these studies was to use the PCs or minimally correlated TIs derived from VC in QSAR/QSPR studies.^{1,2,9–15} In the 1986–1988 studies, 90 TIs were used by Basak *et al.* for the creation of structure spaces for a diverse subset of 3692 industrial chemicals by means of PCA. Such principal components (PCs) have been used in defining structural similarity and selection of structural analogs of chemicals.^{15–27} Another related use of PCs/TIs has been in the clustering of large sets of chemicals to bring down the size of the problem in chemical design, drug discovery, and predictive toxicology.^{5,9,47}

Other analyses showing how TIs may be grouped together have been described by Motoc and Balaban,⁴⁹ by Randić,⁵⁰ by Todeschini *et al.*,⁵¹ by Bertz,⁵² by Ivanciuc *et al.*,⁵³ and by Balaban *et al.*^{54,55} In the last type of analysis,^{50–55} alkanes with up to nine carbon atoms were

found to be ordered differently by various TIs and this fact allowed related TIs to be grouped together.

A perusal of the above and other pertinent literature shows that TIs are being used in many diverse situations such as lead optimization in drug discovery,⁵⁶ QSAR/ QSPR/ QSTR,^{1–6,9–14,35,36,38,48} analog selection,^{15–25,27,48} molecular similarity-based estimation of properties,^{15,18–27,48} clustering of large sets of chemicals for molecular and pharmaceutical design,^{5,57} or the investigation of relationships between transfer RNAs of bacteria, providing support for the coevolution theory of the genetic code,⁵⁸ to name just a few. Therefore, we need to know the degree of intercorrelation of the various TIs. Although different TIs are derived from different matrices defined on various types of molecular graphs, and are based on diverse theoretical rationales, many of these TIs are strongly correlated. Practical application of TIs requires that we know which of the several hundred indices are least correlated, *i.e.*, which ones encode relatively independent structural information.

With this end in view, in a previous publication⁴⁷ we studied the mutual relatedness of a set consisting initially of 202 TIs calculated for a group of 139 hydrocarbons and a group of 1029 diverse chemicals taken from the TSCA Inventory. From the set of TIs we selected 162 weakly intercorrelated TIs for the former group, and 176 TIs for the latter group of chemicals. Application of the VARCLUS program allowed the analysis and visual presentation of clustering for these two databases.

The set of TIs studied in our previous paper⁴⁷ did not include some important TIs such as the kappa and electrotopological indices.⁵⁶ Therefore, in this paper we have studied the mutual relatedness of an expanded set of 318 TIs calculated for three groups of molecules, two of which are those examined in the previous study: (i) the same relatively homogeneous group of 139 hydrocarbons, (ii) the same group of 1029 diverse chemicals, and (iii) a new, larger and more diverse group of 2887 molecules taken from the US EPA ASTER System.⁵⁰

The major objectives of this paper are: (i) to determine which of the large number of TIs are minimally correlated with each other so that they can constitute the starting subset of indices for QSAR/QSPR/ QSTR studies, analog selection, quantification of structural similarity/dissimilarity, and clustering of large real and virtual libraries of chemicals, and (ii) to analyze the »intrinsic dimensionality« of structure spaces created by TIs for congeneric *versus* structurally diverse and non-congeneric groups of chemicals.

The chemical classes of structures in the database with 2887 diverse chemicals are presented in Table I. The analogous partitions for the databases with 139 hydrocarbons and with 1037 diverse chemicals pub-

TABLE I. Summary of chemical classes or features in databases analyzed

Chemical classes or features	Database (Total number of compounds)		
	A (139) ^(a)	B (1029) ^(b)	C (2887) ^(c)
Hydrocarbons	139	563	574
Alkanes, Cycloalkanes	73	415	425
Aromatics	66	148	149
Alkylbenzenes	29	91	92
Polycyclic aromatics	37	57	57
Non-hydrocarbons	0	466	2313
Substituted aromatics		140	728
Halogenated compounds		236	239
Bromine only		75	76
Chlorine only		91	93
Fluorine only		20	20
Iodine only		25	25
Mixed halides		25	25
Sulfides		54	56
Thiols		19	19
Mixed heteroatoms		17	495
Alcohols			196
Aldehydes			46
Amides			21
Amines			181
Carboxylic acids			33
Esters			179
Ketones			48
Nitriles			44
Nitro derivatives			28

^(a) Hydrocarbons. ^(b) Diverse. ^(c) Diverse.

lished earlier⁴⁷ are repeated here. The selected topological indices with their abbreviations are indicated in Table II.

METHODS

Chemical Databases

There were three sets of chemicals analyzed in this study: a set of 139 hydrocarbons (for which many physical data were available) to represent a moderately homogeneous set of chemicals and a set of 1037 diverse chemicals. The hydrocarbons consisted of 73 C₃–C₉ alkanes, 29 alkylbenzenes, and 37 polycyclic aromatic hydrocarbons.^{60–62} The diverse set of 1037 compounds consists of those chemicals from TSCA and the US EPA ASTER system⁵⁹ for which a measured boiling point was available and for which there was no hydrogen bonding potential (as measured by HB1 = 0). The final set of 2887 compounds comprises the full set of boiling point data from the ASTER system for which topological indices could be calculated using all three programs (POLLY 2.3, Triplet and Molconn-Z 3.51). We consider that by adding not only further topological indices but also a

new set of diverse compounds, the present paper consolidates the conclusions of the preceding paper.⁴⁷ It must be noted here that any compounds composed of three or fewer non-hydrogen atoms were removed from the data set due to the nature of the Triplet index calculations.

Calculation of TIs

The TIs calculated for this study include the Wiener number W ,²⁸ molecular connectivity indices as calculated by Randić³⁰ and Kier and Hall,^{38,39,45} frequency of path lengths of varying size, information theoretic indices defined on distance matrices of graphs using the methods of Bonchev and Trinajstić,³⁷ Roy *et al.*,⁴² Basak *et al.*,^{25,26} as well as those of Raychaudhury *et al.*,⁴¹ parameters defined on the neighborhood complexity of vertices in hydrogen-filled molecular graphs,^{25,26,41} and Balaban's J indices^{31–34} as well as triplet indices.⁶³ Ninety-eight of the TIs were calculated using the program POLLY 2.3.⁶⁴ The J indices and triplet indices were calculated using software developed in-house by the authors and the additional 167 indices were calculated using Molconn-Z 3.51 developed by Hall and Associates Consulting.⁶⁵

TABLE II. Symbols and definitions of topological indices

Topostructural indices	
IDW	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
MIDW	Mean information index for the magnitude of distance
<i>W</i>	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
ID	Degree complexity
HV	Graph vertex complexity
HD	Graph distance complexity
IC _{bar}	Information content of the distance matrix partitioned by frequency of occurrences of distance <i>h</i>
M1	A Zagreb group parameter = sum of square of degree over all vertices
M2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
<i>Sh</i>	Path connectivity index of order <i>h</i> = 0–6
<i>SCh</i>	Cluster connectivity index of order <i>h</i> = 3–6
<i>SCYh</i>	Chain connectivity index of order <i>h</i> = 3–6
<i>SPCh</i>	Path-cluster connectivity index of order <i>h</i> = 4–6
<i>Kh</i>	Number of paths of length <i>h</i> = 0–10
<i>J</i>	Balaban's <i>J</i> index based on distance
Nrings	Number of rings in a graph
Ncirc	Number of circuits in a graph
DN2Sy	Triplet index from distance matrix, square of graph order (# of non-H atoms), and distance sum; operation <i>y</i> = 1–5
DN21y	Triplet index from distance matrix, square of graph order, and number 1; operation <i>y</i> = 1–5
AS1y	Triplet index from adjacency matrix, distance sum, and number 1; operation <i>y</i> = 1–5
DS1y	Triplet index from distance matrix, distance sum, and number 1; operation <i>y</i> = 1–5
ASNy	Triplet index from adjacency matrix, distance sum, and graph order; operation <i>y</i> = 1–5
DSNy	Triplet index from distance matrix, distance sum, and graph order; operation <i>y</i> = 1–5
DN2Ny	Triplet index from distance matrix, square of graph order, and graph order; operation <i>y</i> = 1–5
ANSy	Triplet index from adjacency matrix, graph order, and distance sum; operation <i>y</i> = 1–5
AN1y	Triplet index from adjacency matrix, graph order, and number 1; operation <i>y</i> = 1–5
ANNy	Triplet index from adjacency matrix, graph order, and graph order again; operation <i>y</i> = 1–5
ASVy	Triplet index from adjacency matrix, distance sum, and vertex degree; operation <i>y</i> = 1–5
DSVy	Triplet index from distance matrix, distance sum, and vertex degree; operation <i>y</i> = 1–5
ANVy	Triplet index from adjacency matrix, graph order, and vertex degree; operation <i>y</i> = 1–5
Topochemical indices	
I _{Orb}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
Max _{IC}	Order of neighborhood when IC _r reaches its maximum value for the hydrogen-filled graph
Max _{Orb}	Order of neighborhood when IC _r reaches its maximum value for the hydrogen-suppressed graph
IC _r	Mean information content or complexity of a graph based on the <i>r</i> th (<i>r</i> = 0–6) order neighborhood of vertices in a hydrogen-filled graph
SIC _r	Structural information content for <i>r</i> th (<i>r</i> = 0–6) order neighborhood of vertices in a hydrogen-filled graph
CIC _r	Complementary information content for <i>r</i> th (<i>r</i> = 0–6) order neighborhood of vertices in a hydrogen-filled graph
<i>Bh</i>	Bond path connectivity index of order <i>h</i> = 0–6
<i>BCh</i>	Bond cluster connectivity index of order <i>h</i> = 3–6
<i>BCYh</i>	Bond chain connectivity index of order <i>h</i> = 3–6
<i>BPCh</i>	Bond path-cluster connectivity index of order <i>h</i> = 4–6
<i>Vh</i>	Valence path connectivity index of order <i>h</i> = 0–6
<i>VCh</i>	Valence cluster connectivity index of order <i>h</i> = 3–6
<i>VCYh</i>	Valence chain connectivity index of order <i>h</i> = 3–6
<i>VPCh</i>	Valence path-cluster connectivity index of order <i>h</i> = 4–6
JB	Balaban's <i>J</i> index based on bond types
JX	Balaban's <i>J</i> index based on relative electronegativities
JY	Balaban's <i>J</i> index based on relative covalent radii
AZVy	Triplet index from adjacency matrix, atomic number, and vertex degree; operation <i>y</i> = 1–5
AZSy	Triplet index from adjacency matrix, atomic number, and distance sum; operation <i>y</i> = 1–5

(cont.)

ASZ _y	Triplet index from adjacency matrix, distance sum, and atomic number; operation $y = 1-5$
AZN _y	Triplet index from adjacency matrix, atomic number, and graph order; operation $y = 1-5$
ANZ _y	Triplet index from adjacency matrix, graph order, and atomic number; operation $y = 1-5$
DSZ _y	Triplet index from distance matrix, distance sum, and atomic number; operation $y = 1-5$
DN2Z _y	Triplet index from distance matrix, square of graph order, and atomic number; operation $y = 1-5$
Nvx	Number of non-hydrogen atoms in a molecule
Nelem	Number of elements in a molecule
Fw	Molecular weight
XPh	Valence path connectivity index of order $h = 7-10$
XChh	Valence chain connectivity index of order $h = 7-10$
Si	Shannon information index
Totop	Total Topological Index t
SumI	Sum of the intrinsic state values I
SumdelI	Sum of delta-I values
Tets2	Total topological state index based on electrotopological state indices
Phi _a	Flexibility index ($kp1 * kp2/nvx$)
IDC _{bar}	Bonchev-Trinajstić mean information index
IDC	Bonchev-Trinajstić information index
Wp	Wiener p
Pf	Platt f
Wt	Total Wiener number
Knotp	Difference of chi-cluster-3 and path/cluster-4
Knotpv	Valence difference of chi-cluster-3 and path/cluster-4
Nclass	Number of classes of topologically (symmetry) equivalent graph vertices
NumHBd	Number of hydrogen bond donors
NumHBa	Number of hydrogen bond acceptors
SHCsats	E-State of C sp^3 bonded to other saturated C atoms
SHCsatu	E-State of C sp^3 bonded to unsaturated C atoms
Shvin	E-State of C atoms in the vinyl group, =CH-
Shtvin	E-State of C atoms in the terminal vinyl group, =CH ₂
Shavin	E-State of C atoms in the vinyl group, =CH-, bonded to an aromatic C
Sharom	E-State of C sp^2 which are part of an aromatic system
SHHBd	Hydrogen bond donor index, sum of Hydrogen E-State values for -OH, =NH, -NH ₂ , -NH-, -SH, and #CH
SHWHBd	Weak hydrogen bond donor index, sum of C-H Hydrogen E-State values for hydrogen atoms on a C to which a F and/or Cl are also bonded
SHHBa	Hydrogen bond acceptor index, sum of the E-State values for -OH, =NH, -NH ₂ , -NH-, >N-, -O-, -S-, along with -F and -Cl
Qv	General Polarity descriptor
NHBinty	Count of potential internal hydrogen bonders ($y = 2-10$)
SHBinty	E-State descriptors of potential internal hydrogen bond strength ($y = 2-10$) Electrotopological State index values for atoms types: SHsOH, SHdNH, SHsSH, SHsNH ₂ , SHssNH, SHtCH, Shother, SHCHnX, Hmax Gmax, Hmin, Gmin, Hmaxpos, Hminneg, SsLi, SssBe, SssssBem, SssBH, SssssB, SssssBm, SsCH ₃ , SdCH ₂ , SssCH ₂ , StCH, SdsCH, SaaCH, SssCH, SddC, StsC, SdssC, SaasC, SaaaC, SssssC, SsNH ₃ p, SsNH ₂ , SssNH ₂ p, SdNH, SssNH, SaaNH, StN, SssNHp, SdsN, SaaN, SssN, SddsN, SaasN, SssssNp, SsOH, SdO, SssO, SaaO, SsF, SsSiH ₃ , SssSiH ₂ , SssssSiH, SssssSi, SsPH ₂ , SssPH, SssssP, SdssP, SssssP, SsSH, SdS, SssS, SaaS, SdssS, SddssS, SssssssS, SsCl, SsGeH ₃ , SssGeH ₂ , SssssGeH, SssssGe, SsAsH ₂ , SssAsH, SssssAs, SdssAs, SssssAs, SsSeH, SdSe, SssSe, SaaSe, SdssSe, SddssSe, SsBr, SsSnH ₃ , SssSnH ₂ , SssssSnH, SssssSn, SsI, SsPbH ₃ , SssPbH ₂ , SssssPbH, SssssPb
kp0	Kappa zero
kp1-kp3	Kappa simple indices
ka1-ka3	Kappa alpha indices

Statistical Analysis

The number of TIs calculated *via* the POLLY, Triplet, and Molconn-Z programs, before deleting those TIs that were completely collinear with other indices and those that had zero values for all chemicals in the data set, was 369. After deletions were carried out, the number of remaining descriptors was between 150 and 300, depending on the diversity of the data set.

Once the zero value and redundant indices were removed, the computed TIs were transformed by the natural logarithm of the index plus a constant, generally one. This was done since the scale of some indices may be several orders of magnitude greater than that of other indices.

For each set, a technique known as variable clustering was performed using SAS procedure VARCLUS, which presents the advantage that it combines clustering iteratively with principal component analysis (PCA) techniques.⁶⁶ The variable clustering procedure divides the set of indices into disjoint clusters, such that each cluster is essentially unidimensional. This is accomplished by a repeated principal components analysis of the sets of indices. The initial PCA examines all indices and defines two principal components or eigenvectors. If the eigenvalue for the second component is > 1.0 , the indices are split into separate clusters by correlating the indices with the first and second principal component. Those indices most correlated with the first component form one cluster and those indices most correlated with the second component form another cluster, thus forming two disjoint clusters. A PCA is then performed for each cluster of indices, with the cluster being split if the eigenvalue for the second component is > 1.0 . The procedure is repeated until the second eigenvalue is < 1.0 for all clusters.

RESULTS – COMPARISON OF PRESENT AND PREVIOUS CLUSTERING OF TOPOLOGICAL INDICES

139 Hydrocarbons

Analysis of the set of indices for the 139 hydrocarbons showed that 157 of the calculated indices were completely correlated with another index that was retained in the set or had zero values for all compounds. A total of 12 POLLY indices, mainly associated with strained ring systems (3–4 membered rings) were removed since no strained ring compounds were present in the data set. Of the 150 indices calculated by Triplet, 32 were removed. Many of the MolConn-Z parameters new to this study were removed (113) as a wide variety of atom-types were not represented in our hydrocarbon data set, thus the atom-type indices had zero values for all compounds. This left us with a set of 212 TIs for variable clustering. The present clustering of the 212 TIs including TIs from Molconn-Z for the 139 hydrocarbons-database afforded sixteen clusters, denoted by H1 through H16. For the same database, the previous cluster analysis using 162 TIs⁴⁷ had yielded fourteen clusters, denoted by A1 through A14. There is a close correspondence between the clusters of variables (TIs) found now and those reported in the previous paper,⁴⁷ as seen by examining Figure 1. The clusters are listed in the order of decreasing numbers of TIs in each cluster, and are ordered vertically on the right and left sides of Figure 1. The numbers of TIs in each cluster are written in brackets close to each cluster. Each line in Fig. 1 connects clusters sharing at least one TI in common and the number of the shared TIs is written close to each line. Num-

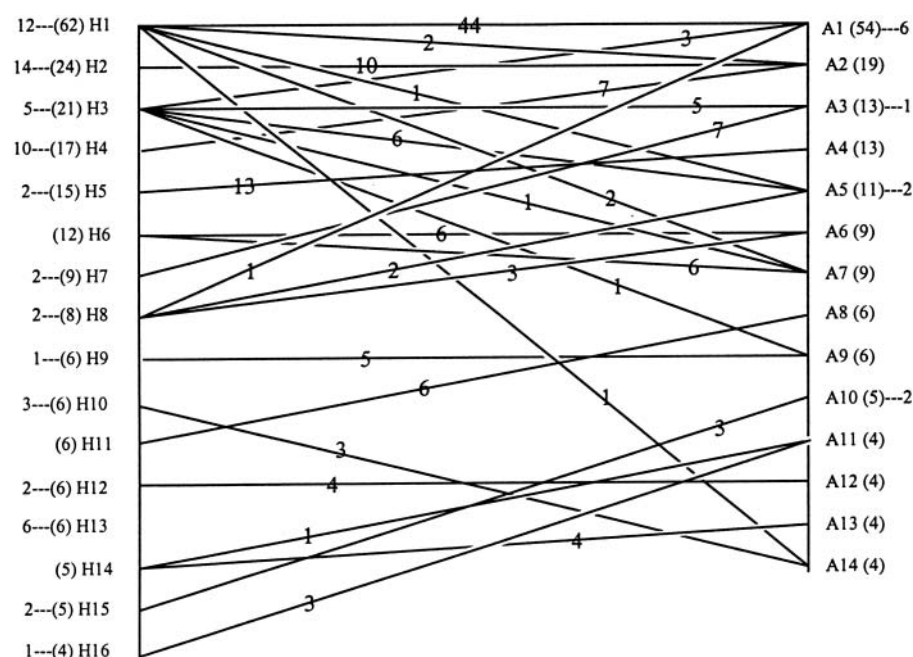


Figure 1. Associations between clusters for the hydrocarbon database using the present set of molecular descriptors (H-type clusters) and the previous set⁴⁷ (A-type clusters). The number of descriptors in each cluster is indicated in brackets. Solid lines connect clusters that have common descriptors, and their numbers are indicated on each line. Dashed lateral lines indicate descriptors that have no correspondence for the other type.

TABLE III. Clustering of 139 hydrocarbons^{(a), (b)}

With Molconn-Z (16 clusters, 212 parameters)						Without Molconn Z (14 clusters, 158 parameters)					
# of TIs in cluster	TI most correlated with cluster	TI least correlated with cluster		Cluster		# of TIs in cluster	TI most correlated with cluster	TI least correlated with cluster		Cluster	
62	DN2Z4	0.9995	B4	0.8975	H1	53	DN2Z4	0.9996	B4	0.9035	A1
24	XP7	0.9912	JB	0.5893	H2	18	S6	0.9884	JB	0.6325	A2
21	ANS2	0.9800	KA1	0.7837	H3	12	V0	0.9390	ANS1	0.0670	A3
17	QV	0.9725	SSSSCH	0.3951	H4	–	–	–	–	–	–
15	SIC6	0.9820	MAXORB	0.6888	H5	13	SIC6	0.9889	MAXORB	0.6789	A4
12	DSZ3	0.9902	SPC5	0.9346	H6	13		0.9866	ANV2	0.8016	A6
9	SUMDELI	0.9337	SPC4	0.8609	H7	9	DSN3	0.9866	ANS2	0.1385	A7
8	DSV1	0.9488	IC_BAR	0.7931	H8	10	DSZ1	0.9881	DSV5	0.9247	A5
6	DSV2	0.9717	DSN2	0.8158	H9	5	DSZ2	0.9778	ASV2	0.8850	A9
6	PHIA	0.9105	KA2	0.6363	H10	–	–	–	–	–	–
6	VC5	0.9456	SC6	0.6750	H11	6	VC5	0.9456	SC6	0.6750	A8
6	SCY5	0.9984	SDSCH	0.0850	H12	3	SCY5	0.9763	AZS1	0.2899	A14
6	BC3	0.9626	SC3	0.6948	H13	4	VC3	0.9860	SC4	0.7994	A12
5	SIC1	0.8910	IC1	0.7815	H14	5	SIC1	0.8365	ASV3	0.6817	A10
5	SIC2	0.9578	CIC2	0.7392	H15	4	SIC3	0.9355	SIC2	0.9092	A13
4	SIC0	0.9706	GMAX	0.8613	H16	4	CIC1	0.9559	CIC2	0.6339	A11

^(a) Published earlier ⁴⁷ without Molconn-Z; the present paper includes Molconn-Z.

^(b) In Figure 1, clusters H1–H16 (here: left column, top to bottom) are linked to clusters A1–A14 (right column) mainly by sharing descriptors when the clusters are on the same horizontal line in the present table.

bers of TIs that are not shared are listed on the outside of the central part of Figure 1.

A more detailed account of how the TIs computed with or without electrotopological state and other parameters computed by the Molconn-Z program for the hydrocarbon database is presented in Table III, which contains also the TIs with the highest and lowest eigenvalues for each cluster. One can see that often these two TIs with highest and lowest eigenvalues in the two classes of clusters (with/without Molconn-Z indices) coincide, especially when the number of TIs in clusters is not large.

Four clusters in the previous and present papers have a unique counterpart: H5 is paired with A4, H11 with A8, H12 with A12, and H15 with A10; however, clusters H5, H12 and H15 in the present analysis, as well as cluster A10 in the previous one, include a few TIs that do not appear in the other analysis.

In both analyses, the first clusters are the most populated, and the 44 TIs that they share represent 81 % of A1's population and 71 % of H1's population of TIs.

One can conclude that for hydrocarbons the addition of Molconn-Z parameters does not change appreciably the clustering of molecular descriptors, *i.e.*, the intrinsic dimensionality of the structure space remains practically the same in spite of increasing the number of TIs from 162 to 212. This was to be expected, because the Molconn-Z parameters contain considerable information about heteroatoms.

In light of the results presented here and the previous study by Basak *et al.* on hydrocarbons,⁴⁷ it is proposed that QSAR/QSPR studies and clustering of hydrocarbons could start with the following sixteen indices most correlated with the sixteen clusters (H1–H16) from this study: DN2Z4, XP7, ANS2, QV, DSZ3, SUMDELI, DSV1, DSV2, PHIA, VC5, VCY5, BC3, SIC0–SIC3, SIC6. Of course, when there are fewer atoms than specified for the subgraph, the indices that do not apply should be disregarded.

The above sixteen indices encode the least correlated and most information-rich subset of the 212 TIs analyzed in this study. Some additional information can be gained by supplementing the above group of sixteen TIs with one or more indices least correlated with the individual clusters.

1029 Diverse Compounds

Data reduction on the set of indices calculated for the 1029 set of TSCA chemicals resulted in the removal of 114 indices. In this instance, all of the POLLY indices were retained for clustering while 30 of the Triplet indices were removed. Far fewer of the atom-type parameters had to be removed (84) as the TSCA set shows a greater diversity of atom types than the hydrocarbon database. The present clustering of the 255 TIs including TIs from Molconn-Z for the 1029 diverse compounds afforded 37 clusters, denoted by N1 through N37, whereas the pre-

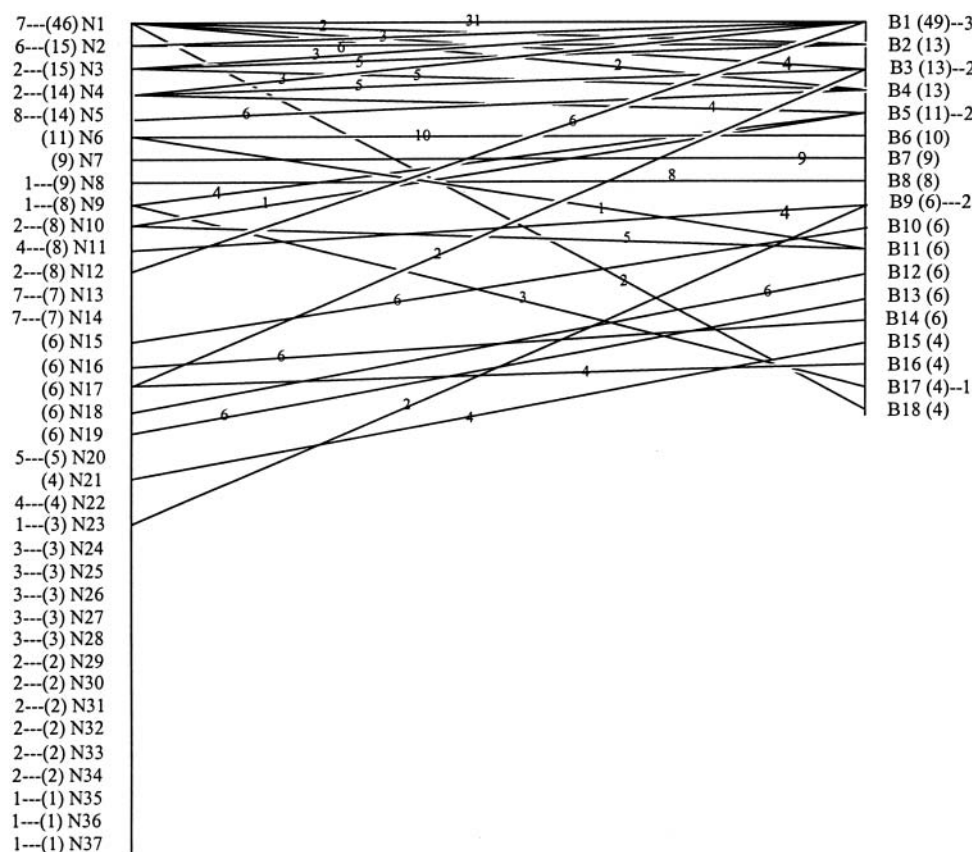


Figure 2. Same as Figure 1, but with associations between clusters for the diverse compound database using the present set of molecular descriptors (N-type clusters) and the previous set⁴⁷ (B-type clusters).

ceding analysis of nearly identical database (1037 compounds) with 176 TIs had yielded only 18 clusters denoted by B1 through B18. The clusters are ordered according to the decreasing numbers of TIs in each cluster. The associations between these two sets of clusters are presented in Figure 2, using the same conventions as in Figure 1. Table IV presents part of the N-type clusters in correspondence with all the B-type clusters, indicating as in Table III the TIs with the highest and lowest eigenvalues.

Again, there is a fair degree of similarity between these two clusterings. Cluster N1 is mainly associated with cluster B1. Some clusters are totally interassociated (N7 with B7, N8 with B8, N15 with B10, N16 with B14, N18 with B12, N19 with B13, and N21 with B15). Other N-type clusters are associated with only one B-type cluster, and have, in addition, several Molconn-Z-type indices (N5 with B4, N8 with B8, N11 with B9, N23 with B9). Many N-type indices yield clusters containing only Molconn-Z indices (N13, N14, N20, N22, and N24 through N38), probably because the corresponding heteroatom types in the compound database are better taken into account by the electrotopological state indices.

On the other hand, one can see that clusters N1 through N4 and N9 have several connections with clusters B1 through B5, so that one can conclude that adding Molconn-Z indices for diverse compounds results in much smaller clusters.

2887 Diverse Compounds

In the previous paper,⁴⁷ only one database with diverse chemicals was analyzed, without Molconn-Z parameters. Now we present in Table V and Figure 3 the comparison for two databases with diverse compounds between clustering of TIs that include Molconn-Z parameters in both cases. The partition of 255 TIs grouped into the same 37 clusters denoted by N1 through N37 for the database with 1029 compounds that was discussed in the preceding section is now compared with the clustering of 293 TIs for a larger and more structurally diverse database of 2887 compounds. As is obvious from the previous discussions, even greater atom-type diversity is evident in this data set. The same deletions were made concerning the POLLY and Triplet indices as in the previous set, all POLLY indices were retained and the same 30 Triplet indices were removed. As a result of the greater diversity of the database, only 46 of the Molconn-Z indices were removed from the set of indices and most of those were indices calculated for non-organic atom types such as silicon, germanium, selenium, arsenic and lead. From this set of 293 indices, we obtained 56 clusters (denoted by T1 through T56) instead of the 37 N-type clusters. This difference is no longer due to the inclusion of new TIs as in the preceding two cases, but to the increased diversity of compounds in the database. The

TABLE IV. Clustering of 1000+ diverse compounds (1029 with Molconn-Z, 1037 without Molconn-Z)^(a), ^(b)

Cluster	With Molconn-Z (37 clusters, 255 parameters)				Without Molconn-Z (18 clusters, 158 parameters)				
	# of TIs in cluster	TI most correlated with cluster	TI least correlated with cluster	Cluster	# of TIs in cluster	TI most correlated with cluster	TI least correlated with cluster		
N1	46	IDW	0.9970	B3	49	K0	0.9966	V2	0.6255
N2	15	DN2S4	0.9935	SUMI	13	ANV1	0.9634	ANV5	0.7927
N3	15	WT	0.9638	SHOTHER	12	S6	0.9406	DN2S3	0.3290
N4	14	HV	0.9807	MAXORB	13	AS11	0.9902	ASV2	0.7852
N5	14	XP8	0.9213	K7	6	IC4	0.9850	IC2	0.7332
N6	11	SIC3	0.9624	IC1	10	SIC3	0.9480	IC1	0.5836
N7	9	BPC5	0.9380	VPC4	9	BPC5	0.9383	VPC4	0.7051
N8	9	ASZ2	0.9722	SSBR	8	ASZ2	0.9687	ANZ1	0.6860
N9	8	DN2N1	0.9513	AZN4	4	DN213	0.9715	AZN4	0.4600
N10	8	IC6	0.9749	MAX_IC	11	ASN5	0.9745	ASV3	0.5682
N11	8	SC6	0.8958	SSF	6	BC5	0.9537	VC6	0.5282
N12	8	V0	0.9702	FW	–	–	–	–	–
N13	7	KA2	0.9045	SHCSATS	–	–	–	–	–
N14	7	SHAROM	0.8615	SAASC	–	–	–	–	–
N15	6	SCY3	0.9030	VCY3	6	SCY3	0.9030	VCY3	0.8314
N16	6	BC3	0.9161	VC4	6	BC3	0.9160	VC4	0.6765
N17	6	AS12	0.9269	ASV5	4	AS12	0.9885	DSV2	0.8905
N18	6	CIC1	0.9708	IC0	6	CIC1	0.9602	IC0	0.7348
N19	6	VCY6	0.8980	SCY6	6	VCY6	0.8981	SCY6	0.7873
N20	5	SHBBA	0.9414	NELEM	–	–	–	–	–
N21	4	JB	0.9797	J	4	JB	0.9810	J	0.7923

^(a) Published earlier⁴⁷ without Molconn-Z; the present paper includes Molconn-Z.^(b) Only the first 21 clusters are listed here (N1–N21, left column, top to bottom); they are linked to B1–B17 (right column) *via* multiple connections, the most significant one being on the same horizontal line.

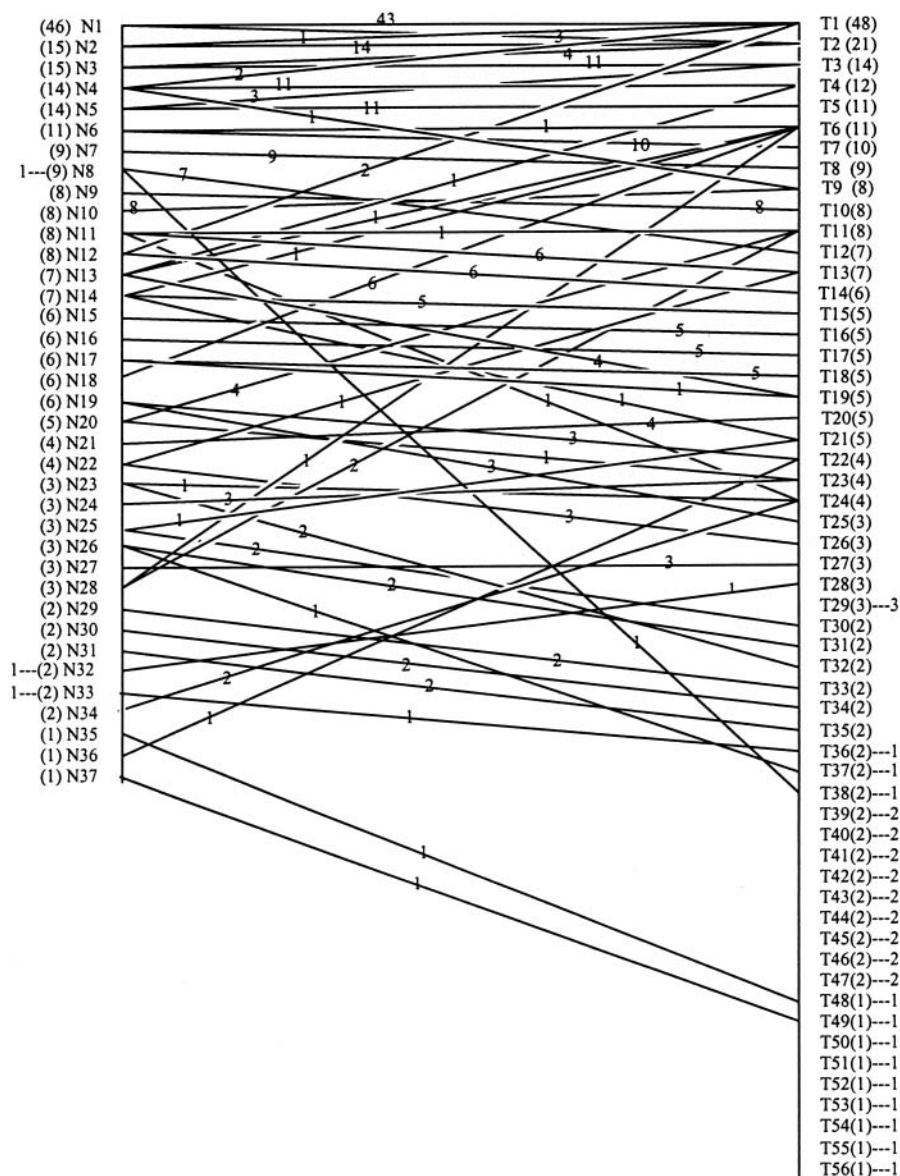


Figure 3. Associations between clusters for the two diverse compound databases having 1029 diverse chemicals (N-type clusters) and 2887 diverse chemicals (T-type clusters). In both cases, the present set of molecular descriptors has been used. The remaining explanations are as in Figure 1.

results are presented in Figure 3 for all clusters and in Table V for a limited, overlapping set of clusters.

Again, as in the preceding cases, the first cluster in each class is the most numerous, and again, as seen in Figure 3, the upper horizontal line indicates that these two clusters share most of their TIs. Addition of new compounds in the database increases the number of T-type clusters with only one or two parameters that are not shared by N-type parameters, apparently because such compounds are uniquely associated with Molconn-Z descriptors.

Table V demonstrates that the two databases having 1029 and 2887 diverse chemicals give rise to very similar clusters: 15 out of the first 23 clusters have exactly the same TIs that are most correlated with their own clusters; the 3 other clusters have closely related TIs that are most correlated with their own cluster (CIC1 and SIC1, SCY3 and SCY4, VCY6 and SCY5).

Concerning the third cluster (N3) in Table V, one should note (Appendix 4) that the molecular descriptor S5 in this cluster (the same TI S5 that is most correlated with its own cluster T3) has $R^2 = 0.9299$, a value that is not much lower than the maximal one ($R^2 = 0.9638$) for index WT. Thus, the third clusters N3 and T3 can also be considered to share S5 as a central TI, and therefore the overwhelming majority of clusters (for all the major clusters) have the same »central« molecular descriptors.

In the light of the above findings for structurally diverse compounds, it follows that one could start with the following 18 indices whenever attempting QSAR or OSPR studies for structurally diverse chemicals: IDW, DN2S4, S5, HV, XP8, SIC1, SIC3, BPC5, IC6, SHHBA, DN2N1, ASZ2, SC6, BC3, SCY4, V0, JB, AS12.

It is interesting to see that only a few of these indices (DN2S4, BC3, SCY5, SIC1) coincide with those selected for hydrocarbons. A few other indices selected for

TABLE V. Clustering of 1029 and 2887 diverse compounds (with Molconn-Z) (a), (b)

1029 compounds (37 clusters, 255 parameters) Fig. 3, columns N1–N37				2887 compounds (56 clusters, 293 parameters) Fig. 3, columns T1–T56			
Cluster	# of TIs in cluster	TI most correlated with cluster	TI least correlated with cluster	Cluster	# of TIs in cluster	TI most correlated with cluster	TI least correlated with cluster
N1	46	IDW	B3	T1	48	IDW	SUMI
N2	15	DN2S4	SUMI	T2	21	DN2S4	TETS2
N3	15	WT	SHOTHER	T3	14	S5	SHOTHER
N4	14	HV	MAXORB	T4	12	HV	IC_BAR
N5	14	XP8	K7	T5	11	XP8	XVP10
N18	6	CIC1	IC0	T6	11	SIC1	HMIN
N6	11	SIC3	IC1	T7	10	SIC3	IC2
N7	9	BPC5	VPC4	T8	9	BPC5	VPC4
N10	8	IC6	MAX_IC	T9	9	IC6	MAXORB
N20	5	SHBBA	NELEM	T10	8	SHBBA	SSSSCH
N9	8	DN2N1	AZN4	T11	8	DN2N1	AZN4
N8	9	ASZ2	SSBR	T12	7	ASZ2	ANZ2
N11	8	SC6	SSF	T13	7	SC6	SSF
N16	6	BC3	VC4	T14	6	BC3	VC4
N15	6	SCY3	VCY3	T15	6	SCY4	VCY4
N12	8	V0	FW	T16	6	V0	FW
–	–	–	–	T17	5	SHHBD	HMAX
N21	4	JB	J	T18	5	JB	J
N14	7	SHAROM	SAASC	T19	5	SAACH	SAASC
N17	6	AS12	ASV5	T20	5	AS12	DSV2
N13	7	KA2	SHCSATS	T21	5	KP3	ASV5
N19	6	VCY6	SCY6	T22	4	SCY5	SAAS
–	–	–	–	T23	3	BCY6	SCY6

(a) Only the first 23 clusters are listed, ordered according to the latter database results (clusters T1–T23, right-hand column, top to bottom).

(b) In Figure 3, clusters T1–T56 are linked to clusters N1–N37 via multiple connections, the most significant ones being those that are on the same horizontal line in the present table.

hydrocarbons are related to, but do not coincide with selected indices for diverse compounds: XP7 and QV for hydrocarbons are related to XP8 and HV, respectively, for diverse chemicals.

CONCLUSIONS

We have presented a variable cluster analysis of topological indices (including also the Kier-Hall indices available in the Molconn-Z program) for three databases: 139 hydrocarbons, 1029 diverse compounds, and 2887 diverse compounds, resulting in clusters denoted as H1–H16, N1–N37, and T1–T56, respectively. In a preceding paper,⁴⁷ the first two databases were analyzed similarly, but without the Molconn-Z indices, affording clusters A1–A14 and B1–B18. In the first studies of topological index intercorrelation⁴⁹ and clustering,⁶⁷ only a small number of TIs were analyzed, and only hydrocarbons had been taken into account.

We have also presented visual comparisons of connections between clusters (such that clusters sharing the same descriptors become connected by lines indicating how many descriptors are shared) originating with the same databases, but with descriptors augmented by Molconn-Z indices (clusters A with H, and B with N) or between clusters with the same set of descriptors, but with databases of 1029 and 2887 diverse compounds (clusters N with T).

The usefulness of the present data will consist in having at hand (in this text, figures, and in the Supplementary Material) a rich source of data on how various topological indices become associated in clusters when applied to homogeneous or heterogeneous databases.

It is evident that the inclusion of Molconn-Z indices (with specific descriptors for various types of heteroatoms and multiple bonding) practically doubles the number of clusters for the database with 1029 compounds (from 18 to 37), but does not substantially increase the number of clusters for hydrocarbons (from 14 to 16). Also, an increase of the number of compounds in databases with diverse compounds (from 1029 compounds to 2887 compounds, with the same set of descriptors but including also Molconn-Z descriptors) results in a marked increase of the number of clusters (from 37 to 56). These increased numbers are due to small clusters having only 2–4 descriptors.

More details about the clusters in each of the three databases can be found in the Supplementary Material (Appendix 1 through 6)

Supplementary Materials. – For each of the three data bases, one table presents the numbers of descriptors in each cluster, and for each cluster, the variation explained, the proportion explained, and the second eigenvalue; another table indicates all descriptors in each cluster and their correlation factors with the own and next closest cluster, as well as the $(1 - R^2)$ ratio.

These data are available *via* the Web under <http://pubwww.srce.hr/ccacaa> or may be obtained from the author.

REFERENCES

1. S. C. Basak, B. D. Gute, and G. D. Grunwald, in: J. Devillers and A. T. Balaban (Eds.), *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon & Breach Science Publishers, The Netherlands, 1999, pp. 675–696.
2. S. C. Basak, G. D. Grunwald, and G. J. Niemi, in: A. T. Balaban (Ed.), *From Chemical Topology to Three Dimensional Molecular Geometry*; Plenum Press, New York, 1997, pp. 73–116.
3. A. R. Katritzky, R. Petrukhin, D. Tatham, S. C. Basak, E. Benfenati, M. Karelson, and U. Maran, *J. Chem. Inf. Comput. Sci.* **41** (2001) 679–685.
4. A. R. Katritzky, U. Maran, V. S. Lobanov, and M. Karelson, *J. Chem. Inf. Comput. Sci.* **40** (2000) 1–8.
5. S. C. Basak, D. Mills, B. D. Gute, A. T. Balaban, S. C. Basak, and G. D. Grunwald, in: D. K. Sinha, S. C. Basak, R. K. Mohanty, and I. N. Basumallick (Eds.), *Some Aspects of Mathematical Chemistry*; Visva-Bharati University: Santiniketan, West Bengal, India, in press.
6. A. T. Balaban and O. Ivanciuc, *Topological Indices and Related Descriptors*, in: J. Devillers and A. T. Balaban (Eds.), *QSAR and QSPR*, Gordon & Breach Science Publishers, The Netherlands, 1999, pp. 21–57.
7. F. Harary, *Graph Theory*, Addison Wesley Publ., Reading, Massachusetts, 1969.
8. N. Trinajstić, *Chemical Graph Theory*, 2nd ed., CRC Press, Boca Raton, FL, 1992.
9. S. C. Basak, G. D. Grunwald, B. D. Gute, K. Balasubramanian, and D. Opitz, *J. Chem. Inf. Comput. Sci.*, **40** (2000) 885–890.
10. S. C. Basak, B. D. Gute, and G. D. Grunwald, in: F. Chen and G. Schuurman (Eds.), *Quantitative Structure-Activity Relationships in Environmental Sciences*, Vol. 7, SETAC Press, Pensacola, FL, 1997, Chapter 17, pp. 245–261.
11. S. C. Basak, B. D. Gute, G. D. Grunwald, D. W. Opitz, and K. Balasubramanian, in: *Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools – Papers from the 1999 AAAI Symposium*; AAAI Press: Menlo Park, CA, 1999, pp. 108–111.
12. S. C. Basak, D. Mills, A. T. Balaban, and B. D. Gute, *J. Chem. Inf. Comput. Sci.* **41** (2001) 671–678.
13. B. D. Gute and S. C. Basak, *SAR QSAR Environ. Res.* **7** (1997) 117–131.
14. B. D. Gute, G. D. Grunwald, and S. C. Basak, *SAR QSAR Environ. Res.* **10** (1999) 1–15.
15. S. C. Basak, S. Bertelsen, and G. D. Grunwald, *J. Chem. Inf. Comput. Sci.* **34** (1994) 270–276.
16. S. C. Basak, S. Bertelsen, and G. D. Grunwald, *Toxicol. Lett.* **79** (1995) 239–250.
17. S. C. Basak and G. D. Grunwald, *Math. Model. Sci. Comput.* **4** (1994) 464–469.
18. S. C. Basak and G. D. Grunwald, *New J. Chem.* **19** (1995) 231–237.
19. S. C. Basak and G. D. Grunwald, *Chemosphere* **31** (1995) 2529–2546.
20. S. C. Basak, G. D. Grunwald, G. E. Host, G. J. Niemi, and S. P. Bradbury, *Environ. Toxicol. Chem.* **17** (1998) 1056–1064.
21. S. C. Basak and B. D. Gute, in: B. L. Johnson, C. Xintaras, and J. S. Andrews, Jr. (Eds.), *Proceedings of the International Congress on Hazardous Waste: Impact on Human and*

- Ecological Health*; Princeton Scientific Publishing Co., Princeton, NJ, 1997 pp. 492–504.
22. S. C. Basak, B. D. Gute, and G. D. Grunwald, in: R. Carbo-Dorca and P. G. Mezey (Eds.), *Advances in Molecular Similarity*, Vol. 2; JAI Press: Stanford, Connecticut, 1998 pp. 171–185.
 23. S. C. Basak, B. D. Gute, and G. D. Grunwald, *SAR QSAR Environ. Res.* **10** (1999) 117–129.
 24. S. C. Basak, B. D. Gute, and G. D. Grunwald, in: P. Hansen, P. Fowler, and M. Zheng (Eds.), *Discrete Mathematical Chemistry*, DIMACS Series 51, American Mathematical Society, Providence, Rhode Island, 2000 pp. 9–24.
 25. S. C. Basak, V. R. Magnuson, G. J. Niemi, and R. R. Regal, *Discrete Appl. Math.* **19** (1988) 17–44.
 26. B. D. Gute, G. D. Grunwald, D. Mills, and S. C. Basak, *SAR QSAR Environ. Res.* **11** (2001) 363–382.
 27. B. D. Gute and S. C. Basak, *J. Mol. Graphics Modell.* **20** (2001) 95–109.
 28. H. Wiener, *J. Am. Chem. Soc.* **69** (1947) 17–20.
 29. H. Hosoya, *Bull. Chem. Soc. Jpn.* **44** (1971) 2332–2339.
 30. M. Randić, *J. Am. Chem. Soc.* **97** (1975) 6609–6615.
 31. A. T. Balaban, *Chem. Phys. Lett.* **89** (1982) 399–404.
 32. A. T. Balaban, *Pure Appl. Chem.* **55** (1983) 199–206.
 33. A. T. Balaban, *Math. Chem. (MATCH)* **21** (1986) 115–122.
 34. O. Ivanciuc, T. Ivanciuc, and A. T. Balaban, *J. Chem. Inf. Comput. Sci.* **38** (1998) 395–401.
 35. S. C. Basak, A. B. Roy, and J. J. Ghosh, in: X. J. R. Avula, R. Bellman, Y. L. Luke, and A. K. Rigler (Eds.), *Proceedings of the 2nd International Conference on Mathematical Modelling*, University of Missouri-Rolla, Rolla, Missouri, Vol. 2, 1980, pp. 851–856.
 36. S. C. Basak and V. R. Magnuson, *Arzneim. Forsch.* **33** (1983) 501–503.
 37. D. Bonchev and N. Trinajstić, *J. Chem. Phys.* **67** (1977) 4517–4533.
 38. L. B. Kier, W. J. Murray, M. Randić, and L. H. Hall, *J. Pharm. Sci.* **65** (1975) 1226–1230.
 39. L. B. Kier and L. H. Hall, *Molecular Connectivity in Structure-Activity Analysis*, Research Studies Press, Letchworth, Hertfordshire, U.K., 1986.
 40. N. Rashevsky, *Bull. Math. Biophys.* **17** (1955) 229–235.
 41. C. Raychaudhury, S. K. Ray, J. J. Ghosh, A. B. Roy, and S. C. Basak, *J. Comput. Chem.* **5** (1984) 581–588.
 42. A. B. Roy, S. C. Basak, D. K. Harriss, and V. R. Magnuson, in: X. J. R. Avula, R. E. Kalman, A. I. Lipais, and E. I. Rodin (Eds.), *Mathematical Modelling in Science and Technology*, Pergamon Press, New York, 1984, pp. 745–750.
 43. R. Sarkar, A. B. Roy, and R. K. Sarkar, *Math. Biosci.* **39** (1978) 299–312.
 44. C. E. Shannon, *Bell Syst. Tech. J.* **27** (1948) 379–423.
 45. L. B. Kier and L. H. Hall, *Molecular Structure Description: The Electrotological State*, Academic Press, San Diego, CA, 1999.
 46. M. Randić, X. Guo, and S. Bobst, in: P. Hansen, P. Fowler, and M. Zheng, (Eds.), *Discrete Mathematical Chemistry*, DIMACS Series 51, American Mathematical Society, Providence, Rhode Island, 2000, pp. 305–322.
 47. S. C. Basak, A. T. Balaban, G. D. Grunwald, and B. D. Gute, *J. Chem. Inf. Comput. Sci.* **40** (2000) 891–898.
 48. S. C. Basak, D. Mills, B. D. Gute, G. D. Grunwald, and A. T. Balaban, in: D. H. Rouvray and R. B. King (Eds.), *Topology in Chemistry: Discrete Mathematics of Molecules*, Horwood Publ., Chichester, 2002, pp. 113–184.
 49. I. Motoc and A. T. Balaban, *Rev. Roum. Chim.* **26** (1981) 593–600.
 50. M. Randić, *J. Math. Chem.* **24** (1998) 345–358.
 51. R. Todeschini, R. Cazar, and E. Collina, *Chemom. Intell. Lab. Syst.* **15** (1992) 51–59.
 52. S. H. Bertz, *Discrete Appl. Math.* **19** (1988) 65–83.
 53. O. Ivanciuc, T. Ivanciuc, D. Cabrol-Bass, and A. T. Balaban, *Commun. Math. Chem. (MATCH)* **42** (2000) 155–180.
 54. A. T. Balaban, D. Mills, and S. C. Basak, *Commun. Math. Chem. (MATCH)* **45** (2002) 5–26.
 55. A. T. Balaban, in: D. H. Rouvray and R. B. King (Eds.), *Topology in Chemistry: Discrete Mathematics of Molecules*, Horwood Publ., Chichester, 2002, pp. 89–112.
 56. G. Grassy, B. Calas, A. Yasri, R. Lahana, J. Woo, S. Iyer, M. Kaczorek, R. Floc'h, and R. Buelow, *Nature Biotechnol.* **16** (1998) 748–752.
 57. M. Lajiness, in: D. H. Rouvray (Ed.), *Computational Chemical Graph Theory*, Nova, New York, 1990, pp. 299–316.
 58. C. Bermudez, E. E. Daza, and E. Andrade, *J. Theor. Biol.* **197** (1999) 193–205.
 59. C. L. Russom, E. B. Anderson, B. E. Greenwood, and A. Pilli, *Sci. Total Environ.* 109/110 (1991) 667–670.
 60. D. E. Needham, I. C. Wei, and P. G. Seybold, *J. Am. Chem. Soc.* **110** (1998) 4186–4194.
 61. O. Mekenyan, D. Bonchev, and N. Trinajstić, *Int. J. Quantum Chem.* **18** (1980) 369–380.
 62. W. Karcher, *Spectral Atlas of Polycyclic Aromatic Hydrocarbons*, Vol. 2. Kluwer, Dordrecht, 1988, pp. 16–19.
 63. P. A. Filip, T. S. Balaban, and A. T. Balaban, *J. Math. Chem.* **1** (1987) 61–83.
 64. S. C. Basak, D. K. Harriss, and V. R. Magnuson, POLLY 2.3, copyright of the University of Minnesota, 1988.
 65. Molconn-Z v. 3.50, Hall Associates Consulting, Quincy, MA, 2000.
 66. The VARCLUS Procedure, in SAS/STAT® User's Guide, Version 6, 4th edn. Vol. 2, Cary, NC. SAS Institute Inc., 1989.
 67. I. Motoc, A. T. Balaban, O. Mekenyan, and D. Bonchev, *Math. Chem. (MATCH)*, **13** (1982) 369–404.

SAŽETAK**Međuodnos glavnih topologijskih indeksa pokazan pomoću okupljanja u grozdove****Subhash C. Basak, Brian D. Gute i Alexandru T. Balaban**

U članku se razmatra međuodnos 318 najčešće rabljenih topologijskih indeksa (TI) za tri skupine molekula: (i) 139 ugljikovodika, (ii) 1029 različitih molekula i (iii) 2887 različitih molekula. Nakon uklanjanja onih TI za koje su sve vrijednosti neke molekule jednake nuli i onih TI koji su potpuno korelirani s nekim drugim TI, metoda koja se temelji na okupljanju u grozdove primijenjena je na preostale TI. Dobiveni su grozdovi od 16, 37 i 56 TI za tri skupine razmatranih molekula. Analiziran je odnos među trima grozdovima s ciljem razumijevanja strukturnih karakteristika različitih TI.

Interrelationship of Major Topological Indices Evidenced by Clustering

Subhash C. Basak, Brian D. Gute,
and Alexandru T. Balaban

SUPPLEMENT

Appendix 1. Clustering of the 139 hydrocarbons

Cluster	Members	Variation explained	Proportion explained	Second eigenvalue
H1	62	60.5485	0.9766	0.5007
H2	24	22.0955	0.9206	0.7376
H3	21	19.6624	0.9363	0.5491
H4	17	15.0351	0.8844	0.8544
H5	15	13.6322	0.9088	0.5744
H6	12	11.6814	0.9734	0.1297
H7	9	7.6352	0.8484	0.5892
H8	8	7.1152	0.8894	0.3401
H9	6	5.0149	0.8358	0.6081
H10	6	4.7695	0.7949	0.6531
H11	6	5.0076	0.8346	0.7395
H12	6	4.9728	0.8288	0.9371
H13	6	5.0731	0.8455	0.6305
H14	5	4.1217	0.8243	0.5632
H15	5	4.3705	0.8741	0.4453

Appendix 2. Clustering of the 139 hydrocarbons

Cluster	T. I.	R^2 with own cluster	R^2 with next closest	$1-R^2$ ratio
H1	IDW	0.9831	0.9646	0.4796
	MIDW	0.9504	0.9421	0.8571
	W	0.9863	0.9598	0.3413
	ID	0.9821	0.9291	0.2533
	HD	0.9559	0.9404	0.7407
	M1	0.9838	0.9151	0.1906
	M2	0.9773	0.9227	0.2932
	S0	0.9679	0.9551	0.7125
	S1	0.9936	0.8992	0.0634
	S2	0.9392	0.9055	0.6433
	S3	0.9495	0.8936	0.4741
	S4	0.9365	0.8901	0.5777
	B4	0.8975	0.8501	0.6841
	V4	0.9165	0.8538	0.5712
	K0	0.9979	0.9253	0.0282
	K1	0.9971	0.8833	0.0261
	K2	0.9651	0.9355	0.5431
	K3	0.9486	0.8837	0.4415
	K4	0.9176	0.8595	0.5865
	AZV1	0.9881	0.9052	0.1254
	AZV2	0.9631	0.9185	0.4531
	AZV3	0.9928	0.8821	0.0609
	AZV4	0.9985	0.9166	0.0184
	AZV5	0.9572	0.9076	0.4627
	AZS1	0.9846	0.9591	0.3764
	AZS2	0.9751	0.9691	0.8081
	ASZ4	0.9954	0.9188	0.0565
	DN2S3	0.9853	0.9545	0.3222
	DN2S4	0.9854	0.9365	0.2293
	DN2Z4	0.9995	0.9181	0.0067
	DSZ4	0.9878	0.9319	0.1789
	ASN3	0.9908	0.9219	0.1185
	ASN4	0.9908	0.9171	0.1112
DSN3	0.9813	0.9401	0.3117	
DSN4	0.9741	0.9295	0.3689	
DN2N3	0.9975	0.9048	0.0261	
DN2N4	0.9992	0.9131	0.0095	
ANS1	0.9731	0.9715	0.9478	
ANV1	0.9425	0.8948	0.5465	
ANV3	0.9887	0.8828	0.0966	
ANV4	0.9976	0.9085	0.0265	
AZN1	0.9952	0.9247	0.0639	
AZN2	0.9921	0.9193	0.0995	
AZN3	0.9966	0.9261	0.0463	
AZN5	0.9958	0.9235	0.0555	
ANZ2	0.9887	0.9141	0.1311	
ANZ3	0.9951	0.9401	0.0811	
ANZ4	0.9957	0.8923	0.0404	
ANN1	0.9952	0.9328	0.0708	
ANN2	0.9895	0.9314	0.1529	

	ANN3	0.9968	0.9305	0.0455		KP0	0.8618	0.8366	0.8455
	ANN4	0.9855	0.9127	0.1666		KP1	0.9608	0.8128	0.2096
	ANN5	0.9958	0.9371	0.0662		KA1	0.7837	0.7783	0.9753
	NVX	0.9979	0.9253	0.0282	H4	S5	0.9516	0.9095	0.5348
	FW	0.9854	0.9605	0.3705		SCY6	0.9155	0.8719	0.6594
	TOTOP	0.9628	0.9133	0.4296		BCY6	0.8536	0.7932	0.7081
	SUMI	0.9952	0.8941	0.0457		VCY6	0.8712	0.8462	0.8381
	TETS2	0.9451	0.9414	0.9366		K5	0.9306	0.8223	0.3908
	IDC	0.9801	0.9637	0.5506		K6	0.9531	0.8654	0.3488
	WP	0.9486	0.8837	0.4415		J	0.8641	0.7145	0.4764
	PF	0.9613	0.9278	0.5357		ANV5	0.8671	0.8508	0.8906
	WT	0.9566	0.9504	0.8743		NRINGS	0.9842	0.8871	0.1401
H2	S6	0.9688	0.9146	0.3655		SHOTHER	0.9375	0.9033	0.6467
	B5	0.8811	0.8703	0.9172		HMAX	0.9009	0.6951	0.3251
	B6	0.9525	0.8608	0.3414		SSCH3	0.8946	0.8131	0.5638
	V5	0.9089	0.8815	0.7688		SAACH	0.9512	0.7141	0.1706
	V6	0.9694	0.8658	0.2279		SSSSCH	0.3951	0.2882	0.8498
	K7	0.9377	0.9215	0.7928		SHCSATS	0.8215	0.5924	0.4381
	K8	0.9614	0.8525	0.2621		SHAROM	0.9709	0.7508	0.1168
	K9	0.9695	0.8243	0.1736		QV	0.9725	0.7261	0.1006
	K10	0.9727	0.8209	0.1524	H5	MAX_IC	0.8174	0.4361	0.3237
	JB	0.5893	0.5603	0.9342		I_ORB	0.9431	0.5061	0.1155
	NCIRC	0.9726	0.9114	0.3097		MAX_ORB	0.6888	0.4616	0.5781
	XP7	0.9912	0.8549	0.0605		IC3	0.8695	0.6734	0.3997
	XP8	0.9903	0.8322	0.0576		IC4	0.9589	0.5423	0.0899
	XP9	0.9857	0.8164	0.0777		IC5	0.9729	0.5156	0.0559
	XP10	0.9671	0.7901	0.1569		IC6	0.9759	0.5179	0.0501
	XVP7	0.9904	0.8388	0.0595		SIC4	0.9508	0.7057	0.1672
	XVP8	0.9818	0.8092	0.0956		SIC5	0.9767	0.6489	0.0664
	XVP9	0.9591	0.7755	0.1823		SIC6	0.9821	0.6347	0.0493
	XVP10	0.9179	0.7292	0.3031		CIC4	0.8603	0.7025	0.4698
	XCH10	0.8618	0.6867	0.4411		CIC5	0.8821	0.5792	0.2803
	XVCH10	0.8434	0.6646	0.4668		CIC6	0.8725	0.5346	0.2739
	HMIN	0.8761	0.8611	0.8915		SI	0.9571	0.5377	0.0929
	SAAAC	0.9703	0.8207	0.1657		NCLASS	0.9245	0.5401	0.1641
	KNOTP	0.6766	0.6016	0.8118	H6	SPC5	0.9346	0.8855	0.5713
H3	B0	0.9578	0.8643	0.3108		SPC6	0.9501	0.9099	0.5554
	B1	0.9779	0.9091	0.2431		ASZ3	0.9883	0.9223	0.1509
	B3	0.8621	0.7997	0.6887		DN2Z3	0.9848	0.7849	0.0708
	V0	0.9584	0.8717	0.3246		DSZ3	0.9902	0.8135	0.0523
	V1	0.9732	0.9188	0.3304		ASN1	0.9876	0.9264	0.1691
	V3	0.8772	0.8366	0.7519		ASN5	0.9901	0.9199	0.1231
	ASZ1	0.9501	0.8581	0.3527		DSN1	0.9862	0.8302	0.0811
	ASZ2	0.9223	0.8131	0.4157		DSN5	0.9811	0.8022	0.0955
	ASZ5	0.9491	0.8651	0.3783		DN2N1	0.9865	0.7902	0.0643
	DN2S1	0.9691	0.9319	0.4531		DN2N5	0.9853	0.7862	0.0689
	DN2S5	0.9671	0.9387	0.5393		AZN4	0.9167	0.8459	0.5401
	DN2Z1	0.9608	0.9429	0.6873	H7	SPC4	0.8609	0.7568	0.5721
	DSZ1	0.9491	0.8491	0.3368		B2	0.8751	0.7441	0.4881
	DSZ5	0.9291	0.8811	0.5965		BPC5	0.8767	0.5521	0.2753
	DN2N2	0.9483	0.8799	0.4307		BPC6	0.9082	0.8114	0.4865
	ANS2	0.9801	0.9462	0.3714		V2	0.8879	0.7726	0.4927
	ANZ1	0.9469	0.8633	0.3882		VPC5	0.9227	0.6334	0.2109
	ANZ5	0.9782	0.8714	0.1698		VPC6	0.9062	0.8461	0.6097

	SUMDELI	0.9337	0.7605	0.2767
	KNOTPV	0.4638	0.3459	0.8196
H8	HV	0.9257	0.7835	0.3431
	IC_BAR	0.7931	0.5385	0.4483
	ASV1	0.8618	0.5535	0.3096
	DSV1	0.9488	0.7888	0.2426
	DN2S2	0.9244	0.5833	0.1815
	ANV2	0.8434	0.7229	0.5653
	KP2	0.8665	0.4387	0.2377
	IDCBAR	0.9515	0.8187	0.2672
H9	ASV2	0.9631	0.6777	0.1146
	ASV5	0.5047	0.2721	0.6803
	DSV2	0.9717	0.6121	0.0729
	DN2Z2	0.8331	0.5521	0.3726
	DSZ2	0.9264	0.7426	0.2859
	DSN2	0.8158	0.7742	0.8155
H10	ASN2	0.8366	0.3421	0.2484
	KP3	0.8216	0.4045	0.2996
	KA2	0.6363	0.5836	0.8736
	KA3	0.8221	0.5121	0.3646
	PHIA	0.9105	0.3988	0.1488
	SSSCH2	0.7423	0.4974	0.5127
H11	SC5	0.8914	0.4314	0.1909
	SC6	0.6751	0.2501	0.4333
	BC5	0.9308	0.2369	0.0907
	BPC4	0.7906	0.5506	0.4659
	VC5	0.9456	0.2296	0.0706
	VPC4	0.7741	0.6328	0.6153
H12	SCY5	0.9984	0.2748	0.0022
	BCY5	0.9738	0.2623	0.0356
	VCY5	0.9809	0.2647	0.0261
	XCH9	0.9587	0.2771	0.0571
	XVCH9	0.9761	0.2771	0.0331
	SDSCH	0.0851	0.0232	0.9368
H13	SC3	0.6948	0.5242	0.6416
	SC4	0.8785	0.1885	0.1497
	BC3	0.9626	0.3331	0.0561
	VC3	0.9498	0.3496	0.0772
	GMIN	0.7381	0.5117	0.5366
	SSSSSC	0.8493	0.2197	0.1931
H14	IC0	0.7909	0.5794	0.4971
	IC1	0.7815	0.5524	0.4881
	SIC1	0.891	0.4721	0.2065
	SAASC	0.8541	0.3128	0.2125
	SHCSATU	0.8043	0.2765	0.2705
H15	IC2	0.8793	0.6289	0.3252
	SIC2	0.9578	0.4559	0.0775
	SIC3	0.8731	0.7961	0.6224
	CIC2	0.7392	0.4549	0.4784
	CIC3	0.9211	0.6204	0.2081
H16	SIC0	0.9706	0.3233	0.0434
	CIC0	0.8859	0.6496	0.3257
	CIC1	0.8734	0.4854	0.2459
	GMAX	0.8613	0.2571	0.1867

Appendix 3. Clustering of 1029 diverse compounds

Cluster	Members	Variation explained	Proportion explained	Second eigenvalue
N1	46	44.2065	0.9611	0.5858
N2	15	13.9663	0.9311	0.4987
N3	15	13.0423	0.8695	0.5795
N4	14	12.5807	0.8986	0.4083
N5	14	11.8714	0.8481	0.8892
N6	11	9.2783	0.8435	0.7222
N7	9	7.7103	0.8567	0.6288
N8	9	7.4488	0.8276	0.9035
N9	8	6.8944	0.8618	0.6881
N10	8	7.0281	0.8785	0.6847
N11	8	5.5782	0.6973	0.9961
N12	8	6.7319	0.8415	0.4523
N13	7	5.6734	0.8105	0.7507
N14	7	4.3811	0.6259	0.9746
N15	6	5.2692	0.8782	0.6882
N16	6	4.9684	0.8281	0.5074
N17	6	4.8336	0.8056	0.8432
N18	6	5.2931	0.8822	0.4115
N19	6	5.2584	0.8764	0.6231
N20	5	3.8355	0.7671	0.5241
N21	4	3.6853	0.9213	0.2889
N22	4	2.4384	0.6096	0.9888
N23	3	2.1977	0.7326	0.7168
N24	3	2.9701	0.9901	0.0268
N25	3	2.9955	0.9985	0.0033
N26	3	1.9115	0.6372	0.9815
N27	3	1.8209	0.6071	0.9315
N28	3	2.0201	0.6733	0.7263
N29	2	1.9316	0.9658	0.0684
N30	2	1.9241	0.9621	0.0759
N31	2	1.9961	0.9981	0.0039
N32	2	1.2113	0.6056	0.7887
N33	2	1.0074	0.5037	0.9926
N34	2	1.5112	0.7556	0.4888
N35	1	1	1	0
N36	1	1	1	0
N37	1	1	1	0

Appendix 4. Clustering of 1029 diverse compounds

Cluster	T. I.	R^2 with own cluster	R^2 with next closest	$1-R^2$ ratio					
N1	IDW	0.9971	0.8605	0.0212	N3	K2	0.9759	0.8231	0.1362
	MIDW	0.9461	0.9141	0.6274		K3	0.8979	0.8273	0.5912
	W	0.9925	0.8734	0.0591		AZV2	0.9333	0.8971	0.6481
	ID	0.9769	0.8944	0.2187		DN2S4	0.9935	0.8704	0.0499
	HD	0.9468	0.9236	0.6962		ASN3	0.9865	0.9145	0.1578
	S0	0.9751	0.8605	0.1792		ANN4	0.9597	0.9252	0.5394
	S1	0.9858	0.8698	0.1091		TOTOP	0.9215	0.7968	0.3864
	B0	0.9384	0.8921	0.5706		SUMI	0.7191	0.6644	0.8371
	B1	0.9323	0.8819	0.5731		TETS2	0.8783	0.7608	0.5088
	B3	0.8223	0.8041	0.9068		WP	0.8981	0.8273	0.5911
	K0	0.9954	0.9026	0.0471		PF	0.9701	0.8172	0.1644
	K1	0.9667	0.9553	0.7455		S4	0.9561	0.8552	0.3037
	AZV1	0.9638	0.9064	0.3869		S5	0.9299	0.7752	0.3121
	AZV3	0.9721	0.8942	0.2646		S6	0.8461	0.7757	0.6868
	AZS1	0.9792	0.8792	0.1722		B4	0.8589	0.8051	0.7237
	AZS2	0.9655	0.8893	0.3121		B5	0.8719	0.7551	0.5231
	DN2S3	0.9806	0.8891	0.1747		V4	0.8079	0.7064	0.6541
	DN211	0.9253	0.8591	0.5295		V5	0.8368	0.6801	0.5099
	DN214	0.9932	0.9223	0.0874		K4	0.9091	0.8181	0.4997
	AS14	0.9931	0.8671	0.0521		K5	0.9036	0.7582	0.3987
	DS11	0.9101	0.8712	0.6991		K6	0.8664	0.7347	0.5037
	ASN4	0.9804	0.8578	0.1379		AZV5	0.8853	0.8731	0.9038
	DN2N2	0.9251	0.8257	0.4304		ANV1	0.9149	0.8457	0.5514
	DN2N3	0.9849	0.9352	0.2339		ANV5	0.8107	0.6434	0.5309
	DN2N4	0.9902	0.9256	0.1313		SHOTHER	0.6811	0.5445	0.7001
	ANS1	0.9754	0.9053	0.2602		WT	0.9638	0.9295	0.5131
	ANS2	0.9457	0.9246	0.7207		HV	0.9807	0.7961	0.0946
	ANV3	0.9571	0.9155	0.5091		IC_BAR	0.7432	0.5347	0.5521
	ANV4	0.9895	0.9029	0.1081		MAX_ORB	0.6848	0.6257	0.8422
	AZN1	0.9851	0.8659	0.1108		ASV1	0.8942	0.7984	0.5247
	AZN2	0.9769	0.8456	0.1496		DSV1	0.9593	0.8021	0.2056
	AZN3	0.9908	0.8815	0.0778		DN2S1	0.9473	0.9096	0.5831
	AZN5	0.9754	0.8579	0.1733		DN2S2	0.8892	0.7695	0.4806
AN11	0.8818	0.7952	0.5773	DN2S5	0.9444	0.9133	0.6421		
AN12	0.9002	0.8801	0.8321	AS11	0.9651	0.9152	0.4134		
AN13	0.9957	0.8912	0.0395	AS15	0.9653	0.9125	0.3971		
AN14	0.9741	0.9478	0.4985	DSN2	0.9063	0.7533	0.3801		
AN15	0.9351	0.8177	0.3561	ANV2	0.8531	0.6165	0.3831		
ANN1	0.9945	0.8918	0.0505	KP2	0.8948	0.8241	0.5975		
ANN2	0.9849	0.8756	0.1211	IDCBAR	0.9532	0.8082	0.2441		
ANN3	0.9961	0.8976	0.0393	N5	B6	0.8075	0.7394	0.7386	
ANN5	0.9969	0.8929	0.0285		V6	0.8038	0.7341	0.7378	
NVX	0.9765	0.8801	0.1961		K7	0.7536	0.7535	0.9996	
KP0	0.8546	0.7505	0.5828		K8	0.8691	0.5933	0.3222	
KP1	0.8943	0.8781	0.8672		K9	0.8874	0.5177	0.2334	
IDC	0.9878	0.8623	0.0884		K10	0.8977	0.4676	0.1922	
N2	M1	0.9877	0.8969		0.1197	XP7	0.9071	0.6781	0.2885
	M2	0.9821	0.8427		0.1147	XP8	0.9213	0.5286	0.1668
	S2	0.9357	0.8609		0.4626	XP9	0.8734	0.4478	0.2294
	S3	0.9274	0.8504		0.4858	XP10	0.7821	0.3625	0.3421
					XVP7	0.9032	0.5901	0.2362	
					XVP8	0.8914	0.5278	0.2301	
				XVP9	0.8421	0.5096	0.3221		
				XVP10	0.7321	0.5143	0.5519		

N6	IC1	0.5968	0.4021	0.6742		V0	0.9702	0.8231	0.1685
	IC2	0.6477	0.4705	0.6654		V1	0.9275	0.7851	0.3371
	SIC2	0.8101	0.5231	0.3982		V2	0.8468	0.5532	0.3428
	SIC3	0.9624	0.2533	0.0504		V3	0.7678	0.6721	0.7077
	SIC4	0.9457	0.3867	0.0886		AZV4	0.9098	0.8312	0.5343
	SIC5	0.8971	0.4878	0.2011		FW	0.6427	0.4176	0.6134
	SIC6	0.8376	0.5557	0.3655		KA1	0.8505	0.8304	0.8814
	CIC3	0.8464	0.5498	0.3413	N13	ASN2	0.8973	0.7215	0.3688
	CIC4	0.9221	0.4123	0.1325		KP3	0.8594	0.4992	0.2808
	CIC5	0.9195	0.3469	0.1232		KA2	0.9045	0.8036	0.4863
	CIC6	0.8931	0.3028	0.1535		KA3	0.8525	0.4126	0.2511
N7	SPC4	0.8903	0.7343	0.4129		PHIA	0.9012	0.5871	0.2393
	SPC5	0.9328	0.8089	0.3516		SHCSATS	0.6096	0.4053	0.6564
	SPC6	0.8352	0.7638	0.6976		SSSCH2	0.6489	0.4351	0.6213
	BPC4	0.8317	0.5494	0.3734	N14	NRINGS	0.7372	0.5873	0.6368
	BPC5	0.9381	0.6321	0.1685		NCIRC	0.7031	0.5811	0.7087
	BPC6	0.8862	0.6211	0.3004		HMAX	0.5945	0.2105	0.5136
	VPC4	0.7028	0.4744	0.5654		HMIN	0.5041	0.2304	0.6444
	VPC5	0.8551	0.5971	0.3596		SAACH	0.8481	0.3077	0.2195
	VPC6	0.8382	0.6175	0.4232		SHAROM	0.8615	0.3143	0.2021
N8	ASZ1	0.9113	0.6997	0.2953		SAASC	0.1325	0.0241	0.8889
	ASZ2	0.9722	0.6277	0.0746	N15	SCY3	0.9031	0.1961	0.1207
	ANZ1	0.7325	0.4331	0.4717		SCY4	0.8722	0.4141	0.2181
	ANZ2	0.9353	0.5869	0.1566		BCY3	0.8964	0.1903	0.1279
	DSZ1	0.8633	0.7608	0.5713		BCY4	0.8806	0.4067	0.2012
	DSZ2	0.9258	0.5721	0.1733		VCY3	0.8314	0.1485	0.1981
	DN2Z1	0.9555	0.6687	0.1343		VCY4	0.8855	0.3566	0.1779
	DN2Z2	0.8581	0.5525	0.3174	N16	SC3	0.8518	0.5333	0.3176
	SSBR	0.2948	0.1072	0.7898		SC4	0.8779	0.3644	0.1921
N9	DN213	0.9486	0.6484	0.1462		BC3	0.9161	0.4412	0.1501
	AS13	0.9024	0.7719	0.4279		BC4	0.8758	0.3704	0.1973
	ASN1	0.8956	0.7738	0.4616		VC3	0.7702	0.2949	0.3259
	ASN5	0.9115	0.7742	0.3919		VC4	0.6767	0.0995	0.3591
	DSN1	0.9493	0.7309	0.1886	N17	ASV2	0.9024	0.5901	0.2379
	DN2N1	0.9513	0.6538	0.1407		ASV5	0.2629	0.1304	0.8476
	DN2N5	0.9481	0.6472	0.1471		DSV2	0.9103	0.5471	0.1981
	AZN4	0.3876	0.1913	0.7573		DN212	0.8868	0.5394	0.2457
N10	MAX_IC	0.7312	0.6197	0.7071		AS12	0.9269	0.7243	0.2652
	I_ORB	0.9357	0.4516	0.1173		DS12	0.9443	0.5425	0.1218
	IC3	0.7664	0.5181	0.4848	N18	IC0	0.7613	0.5008	0.4782
	IC4	0.9011	0.3981	0.1643		SIC0	0.9316	0.6659	0.2047
	IC5	0.9544	0.3762	0.0731		SIC1	0.9369	0.4406	0.1128
	IC6	0.9749	0.3844	0.0408		CIC0	0.9055	0.6938	0.3086
	SI	0.9212	0.4974	0.1567		CIC1	0.9708	0.5931	0.0718
	NCLASS	0.8432	0.5736	0.3678		CIC2	0.7871	0.5858	0.5141
N11	SC5	0.8421	0.4769	0.3021	N19	SCY5	0.8927	0.4427	0.1926
	SC6	0.8958	0.2979	0.1484		SCY6	0.7873	0.3603	0.3325
	BC5	0.8801	0.4372	0.2132		BCY5	0.8972	0.4831	0.1989
	BC6	0.8883	0.2881	0.1569		BCY6	0.8972	0.2186	0.1316
	SUMDELI	0.5651	0.5294	0.9242		VCY5	0.8861	0.4442	0.2051
	SSF	0.4726	0.2975	0.7507		VCY6	0.8981	0.1976	0.1271
	GMIN	0.7632	0.4916	0.4657	N20	NELEM	0.6391	0.5323	0.7718
	KNOTP	0.2711	0.0696	0.7834		GMAX	0.7329	0.3766	0.4284
N12	B2	0.8165	0.7636	0.7763		SSCL	0.6396	0.3633	0.5661

	NUMHBA	0.8825	0.3623	0.1842
	SHHBA	0.9414	0.3981	0.0973
N21	J	0.7727	0.3615	0.3561
	JB	0.9797	0.2158	0.0259
	JX	0.9574	0.1697	0.0513
	JY	0.9755	0.2106	0.0311
N22	SDSCH	0.8785	0.1283	0.1394
	SHCSATU	0.4972	0.1058	0.5623
	SHVIN	0.8583	0.1282	0.1626
	SDSSC	0.2044	0.0436	0.8319
N23	XCH10	0.9513	0.2291	0.0632
	XVCH10	0.6476	0.1355	0.4077
	SAAAC	0.5988	0.2374	0.5261
N24	SHCHNX	0.9935	0.2682	0.0088
	NUMWHBD	0.9821	0.2735	0.0246
	SHWHBD	0.9944	0.2671	0.0076
N25	SHSSH	0.9989	0.0107	0.0011
	SSSH	0.9978	0.0113	0.0022
	NUMHBD	0.9988	0.0119	0.0012
N26	XCH8	0.9432	0.0228	0.0581
	XVCH8	0.9161	0.0265	0.0862
	STSC	0.0522	0.0084	0.9558
N27	VC5	0.8379	0.3085	0.2344
	VC6	0.8426	0.1987	0.1964
	KNOTPV	0.1403	0.0158	0.8735
N28	SSCH3	0.7748	0.3068	0.3249
	QV	0.8138	0.5158	0.3845
	SSSSCH	0.4314	0.0922	0.6263
N29	XCH7	0.9658	0.0594	0.0364
	XVCH7	0.9658	0.0773	0.0371
N30	XCH9	0.9621	0.0391	0.0395
	XVCH9	0.9621	0.0874	0.0416
N31	SDCH2	0.9981	0.1446	0.0023
	SHTVIN	0.9981	0.1441	0.0023
N32	SDDC	0.6056	0.0263	0.4051
	SDS	0.6056	0.0212	0.4029
N33	SSSSB	0.5037	0.0102	0.5014
	SSSS	0.5037	0.0183	0.5056
N34	SSI	0.7556	0.0951	0.2701
	SSSSC	0.7556	0.3881	0.3993
N35	SSSPH	1	0.0036	0
N37	SAAS	1	0.0204	0
N38	SSSP	1	0.0083	0

Appendix 5. Clustering of 2887 compounds

Cluster	Members	Variation explained	Proportion explained	Second eigenvalue
T1	48	45.6682	0.9514	0.7301
T2	21	19.3797	0.9228	0.4927
T3	14	11.6125	0.8295	0.8171
T4	12	10.9466	0.9122	0.4211
T5	11	9.1359	0.8305	0.9921
T6	11	7.3647	0.6695	0.9815
T7	10	8.7445	0.8744	0.6178
T8	9	7.6291	0.8477	0.5576
T9	9	7.7142	0.8571	0.8348
T10	8	7.0956	0.8871	0.5248
T11	8	5.4259	0.6782	0.9571
T12	7	6.4366	0.9195	0.3201
T13	7	4.9436	0.7062	0.9061
T14	6	4.9361	0.8227	0.5209
T15	6	5.1848	0.8641	0.7213
T16	6	5.0117	0.8353	0.4615
T17	5	4.2258	0.8452	0.4443
T18	5	3.7116	0.7423	0.9895
T19	5	3.5557	0.7111	0.9244
T20	5	4.6363	0.9273	0.3136
T21	5	3.9021	0.7804	0.7422
T22	4	2.9563	0.7391	0.9659
T23	4	3.1266	0.7816	0.6816
T24	4	3.3631	0.8407	0.6125
T25	3	2.8434	0.9478	0.1178
T26	3	2.2123	0.7374	0.6833
T27	3	1.9955	0.6652	0.9924
T28	3	1.7525	0.5842	0.8144
T29	3	2.0999	0.7001	0.8568
T30	2	1.9971	0.9985	0.0029
T31	2	1.9222	0.9611	0.0778
T32	2	1.7785	0.8892	0.2215
T33	2	1.9597	0.9799	0.0403
T34	2	1.8271	0.9136	0.1729
T35	2	1.9956	0.9978	0.0044
T36	2	1.6507	0.8254	0.3493
T37	2	1.9654	0.9827	0.0346
T38	2	1.9991	0.9995	0.0011
T39	2	1.9973	0.9986	0.0027
T40	2	1.9766	0.9883	0.0234
T41	2	1.8534	0.9267	0.1466
T42	2	1.9438	0.9719	0.0562
T43	2	1.7366	0.8683	0.2634
T44	2	1.8033	0.9016	0.1967
T45	2	1.5357	0.7679	0.4643
T46	2	1.6881	0.8441	0.3119
T47	2	1.1454	0.5727	0.8546
T48	2	1.0056	0.5028	0.9944
T49	2	1.0007	0.5004	0.9993
T50	2	1.6205	0.8103	0.3795

T51	1	1	1	0
T52	1	1	1	0
T53	1	1	1	0
T54	1	1	1	0
T55	1	1	1	0
T56	1	1	1	0

Appendix 6. Clustering of 2887 compounds

Cluster	T. I.	R^2 with	R^2 with	$1-R^2$ ratio
		Own cluster	Next closest	
T1	IDW	0.9975	0.8656	0.0182
	MIDW	0.9487	0.9166	0.6158
	W	0.9965	0.8397	0.0221
	ID	0.9718	0.9143	0.3293
	HD	0.9484	0.9239	0.6777
	S0	0.9784	0.8355	0.1311
	S1	0.9831	0.8945	0.1598
	B0	0.9362	0.8262	0.3669
	B1	0.9094	0.8272	0.5241
	K0	0.9941	0.9101	0.0665
	AZV1	0.9454	0.9271	0.7482
	AZV3	0.9611	0.9135	0.4512
	AZV4	0.8831	0.8356	0.7112
	AZS1	0.9832	0.8137	0.0901
	AZS2	0.9719	0.8151	0.1521
	DN2S1	0.9257	0.9031	0.7666
	DN2S3	0.9894	0.8139	0.0571
	DN2S5	0.9296	0.8978	0.6893
	DN211	0.9359	0.8647	0.4734
	DN214	0.9876	0.9337	0.1866
	AS14	0.9921	0.8846	0.0685
	DS11	0.9092	0.8531	0.6185
	ASN4	0.9809	0.8548	0.1317
	DN2N2	0.9348	0.8471	0.4265
	DN2N3	0.9789	0.9396	0.3489
	DN2N4	0.9831	0.9397	0.2811
	ANS1	0.9838	0.8338	0.0973
	ANS2	0.9579	0.8677	0.3184
	ANV3	0.9447	0.9401	0.9238
	ANV4	0.9889	0.9128	0.1273
	AZN1	0.9841	0.8827	0.1362
	AZN2	0.9769	0.8654	0.1714
	AZN3	0.9891	0.8948	0.1051
	AZN5	0.9756	0.8755	0.1961
	AN11	0.8987	0.7058	0.3444
	AN13	0.9962	0.8945	0.0365
	AN14	0.9667	0.9538	0.7201
	AN15	0.9411	0.7692	0.2557
	ANN1	0.9954	0.8934	0.0435
	ANN2	0.9894	0.8705	0.0816
	ANN3	0.9955	0.9026	0.0464
	ANN5	0.9966	0.8972	0.0334
NVX	0.9442	0.8527	0.3792	
KP0	0.8765	0.7832	0.5697	
KP1	0.9006	0.8226	0.5602	
KA1	0.8124	0.7921	0.9018	
SUMI	0.6082	0.5687	0.9084	
IDC	0.9901	0.8499	0.0657	
T2	M1	0.9852	0.8821	0.1258
	M2	0.9842	0.8257	0.0909
	S2	0.9034	0.8526	0.6556

	S3	0.9501	0.7992	0.2483		XVP10	0.6821	0.3031	0.4564
	S4	0.8885	0.8706	0.8619	T6	IC0	0.7402	0.6088	0.6643
	B3	0.8004	0.7781	0.8996		IC1	0.5673	0.4751	0.8243
	K1	0.9718	0.9521	0.5883		SIC0	0.8458	0.4191	0.2654
	K2	0.9619	0.8098	0.2003		SIC1	0.9444	0.4233	0.0964
	K3	0.9433	0.7414	0.2191		CIC0	0.7744	0.5665	0.5203
	K4	0.8691	0.8088	0.6852		CIC1	0.9286	0.4364	0.1267
	AZV2	0.9161	0.8721	0.6565		CIC2	0.7573	0.6121	0.6256
	DN2S4	0.9891	0.8561	0.0759		HMIN	0.3701	0.2194	0.8071
	ASN3	0.9774	0.9065	0.2421		SSCH3	0.3804	0.1269	0.7097
	ANV1	0.9004	0.8362	0.6081		SHCSATS	0.5111	0.3561	0.7591
	AN12	0.9181	0.9014	0.8305		SSSCH2	0.5452	0.3254	0.6742
	ANN4	0.9658	0.9128	0.3924	T7	IC2	0.5831	0.4345	0.7373
	TOTOP	0.8441	0.7181	0.5529		SIC2	0.7714	0.5891	0.5562
	TETS2	0.7565	0.6224	0.6449		SIC3	0.9631	0.3463	0.0564
	WP	0.9428	0.7405	0.2204		SIC4	0.9629	0.3705	0.0591
	PF	0.9574	0.8053	0.2186		SIC5	0.9232	0.4637	0.1431
	WT	0.9544	0.9222	0.5863		SIC6	0.8777	0.5177	0.2535
T3	S5	0.9113	0.7952	0.4331		CIC3	0.8737	0.5735	0.2962
	S6	0.8982	0.7318	0.3797		CIC4	0.9393	0.4604	0.1125
	B4	0.8202	0.7753	0.8002		CIC5	0.9362	0.3981	0.1059
	B5	0.9082	0.7076	0.3141		CIC6	0.9139	0.3554	0.1336
	B6	0.8349	0.6558	0.4796	T8	SPC4	0.8458	0.7284	0.5679
	V4	0.7811	0.7211	0.7851		SPC5	0.9138	0.8123	0.4594
	V5	0.8564	0.6481	0.4081		SPC6	0.8395	0.7782	0.7235
	V6	0.7781	0.6404	0.6171		BPC4	0.8222	0.5503	0.3954
	K5	0.8659	0.7948	0.6532		BPC5	0.9294	0.6275	0.1896
	K6	0.8901	0.7661	0.4701		BPC6	0.8981	0.6269	0.2732
	K7	0.8006	0.6552	0.5782		VPC4	0.7267	0.4752	0.5208
	AZV5	0.8821	0.8407	0.7398		VPC5	0.8387	0.5576	0.3646
	ANV5	0.7417	0.6811	0.8099		VPC6	0.8148	0.5663	0.4271
	SHOTHER	0.6437	0.5289	0.7563	T9	MAX_IC	0.7486	0.4545	0.4608
T4	HV	0.9749	0.7683	0.1083		I_ORB	0.9426	0.4577	0.1058
	IC_BAR	0.7761	0.5433	0.4906		MAX_ORB	0.6145	0.5087	0.7847
	ASV1	0.9193	0.7691	0.3494		IC3	0.7601	0.5049	0.4846
	DSV1	0.9692	0.7915	0.1475		IC4	0.8989	0.4023	0.1691
	DN2S2	0.9052	0.6782	0.2945		IC5	0.9518	0.3913	0.0791
	AS11	0.9382	0.9095	0.6831		IC6	0.9706	0.3979	0.0489
	AS15	0.9402	0.9066	0.6401		SI	0.9369	0.4928	0.1245
	DSN2	0.9115	0.7176	0.3134		NCLASS	0.8903	0.5741	0.2576
	ANV2	0.8845	0.6211	0.3047	T10	DN213	0.9521	0.6528	0.1381
	KP2	0.9241	0.7199	0.2714		AS13	0.9181	0.7743	0.3627
	KA2	0.8472	0.8066	0.7901		ASN1	0.9118	0.7761	0.3938
	IDCBAR	0.9564	0.7941	0.2118		ASN5	0.9261	0.7766	0.3309
T5	K8	0.7734	0.6362	0.6228		DSN1	0.9542	0.7365	0.1737
	K9	0.8634	0.4901	0.2679		DN2N1	0.9543	0.6579	0.1336
	K10	0.8844	0.4191	0.1989		DN2N5	0.9514	0.6517	0.1396
	XP7	0.8037	0.7364	0.7448		AZN4	0.5277	0.2921	0.6672
	XP8	0.9144	0.5584	0.1939	T11	NELEM	0.6339	0.4302	0.6424
	XP9	0.9032	0.4163	0.1658		SUMDELI	0.8232	0.2471	0.2348
	XP10	0.8174	0.3211	0.2689		GMAX	0.8283	0.2056	0.2161
	XVP7	0.8349	0.6082	0.4214		SDO	0.4891	0.0983	0.5666
	XVP8	0.8578	0.4498	0.2585		NUMHBA	0.8497	0.2956	0.2134
	XVP9	0.8013	0.3441	0.3031		SHHBA	0.9335	0.2809	0.0925

	QV	0.7621	0.5607	0.5416	T21	ASV5	0.3191	0.1261	0.7791
	SSSSCH	0.1061	0.0646	0.9556		ASN2	0.8698	0.7294	0.4814
T12	ASZ1	0.9228	0.6572	0.2252		KP3	0.9256	0.4864	0.1448
	ASZ2	0.9921	0.6323	0.0216		KA3	0.9201	0.4074	0.1348
	ANZ2	0.8701	0.6754	0.4002		PHIA	0.8675	0.6546	0.3836
	DSZ1	0.9006	0.7524	0.4016	T22	SCY5	0.9731	0.4901	0.0527
	DSZ2	0.9413	0.5759	0.1383		BCY5	0.9681	0.5134	0.0658
	DN2Z1	0.9717	0.6861	0.0901		VCY5	0.9557	0.5247	0.0931
	DN2Z2	0.8381	0.5038	0.3264		SAAS	0.0594	0.0136	0.9536
T13	SC5	0.8331	0.4565	0.3071	T23	SSI	0.4082	0.0727	0.6382
	SC6	0.9177	0.2959	0.1169		SSSSSC	0.8642	0.3551	0.2105
	BC5	0.8927	0.4176	0.1843		KNOTP	0.9354	0.0986	0.0717
	BC6	0.8987	0.2834	0.1414		KNOTPV	0.9188	0.1187	0.0922
	SSF	0.4831	0.1295	0.5938	T24	SHCHNX	0.9646	0.0511	0.0373
	GMIN	0.7299	0.3126	0.3931		SSCL	0.4741	0.1343	0.6076
	SDSSC	0.1885	0.0752	0.8775		NUMWHBD	0.9602	0.0509	0.0419
T14	SC3	0.8248	0.5705	0.4079		SHWHBD	0.9641	0.0526	0.0379
	SC4	0.8739	0.3682	0.1996	T25	SCY6	0.9234	0.4229	0.1327
	BC3	0.9067	0.4972	0.1856		BCY6	0.9703	0.5418	0.0649
	BC4	0.8739	0.3701	0.2003		VCY6	0.9497	0.5278	0.1065
	VC3	0.7861	0.3741	0.3418	T26	XCH10	0.9397	0.2167	0.0771
	VC4	0.6708	0.1538	0.3891		XVCH10	0.6554	0.1351	0.3985
T15	SCY3	0.8752	0.2413	0.1645		SAAAC	0.6172	0.1701	0.4612
	SCY4	0.8808	0.4763	0.2276	T27	SHDNH	0.0176	0.0117	0.9941
	BCY3	0.8717	0.2374	0.1683		NHBINT2	0.9922	0.1095	0.0088
	BCY4	0.8781	0.4675	0.2292		SHBINT2	0.9857	0.1161	0.0161
	VCY3	0.8445	0.2279	0.2014	T28	SDDC	0.7404	0.0114	0.2625
	VCY4	0.8347	0.4446	0.2976		SDSN	0.6281	0.0077	0.3748
T16	B2	0.8045	0.7432	0.7614		SDS	0.3839	0.0064	0.6201
	V0	0.9498	0.8128	0.2679	T29	SDSCH	0.9409	0.1428	0.0691
	V1	0.9346	0.7383	0.2497		SHCSATU	0.2449	0.0624	0.8054
	V2	0.8881	0.5153	0.2311		SHVIN	0.9141	0.1227	0.0979
	V3	0.8273	0.6401	0.4796	T30	SHSSH	0.9985	0.0067	0.0015
	FW	0.6074	0.4556	0.7211		SSSH	0.9985	0.0071	0.0015
T17	SHSOH	0.8765	0.1071	0.1382	T31	XCH8	0.9611	0.0106	0.0393
	HMAX	0.7304	0.1494	0.3171		XVCH8	0.9611	0.0059	0.0391
	SSOH	0.8751	0.0929	0.1377	T32	VC5	0.8892	0.3219	0.1633
	NUMHBD	0.8191	0.2083	0.2286		VC6	0.8892	0.1896	0.1367
	SHHBD	0.9247	0.1704	0.0907	T33	XCH7	0.9799	0.0174	0.0205
T18	J	0.7881	0.3856	0.3451		XVCH7	0.9799	0.0159	0.0205
	JB	0.9802	0.1903	0.0244	T34	XCH9	0.9136	0.0453	0.0905
	JX	0.9518	0.1702	0.0581		XVCH9	0.9136	0.0427	0.0903
	JY	0.9768	0.2013	0.0291	T35	SDCH2	0.9978	0.1391	0.0026
	SDDSN	0.0149	0.0044	0.9895		SHTVIN	0.9978	0.1382	0.0026
T19	NRINGS	0.8469	0.5669	0.3535	T36	STSC	0.8254	0.2957	0.2481
	NCIRC	0.8182	0.5599	0.4132		STN	0.8254	0.0184	0.1779
	SAACH	0.8709	0.3313	0.1931	T37	NHBINT8	0.9827	0.1091	0.0194
	SHAROM	0.8665	0.3401	0.2023		SHBINT8	0.9827	0.1641	0.0207
	SAASC	0.1533	0.0269	0.8701	T38	SHTCH	0.9995	0.0825	0.0005
T20	ASV2	0.9064	0.6258	0.2502		STCH	0.9995	0.0826	0.0005
	DSV2	0.8986	0.5667	0.2341	T39	SHSNH2	0.9986	0.0415	0.0014
	DN212	0.9108	0.5463	0.1966		SSNH2	0.9986	0.0377	0.0014
	AS12	0.9635	0.7491	0.1453	T40	SHSSNH	0.9883	0.0313	0.0121
	DS12	0.9571	0.5524	0.0961		SSSNH	0.9883	0.0229	0.0121

T41	NHBINT9	0.9267	0.0161	0.0745		SDSSSP	0.5727	0.0101	0.4316
	SHBINT9	0.9267	0.0251	0.0752	T48	SAAN	0.5028	0.0502	0.5235
T42	NHBINT3	0.9719	0.1407	0.0327		SAASN	0.5028	0.0059	0.5002
	SHBINT3	0.9719	0.1423	0.0328	T49	SSSS	0.5004	0.0151	0.5073
T43	NHBINT6	0.8683	0.0387	0.1371		SDSSS	0.5004	0.0046	0.5019
	SHBINT6	0.8683	0.0712	0.1418	T50	ANZ1	0.8103	0.5172	0.3931
T44	NHBINT4	0.9016	0.1112	0.1107		SSBR	0.8103	0.2062	0.2391
	SHBINT4	0.9016	0.0821	0.1072	T51	SSSPH	1	0.0028	0
T45	NHBINT5	0.7679	0.0441	0.2428	T52	SAAO	1	0.0106	0
	SHBINT5	0.7679	0.0704	0.2497	T53	SSSSN	1	0.0175	0
T46	NHBINT7	0.8441	0.0625	0.1664	T54	SDDSSS	1	0.0234	0
	SHBINT7	0.8441	0.1861	0.1916	T55	SSSSP	1	0.0051	0
T47	SSSO	0.5727	0.1421	0.4981	T56	SAANH	1	0.0062	0
