Intrinsic Dimensionality of Chemical Space: Characterization and Applications

Subhash C. Basak [1]* Gregory D. Grunwald[2] and Subhabrata Majumdar[3]

[1] University of Minnesota Duluth-Natural Resources Research Institute (UMD-NRRI) and Department of Chemistry and Biochemistry, University of Minnesota Duluth, 5013 Miller Trunk Highway, Duluth, MN 55811, USA; sbasak@nrri.umn.edu

[2] University of Minnesota Duluth-Natural Resources Research Institute (UMD-NRRI), 5013 Miller Trunk Highway, Duluth, MN 55811, USA

[3] School of Statistics, University of Minnesota, Twin Cities, Minneapolis, MN 55414

* Author to whom correspondence should be addressed; E-Mail: sbasak@nrri.umn.edu
   Tel.: +1-218-727-1335

Abstract:

One popular method for the representation and characterization of chemical structure is through the use of their computed mathematical descriptors. Such descriptors, often called molecular descriptors, quantify different aspects of molecular structure, viz., size, shape, branching, cyclicity, bonding patterns, etc. Applications of discrete mathematics in the development of molecular descriptors began in the middle of the twentieth century and the trend is going on in an unabated manner even today. While in the 1970s only a few descriptors could be calculated, currently available software can calculate a large number of descriptors for molecules or biomolecules like DNA/ RNA, proteins. When p molecular descriptors are calculated for n molecules, the data set can be viewed as n vectors in p dimensions, each chemical being represented as a point in R^p. Because many of the descriptors are strongly correlated, the n points in R^p will lie on a subspace of dimension lower than p. Methods like principal components analysis (PCA) can be used to characterize the intrinsic dimensionality of chemical spaces. Since the early 1980s, Basak et al have carried out PCA of various congeneric and diverse data sets relevant to new drug discovery and predictive toxicology. Principal components (PCs) derived from mathematical chemodescriptors have been used in the formulation of quantitative structure-activity relationships (QSARs), clustering of large combinatorial libraries as well as quantitative molecular similarity analysis (QMSA). This presentation will review the results of PCA carried out by Basak and coworkers since the early 1980s to the present time in the characterization and visualization of

chemical spaces with special reference to five data sets, both congeneric and structurally diverse: 1) A large and structurally diverse set of 3,692 chemicals which was a subset of the Toxic Substances Control Act (TSCA) Inventory maintained by the United States Environmental Protection Agency (USEPA), 2) A set of 74 alkanes, 3) A virtual library of 248,832 psoralen derivatives, 4) A congeneric set of 95 aromatic and heteroaromatic amine mutagens, and 5) A structurally diverse collection of 508 chemicals mutagens.

Mol2Net YouTube channel: *http://bit.do/mol2net-tube*

1. Introduction

Mathematical chemistry, or more correctly discrete mathematical chemistry, had its beginning at the middle of the twentieth century probably with the publication of the seminal paper by Harry Wiener [1] on the calculation of .structural indices for the prediction of molecular properties. Although representation of chemical species by graphs was explored by Sylvester [2] as early as 1878, the characterization of molecular structure by graph invariants has made great strides during the past half century or so [3-16] following the seminal work of Wiener [1]. Invariants of graphs associated with molecules and biomolecules quantify certain aspects of their structure and have been used in the characterization and comparison of such structures as well as prediction of their properties [4, 17, 18, 19]. Specifically, such invariants and orthogonal factors like principal components PCs) derived from them have found applications in quantitative structure-activity relationship (QSAR) studies [3-15, 20], quantitative molecular similarity analysis (QMSA) research [21-24],

clustering of large libraries of structures into smaller subsets [23, 24], and in the discrimination of pathological structures like isospectral graphs [17]. One of the authors of this paper (Basak) has been involved since the early 1970s in the development of novel numerical graph invariants or topological indices (TIs) [6, 7, 11, 25-27] as well as biodescriptors derived from DNA/ RNA sequences [28] and proteomics maps [29]. Basak's research [20] carried out with colleagues at the University of Calcutta, India, in the 1970s involved mainly formulation of QSARs of congeneric sets of chemicals using their own information theore5tic indices and topological indices developed by Bonchev & Trinajstić [4, 5], Randic [9-12] & Kier and Hall [3] as well as physical properties like van der Waals' volume, calculated or experimental hydrophobicity (log P, octanol water) [20]. In the early 1980s, after Basak joined the University of Minnesota Duluth, the software POLLY [30] was developed and large scale calculation of TIs for QSAR and QMSA analyses was initiated. In one of the

earliest studies of its kind, Basak et al [31] used POLLY for the calculation of ninety TIs for a collection of 3,692 structurally diverse chemicals which was a subset of the Toxic Substances Control Act (TSCA) Inventory of the United States Environmental Protection Agency (USEPA). The authors carried out principal components analysis (PCA) on this data set and asked the question: *What is the intrinsic dimensionality of chemical structure measured by the large number of TIs*? This line of research, i.e., PCA and use of principal components (PCs) derived from different collection of TIs calculated by POLLY [30], MolConnZ [32], Triplet [33, 34],

2 Results and Discussion

2.1 A large and structurally diverse set of 3,692 chemicals.

For this data set, 90 TIs were calculated by the POLLY [30] software and PCA was performed. For details of the list of the particular TIs calculated for this study see Basak et al [21, 31]. Results showed that first ten PCs with eigenvalues greater than or equal to 1 explained 92.6% of the variance in the data and PC_1-PC_4 explained 78.3% of the variation in the original variables. Regarding the correlation profiles of the original variables or TIs with the first four important PCs, Table 1 below gives the data:

and APProbe [35] in QSAR and QMSA, has continued to this day. This paper summarizes the results and the lessons learned from a few of these studies using both congeneric and structurally diverse sets of chemicals, viz., 1) A large and structurally diverse set of 3,692 chemicals mentioned above, and 2) A data set of 74 alkanes, 3) A virtual library of 248,832 psoralen derivatives, 4) A congeneric set of 95 aromatic and heteroaromatic amine mutagens, and 5) A structurally diverse collection of 508 chemicals mutagens.

.

Table 1: Correlation of the first four PCs with the original variables including topological indices.

PC1	PC2	PC3	PC4
K_1 (.96)	SIC_3 (0.97)	$^4\chi^b_C$ (.69)	$^4\chi_{CH}$ (.85)
$^2\chi$ (.95)	CIC_4 (-.96)	$^4\chi^b_C$ (.69)	$^4\chi^b_{CH}$ (.84)
$^3\chi$ (.95)	CIC_3 (-.95)	$^5\chi^b_C$ (.68)	$^4\chi^v_{CH}$ (.80)
K_2 (.95)	$SIC4$ (.95)	$4\chi_C$ (.68)	$^3\chi_{CH}$ (.75)
K_0 (.95)	SIC_2 (.94)	$^3\chi^v_C$ (.67)	$^3\chi^b_{CH}$ (.75)
$^1\chi$ (.94)	CIC_5 (-.94)	$^5\chi_C$ (.64)	$^4\chi^b_{CH}$ (.74)
$^3\chi^b$ (.94)	CIC_6 (-.92)	$^6\chi_C$ (.64)	$^3\chi^v_{CH}$ (.72)
$^4\chi$ (.94)	$SIC5$ (.92)	$^3\chi_C$ (.61)	$^5\chi_{CH}$ (.71)
$^4\chi^b$ (.93)	$SIC6$ (.89)	$6\chi^b_C$ (.60)	$^5\chi^v_{CH}$ (.67)
$^0\chi$ (.93)	CIC_2 (-.87)	$^5\chi^v_C$ (.60)	$^6\chi^b_{CH}$ (.47)

It is clear from the data in Table 1 that PC1 is strongly correlated with those indices which are related to size of chemicals. It is noteworthy that for the set of 3,692 chemicals PC_1 was also highly correlated (r = 0.81) with molecular weight. PC_2 may be interpreted as an axis of molecular complexity as encoded by the higher order information theoretic indices [27]. PC_3 is most highly related to the cluster/ path-cluster type molecular connectivity indices which quantify information regarding molecular branching. The data in Table 1 clearly show that PC_4 is strongly correlated with the cyclicity terms of the connectivity type.

2.2 A data set of 74 alkanes

For boiling point estimation lf alkanes, twenty six TIs, total surface area (TSA), and volume (V) were calculated for a set of 74 alkanes [36]. Table 2 below gives the three different two-parameter regression models for the prediction of boiling point.

Table 2: Results of three 2-parameter models in predicting the boiling points of 74 alkanes

Parameter	R^2	s. e.	F
$^1\chi$, CIC_2	99.4	3.72	5620
PC_1, PC_3	98.1	6.48	1828
V, TSA	97.0	8.17	1136

It appears from the data in Table 2 that the individual TIs, PCs derived from them as well as the calculated physical properties like volume and total surface area give good QSARs for this congeneric set of molecules. The TIs and PCs derived from them give a little bit superior models as compared to the properties.

2.3 A virtual library of 248,832 psoralen derivatives

A virtual library of 248,832 psoralen derivatives [23] was created and analyzed using PCs derived from TIs. For this study, a set of 92 topological indices was calculated by POLLY [30]. The set of TIs consisted of 37 topostructural and 55 topochemical indices. We define topostructural indices as those invariants which are derived from simple (unweighted) molecular graphs. Such graphs do not distinguish among different types of bonds or atoms. The Wiener index; cluster, path-cluster, and simple connectivity indices; and path length indices are examples of topostructural parameters. Topochemical indices, on the contrary, are indices defined on weighted molecular graphs such that the various types of atoms and bonds are weighted to reflect their nature and contribution to chemical bonding. The SIC, CIC, and IC indices as well as both bonding and valence connectivity indices are all examples of topochemical indices. For this data set, the top 3 PCs explained 89.2% of the variance in the data; first 6 PCs explained 95.5% of the variance of the original calculated indices. The PCs were used to cluster the large set of chemicals into a smaller subset as an exercise of managing combinatorial explosion that can happen in the drug design scenarios when one wants to create a large pool of derivatives of a lead compound. For details of the outcome of clustering of the 248,832 psoralen derivatives, please see [23].

For the large but congeneric set of 248,832 psoralen derivatives, first 6 PCs explained 95.5% of the variance of the original calculated indices.

2.4 A homogeneous set of aromatic amines

<Data description>

Table 3: Top 10 variables behind each of first 4 PCs (loadings in brackets) for 95 compound dataset

PC1	PC2	PC3	PC4
ΔHf (-0.20)	ΔHf (0.75)	E_{HOMO} (-0.52)	fw (-0.76)
$DN^2 1_4$ (-0.14)	$^9\chi$ (0.23)	ΔHf (-0.50)	E_{LUMO} (0.38)
I_D^W (-0.13)	$^{10}\chi$ (0.23)	E_{LUMO} (-0.36)	$SaaN$ (-0.32)
Wt (-0.13)	E_{LUMO} (0.21)	$^{10}\chi$ (0.19)	E_{HOMO-1} (-0.27)
EDH (-0.13)	$^8\chi$ (0.20)	$^9\chi$ (0.16)	SdO (0.24)
$DS1_4$ (-0.13)	E_{HOMO} (0.18)	fw (0.15)	nvx (0.23)
$AS1_4$ (-0.12)	$SaaaC$ (0.12)	$^8\chi$ (0.14)	$phia$ (-0.19)
$DN^2 Z_4$ (-0.12)	$^7\chi$ (0.11)	$SaaN$ (-0.13)	μ (0.14)
AZS_2 (-0.12)	$SHBint3$ (-0.09)	$^7\chi$ (0.12)	ΔHf (0.14)
ED (-0.12)	$XP8$ (0.07)	μ (-0.12)	$^9\chi$ (-0.14)

For the 95 aromatic amine set, PC_1 is correlated with different original variables including some triplet indices and some indices related to molecular size. PC_2 is most strongly correlated with the heat of formation, ΔHf (r = 0.75); the energy of the highest occupied molecular orbital, E_{HOMO} (r = -0.52) and ΔHf (r = -0.50) are most highly corr4elated with PC_3 whereas PC_4 is strongly correlated with fw (r = -0.76) which is the molecular weight of the chemical species.

2.5 A diverse set of 508 chemicals

<Data description>

Table 4: Top 10 variables behind each of first 4 PCs (loadings in brackets) for 508 compound diverse dataset

PC1	PC2	PC3	PC4
$DN^2 1_4$ (-9.89×10⁻²)	ΔHf (-0.77)	ΔHf (-0.62)	E_{HOMO} (0.96)
I_D^W (-9.51×10⁻²)	$^{10}\chi$ (-0.17)	$^8\chi$ (0.18)	E_{HOMO-1} (-0.21)
Wt (-9.47×10⁻²)	$^9\chi$ (-0.17)	$^9\chi$ (0.18)	E_{LUMO} (-0.09)
EDH (-9.38×10⁻²)	$^8\chi$ (-0.16)	$^{10}\chi$ (0.17)	E_{LUMO+1} (-0.07)
$AS1_4$ (-8.94×10⁻²)	$^7\chi$ (-0.15)	$^7\chi$ (0.17)	ΔHf (-0.04)
IdC (-8.59×10⁻²)	$^6\chi$ (-0.12)	$^6\chi$ (0.14)	μ (0.03)
ED (-8.52×10⁻²)	E_{LUMO+1} (0.12)	E_{LUMO} (-0.13)	$^5\chi$ (-0.03)
fw (-8.52×10⁻²)	E_{LUMO} (0.12)	$^5\chi$ (0.13)	DSZ_2 (0.03)
w (-8.43×10⁻²)	$^5\chi$ (-0.11)	$^6\chi_{PC}$ (0.12)	$^6\chi$ (-0.03)
$AN1_4$ (-8.37×10⁻²)	$^6\chi_{PC}$ (-0.09)	E_{LUMO+1} (-0.12)	$^4\chi$ (-0.03)

For the diverse 508 chemical mutagen set, the energy of the highest occupied molecular orbital, E_{HOMO} (r = .96) is most strongly correlated with PC_4; PC_3 is highly correlated with ΔHf (r = -.62) which is also negatively correlated with PC_2 (r = -.77); PC_1 is loaded with some triplet indices and those invariants which reflect molecular size.

3. Materials and Methods
For materials and methods of data collection and
statistical analyses, see [19-21, 23, 27, 30-35].

4. Conclusions
In this paper, we reviewed our over three decades of research on the use of topological indices and principal components analysis in the characterization of five data sets: 1) A large and structurally diverse set of 3,692 industrial chemicals which was a subset of the Toxic Substances Control Act (TSCA) Inventory of the United States Environmental Protection Agency (USEPA), 2) A data set of 74 alkanes, 3) A virtual library of 248,832 psoralen derivatives, 4) A congeneric set of 95 aromatic and heteroaromatic amine mutagens, and 5) A structurally diverse collection of 508 chemicals mutagens.

The results show that the PCs derived from the TIs can be used for the development of QSARs as exemplified in Table 2 with 74 alkanes. PCs derived from TI have been used in the clustering of large set of psoralens [23]. Basak et al also used both PCs and individual TIs for analog selection [24] and characterization of isospectral graphs [17]. The data presented here show the usefulness of TIs and PCs derived from them in the clustering/ characterization of chemical libraries as well as QSAR. Details of QMSA analyses using PCs derived from TIs are not given here for brevity.

Basak [37] recently noted: "Mathematical chemistry or more accurately discrete mathematical chemistry had a tremendous growth spurt in the second half of the twentieth century and the same trend is continuing now. This growth was fueled primarily by two major factors: 1) Novel applications of discrete mathematical concepts to chemical and biological systems, and 2) Availability of high speed computers and associated software whereby *hypothesis driven* as well as *discovery oriented* research on large data sets could be carried out in a timely manner. This led to the development of not only a plethora of new concepts, but also to various useful applications to such important areas as drug discovery, protection of human as well as ecological health, and chemoinformatics. Following the completion of the Human Genome Project in 2003, discrete mathematical methods were applied to the "omics" data to develop descriptors relevant to bioinformatics, toxicoinformatics, and computational biology."

Initially, TIs were used for the discrimination of structure and QSAR studies of congeneric and small sets of structures. For example, Randic's [9] first order connectivity index (1χ), the information theoretic indices developed by Bonchev and Trinajstić [38] and those developed by Raychaudhury et al [7] were used to discriminate the set of alkanes and they worked well in those cases. In the case of 18 octanes, the molecules do not vary from each other with respect to size, but primarily in terms of branching patterns. Therefore, the indices [7, 9, 38] were interpreted based on the data as reflecting molecular branching. But when PCA was carried out with a diverse set of 3,692 chemical structures, the results entered an uncharted territory and were counterintuitive, to say the least.   As shown from the correlation of the original variables with PC_1, 1χ and related indices were now strongly correlated with molecular size in the large and diverse set, not to molecular branching. PC_3 emerged as the axis representing

branching and was strongly correlated to the cluster type molecular connectivity indices. Further studies with both congeneric and diverse data sets are need to understand the utility of TIs and PCs derived from them in structure-activity analyses.

Author Contributions

Since the early 1970sn Subhash C. Basak has been involved in the development of novel topological indices and their applications in QSARs. Gregory D. Grunwald has been a collaborator of Basak since the early 1980s in the use of TIs in QSAR and QMSA analyses. Subhabrata Majumdar has been involved in the applications of calculated mathematical chemodescriptors in QSAR in collaboration with Basak and Grunwald for the past five years and the standardization of novel methods like Interrelated Two-way Clustering (ITC) for use in QSAR.

Conflicts of Interest

 "The authors declare no conflict of interest".

References and Notes

1. Wiener, H. 1947. Structural determination of paraffin boiling points. Journal of the American Chemical Society, 1947, 69, 17-20.

2. Sylvester, J.J. On an application of the new atomic theory to the graphical representation of the invariants and covariants of binary quantics, with three appendices. American Journal of Mathematics, 1878, 1, 64-125.

3. Kier, L.B.; Hall, L. Molecular Structure Description: The Electrotopological State; Academic Press: San Diego, CA, 1999, pp. 245.

4. Trinajstić, N. Chemical Graph Theory, 2nd ed.; CRC Press: Boca Raton, FL, 1992, pp. 352

5. Bonchev, D. Information Theoretic Indices for Characterization of Chemical Structures; Research studies Press: Chichester, U.K.; 1983.

6. Basak, S.C. Use of molecular complexity indices in predictive pharmacology and toxicology: A QSAR approach. Med. Sci. Res., 1987, 15, 605-609.

7. Raychaudhury, C.; Ray, S.K.; Ghosh, J.J.; Roy, A.B.; Basak, S.C. Discrimination of isomeric structures using information-theoretic topological indices. J. Comput. Chem., 1984, 5, 581-588

8. Hosoya, H. Topological Index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. Bull. Chem. Soc. Jpn., 1971, 44, 2332-2339.

9. Randic, M., Characterization of molecular branching. J. Am. Chem. Soc., 1975, 97, 6609-6615.

10. Balaban, A. T. "Distance Connectivity Index." Chem. Phys. Lett., 1982, 89, 399-404.

11. Basak, S. C., Restrepo, G. and Villaveces, J. L., Eds, Advances in Mathematical Chemistry and Applications, volume 1 & 2 , Bentham eBooks, Bentham Science Publishers and Elsevier, , 2015.

12. Devillers, J.; Balaban, A.T., Eds. Topological Indices and Related Descriptors in QSAR and QSPR; Gordon and Breach: Amsterdam, 1999, pp. 811.

13. Karelson, M., Molecular Descriptors in QSAR/QSPR, Wiley-Interscience, New York, 2000.

14. Todeschini, R.; Consonni, V., Molecular Descriptors for Chemoinformatics, Wiley-VCH, New York, 2009.

15. Gonzalez-Diaz, H.; Munteanu, C. R. (Editors), Topological Indices for Medicinal Chemistry, Biology, Parasitology, Neurological and Social Newworks, Transworld Research Neywork, 2011.

16. Janežič, D.; Miličević, A.; Nikolić, S. & Trinajstić, N.: 2007, Graph theoretic matrices in chemistry, Kragujevac: University of Kragujevac.

17. Balasubramanian, K.; Basak, S.C., Characterization of isospectral graphs using graph invariants and derived orthogonal parameters', Journal of Chemical Information and Computer Sciences, 1998, 38, 367-73.

18. Nandy, A.; Harle, M. & Basak, S.C., Mathematical descriptors of DNA sequences: Development and application, Arkivoc, 2006, 9, 211-38.

19. Basak, S. C. Philosophy of Mathematical Chemistry: A Personal Perspective, HYLE--International Journal for Philosophy of Chemistry, 2013, 19, 3-17.

20. Basak, S. C., Mathematical Descriptors for the Prediction of Property, Bioactivity, and Toxicity of Chemicals from their Structure: A Chemical-Cum-Biochemical Approach, Current Computer-Aided Drug Design, 2013, 9, 449-462.

21. Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R., Determining structural similarity of chemicals using graph-theoretic indices. Discrete Appl. Math. 1988, 19, 17–44.

22. Lajiness M (1990) Molecular similarity-based methods for selecting compounds for screening. In *Computational Chemical Graph Theory*, Ed. Rouvray DH (Nova, New York), pp 299-316.

23. Basak, S. C.; Mills, D.; Gute, B. D.; Balaban, A. T.; Basak, K.; Grunwald, G. D. Use of Mathematical Structural Invariants in Analyzing, Combinatorial Libraries: A Case Study with psoralen Derivatives, Current Computer-Aided Drug Design, 6, 240-251, 2010.

24. Basak, S. C. Molecular Similarity and Hazard Assessment of Chemicals: A Comparative Study of Arbitrary and Tailored Similarity Spaces, J. Eng. Sci. Manage. Educ. 7, 178-184, 2014

25. Balaban, A. T.; Mills, D.; Ivanciuc, O.; Basak, S. C. Reverse Wiener indices, Croat. Chim. Acta, 73, 923–941 (2000).

26. Nikolic, S.; Trinajstic, N.; Amic, D.; Beslo, D.; Basak S. C. Modeling the solubility of aliphatic alcohols in water. Graph connectivity indices versus line graph connectivity indices, In QSAR/QSPR Studies by Molecular Descriptors, M.V. Diudea, Ed., Nova Science Publishers, Huntington, New York, USA, pp. 63–81 (2001).

27. Basak, S. C. Information theoretic indices of neighborhood complexity and their applications, In Topological Indices and Related Descriptors in QSAR and QSPR, J. Devillers and A.T. Balaban, Eds., Gordon and Breach Science Publishers, The Netherlands, pp. 563–593 (1999).

28. Randic, M.; Vracko, M.; Nandy, A.; Basak, S. C. On 3-D graphical representation of DNA primary sequences and their numerical characterization, J. Chem. Inf. Comput. Sci., 40, 1235–1244 (2000).

29. Basak, S. C.; Gute, B. D. Mathematical descriptors of proteomics maps: Background and applications, Curr. Opin. Drug Discov. Devel., 11, 320-326 (2008).

30. Basak, S. C.; Harriss, D. K.; Magnuson, V. R. 1988. POLLY v. 2.3: 1988; Copyright of the University of Minnesota.

31. Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R.; Veith, G. D. Topological indices: their nature, mutual relatedness, and applications, Mathematical Modelling, 1987 8, 300–305..

32. MolConnZ, Version 4.05, 2003; Hall Ass. Consult.; Quincy, MA.

33. Basak, S.C.; Grunwald, G.D.; Balaban, A.T.TRIPLET: Copyright of the Regents of the University of Minnesota, 1993

34. Filip, P. A.; Balaban,T. S.; Balaban, A. T. A new approach for devising local graph invariants: derived topological indices with low degeneracy and good correlation ability. J. Math. Chem. 1987, 1, 61-83.

35. Basak, S. C.; Grunwald, G. D., APProbe. 1993; Copyright of the University of Minnesota

36. Basak, S. C.; Grunwald, G. D., A COMPARATIVE STUDY OF GRAPH INVARIANTS, TOTAL SURFACE AREA AND VOLUME IN PREDICTING BOILING POINTS OF ALKANES, Math Modelling & Sci. Computing, 1993, 2, 735-740.

37. Basak, S. C. Mathematical Structural Descriptors of Molecules and Biomolecules: Background and Applications, In, Advances in Mathematical Chemistry and Applications, volume 1, Basak, S. C., Restrepo, G. and Villaveces, J. L., Eds., pp. 3-23, Bentham eBooks, Bentham Science Publishers and Elsevier, , 2015.

38. Bonchev, D. & Trinajstić, N.: 1977, 'Information theory, distance matrix and molecular branching', Journal of Chemical Physics, 67, 4517-33.