Internet Electronic Journal of Molecular Design **2002**, *1*, 374–387 ISSN 1538–6414
BioChem Press http://www.biochempress.com

Inter*net* Electronic Journal of Molecular Design

August 2002, Volume 1, Number 8, Pages 374–387

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Milan Randić on the occasion of the 70th birthday
Part 4

Guest Editor: Mircea V. Diudea

Tailored Similarity Spaces for the Prediction of Physicochemical Properties

Brian D. Gute,[1] Subhash C. Basak,[1] Denise Mills,[1] and Douglas M. Hawkins[2]

[1] Center for Water and the Environment, Natural Resources Research Institute, 5013 Miller Trunk Highway, Duluth, MN, 55811
[2] School of Statistics, University of Minnesota, 224 Church St SE, Minneapolis, MN, 55455

Received: June 27, 2002; Accepted: July 9, 2002; Published: August 31, 2002

Citation of the article:
B. D. Gute, S. C. Basak, D. Mills, and D. M. Hawkins, Tailored Similarity Spaces for the Prediction of Physicochemical Properties, *Internet Electron. J. Mol. Des.* **2002**, *1*, 374–387, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Tailored Similarity Spaces for the Prediction of Physicochemical Properties[#]

Brian D. Gute,[1] Subhash C. Basak,[1,*] Denise Mills,[1] and Douglas M. Hawkins[2]

[1] Center for Water and the Environment, Natural Resources Research Institute, 5013 Miller Trunk Highway, Duluth, MN, 55811

[2] School of Statistics, University of Minnesota, 224 Church St SE, Minneapolis, MN, 55455

Received: June 27, 2002; Accepted: July 9, 2002; Published: August 31, 2002

Internet Electron. J. Mol. Des. **2002**, *1* (8), 374–387

Abstract

Motivation. In the past, molecular similarity spaces have been developed from arbitrary sets of molecular properties or theoretical descriptors and the results of property estimation based on these methods have always been inferior to SAR and QSAR models. Tailored QMSA methods attempt to create similarity spaces specific for a property of interest, rather than being purely arbitrary spaces characterizing the general aspects of all chemicals within the space or intuitively selected structure spaces whose elements are chosen subjectively. To this end, we have created three similarity spaces, two tailored and one non–tailored, for a set of 166 chemicals for which we have both log *P* and normal boiling point (BP) data. The tailored spaces were each tailored to one of the properties, while the other similarity space was developed using standard QMSA methods.

Method. Ridge regression was used to determine which of the available molecular descriptors were most useful in modeling each of the available properties. Fifteen topological descriptors were selected for use as dimensions within each the tailored similarity spaces. The same number of principal components were developed using principal component analysis for the arbitrary similarity space.

Results. The log *P* tailored similarity space was superior to both the arbitrary structure space and the BP tailored space for the estimation of log *P*. Also, the BP tailored similarity space was superior to the arbitrary structure space for the estimation of BP. Interestingly, the space tailored to model log *P* performed as well at modeling BP as did the BP tailored space. This unexpected result is explained by the degree of overlap between the indices used in both of the tailored spaces and in the presence of connectivity indices related to BP in the log *P* model.

Conclusions. The tailored similarity method presents a promising approach to creating property specific similarity spaces derived from structural descriptors based on the results of this study and from a previous study. Further work is necessary to determine to true utility of this method with large, diverse data sets.

Keywords. Quantitative molecular similarity analysis (QMSA); tailored QMSA; arbitrary QMSA; topological indices; lipophilicity; normal boiling point.

Abbreviations and notations

ASTER, Assessment Tools for the Evaluation of Risk	QMSA, quantitative molecular similarity analysis
BP, normal boiling point	QSAR, quantitative structure–activity relationship
ED, Euclidean distance	*R*, regression coefficient
JP–8, jet propellant formulation #8	RR, Ridge regression
*K*NN, *K*–nearest neighbor	*s.e.*, standard error
log *P*, lipophilicity	TI, topological index
PCs, principal components	USEPA, United States Environmental Protection Agency
PCA, principal components analysis	

[#] Dedicated to Professor Milan Randić on the occasion of the 70th birthday.

* Correspondence author; phone: 00–218–720–4230; fax: 00–218–720–4328; E–mail: sbasak@nrri.umn.edu.

1 INTRODUCTION

Quantitative molecular similarity analysis (QMSA) is an important computational tool both for the hazard assessment of environmental pollutants and pharmaceutical drug design. In the area of the hazard estimation of chemicals, QMSA methods are routinely used to assess the potential hazard of a chemical based on the toxicity profiles of analogous chemicals when little or no experimental toxicity data and toxicologically relevant property data are available for the chemical of interest [1–4]. This course of action is generally followed when the structure of the chemical is complex enough that it cannot be unambiguously classified into a particular structural category. If it could be categorized into a specific chemical class, class–specific quantitative structure–activity relationship (QSAR) models would instead be used for hazard assessment. In the area of drug discovery, QMSA techniques are useful for determining whether interesting lead compounds have structural analogs with similar pharmacological and toxicological profiles. The other side of similarity is dissimilarity. Dissimilarity–based clustering of large libraries of real or *in silico* (virtual libraries) of chemicals has been successfully used [5] and suggested [6] as possible methods in the management of combinatorial explosions in various drug design scenarios.

QMSA methods are based on the basic assumption that similar molecular structures usually have similar properties [7]. Two chemicals, X1 and X2, are said to be similar if they resemble each other with respect to some user–defined set of properties or structural attributes, or both. Substructural descriptors [8–17], experimental properties [12,17–19], and theoretical structural invariants [6,7,11–17,19–32] have been widely used in the formulation of QMSA methods and ranking of chemical databases via such techniques.

Our research group has been involved in the development of novel QMSA techniques and their applications in analog selection and the *k*–nearest neighbor (*K*NN) based estimation of properties, as well as the use of similarity spaces in the clustering of chemical databases. Our experience has shown that increasing the intrinsic dimensionality of similarity spaces by the progressive use of more diverse and mutually uncorrelated (or minimally correlated) indices leads to better analog selection as is evident from both a visual inspection of their structures and the predictive power of the selected analogs in property estimation for query chemicals using the *K*NN method.

The stepwise use of increasingly higher dimensional structure spaces, derived from collections of progressively more diverse and comprehensive indices, suffers from the fact that elements of the enhanced spaces do not have any intrinsic relationship to the property of interest that we are attempting to estimate from the chosen analogs. Rather, these spaces are simply a reflection of the chemical diversity within the selected data set. If there is an improvement in the usefulness of analogs selected, that is only by chance, not by design. This is why we have developed the idea of tailored QMSA methods where the structure space is constructed from parameters that are strongly associated with the property of interest [32]. The advantage of such directed spaces over blind or

arbitrary spaces is that analogs selected by the former will be relevant with respect to the property to which they are tailored.

In a previous study, we reported for the first time the development of structure spaces tailored towards two properties, *viz*., log *P* (octanol/water) and Ames mutagenicity, based on calculated topological indices. We also showed that the analogs selected from the tailored similarity spaces gave much better results in *K*NN–based estimation for both of the properties studied, as compared to our previous results using arbitrary similarity spaces. In the current study, we have used a set of 166 chemicals which represent a subset of the known constituents of jet propellant #8 (JP–8), a jet fuel currently in use by the United States Armed Forces. This set of chemicals was of interest for this study since we have data for two physicochemical properties, log *P* and normal boiling point, for this set of chemicals. Three similarity spaces have been constructed for this study. Two of the similarity spaces are tailored spaces, one tailored towards log *P* and the other towards normal boiling point (hereafter simply referred to as BP). The third similarity space is a standard, arbitrary similarity space developed from the set of available molecular descriptors.

2 MATERIALS AND METHODS

Physicochemical property data used in this study represent property values extracted from the ASTER [33] system of the USEPA. These data are predominantly calculated values, rather than experimental values, reflecting the difficulty of obtaining simple physicochemical experimental data for common compounds.

2.1 Chemical Data

The set of chemicals used in this study represents a subset of the known constituents of JP–8 identified through GC/MS [34], a set of 166 hydrocarbons. This subset consisted of all of the chemicals in the full set of 228 chemicals for which log *P* and normal boiling point (BP) were both available from the ASTER database. However, even for the reduced set of 166 chemicals, most of the data values available from ASTER were calculated, not experimental values. This set of chemicals and the data obtained from ASTER are reported in Table 1.

2.2 Calculation of Molecular Descriptors

The topological indices (TIs) used in this study were calculated using three main software programs: POLLY 2.3 [35], MolConn–Z 3.50 [36], and Triplet [37]. Included in the suite of more than 220 indices in this study are: Wiener number [38], molecular connectivity indices as calculated by Randić [39] and Kier and Hall [40], frequency of path lengths of varying size, information theoretic indices defined on distance matrices of graphs using the methods of Bonchev and Trinajstić [41] as well as those of Raychaudhury *et al*. [42], parameters defined on the

neighborhood complexity of vertices in hydrogen–filled molecular graphs [43–45], Balaban's *J* indices [46–48], local orthogonal vertex invariants [37], kappa shape descriptors [49,50], and the electrotopological indices of Kier and Hall [51]. More information on the topological indices calculated by POLLY has been reported in earlier studies [15,20,27,31].

Table 1. Chemicals and Their Physicochemical Property Data for the 166 Identified Components of JP–8

No	Name	log*P*	BP	No	Name	log*P*	BP
1	ISTD (d10–anthracene)	4.49	300	**50**	3,3–dimethylheptane	5.2	137
2	2,2,3–trimethylbutane	4.01	81	**51**	2,4–dimethyl–3–ethylpentane	5.07	137
3	2,3,3–trimethyl–1–butene	3.46	78	**52**	2,3,4–trimethylhexane	5.07	139
4	3,3–dimethylpentane	4.14	86	**53**	2,2,3,3–tetramethylpentane	4.94	140
5	Benzene	2.14	80	**54**	2,3,3,4–tetramethylpentane	4.94	142
6	2–methylhexane	4.27	90	**55**	2,3–dimethylheptane	5.2	141
7	3–ethylpentane	4.27	93.5	**56**	3,4–dimethylheptane	5.2	141
8	*t*–1,3–dimethylcyclopentane	3.83	91	**57**	4–ethylheptane	5.33	141
9	Iso–octane	4.54	99	**58**	Ethylbenzene	3.32	136
10	1–heptene	3.85	94	**59**	4–methyloctane	5.33	141
11	3–heptene	3.85	92.7	**60**	*m*–xylene	3.44	139
12	*n*–heptane	4.4	98	**61**	3–methyloctane	5.33	143
13	2,2–dimethylhexane	4.67	106	**62**	*c*–1,2,3–trimethylcyclohexane	4.91	144
14	1,1,3–trimethylcyclopentane	4.35	105	**63**	3,3–diethylpentane	5.2	146
15	2,3,3–trimethyl–1,4–pentadiene	3.45	125	**64**	1,2,4–trimethylcyclohexane	4.91	142
16	2,4,4–trimethyl–2–pentene	3.99	105	**65**	*c*,*c*,*t*–1,3,5–trimethylcyclohexane	4.91	144
17	2,5–dimethylhexane	4.67	109	**66**	1–nonene	4.91	147
18	2,4–dimethylhexane	4.67	110	**67**	*o*–xylene	3.44	144
19	3,3–dimethylhexane	4.67	112	**68**	4–nonene	4.91	145
20	4–methylcyclohexene	3.33	103	**69**	*n*–nonane	5.46	151
21	*c*,*t*,*c*–1,2,3–trimethylcyclopentane	4.35	123	**70**	*c*,*c*,*t*–1,2,3–trimethylcyclohexane	4.91	144
22	2,3,4–trimethylpentane	4.54	114	**71**	3,3,5–trimethylheptane	5.59	156
23	2,3,3–trimethylpentane	4.54	115	**72**	1–ethyl–1–methylcyclohexane	4.92	144
24	*t*–3,4,4–trimethyl–2–pentene	3.99	119	**73**	1,3,5,5–tetramethyl–1,3–cyclohexadiene	4.52	173
25	1,1,3,3–tetramethylcyclopentane	4.87	114	**74**	t–1,1,3,5–tetramethylcyclohexane	5.43	166
26	2–methylheptane	4.8	118	**75**	Isopropylcyclohexane	4.8	155
27	4–methylheptane	4.8	118	**76**	3,5–dimethyloctane	5.72	160
28	Toluene	2.79	111	**77**	Isopropylbenzene	3.72	152
29	3,4–dimethylhexane	4.67	118	**78**	2,7–dimethyloctane	5.72	160
30	2,2,4,4–tetramethylpentane	4.94	122	**79**	*n*–propylcyclohexane	4.93	157
31	3–methylheptane	4.8	119	**80**	2,6–dimethyloctane	5.72	155
32	3–ethylhexane	4.8	119	**81**	3,4–diethylhexane	5.72	162
33	*t*–1,1,3,4–tetramethylcyclopentane	4.87	144	**82**	3,6–dimethyloctane	5.72	160
34	2–ethyl–1–hexene	4.25	120	**83**	3–ethyl–2–methylheptane	5.72	166
35	2,2,4–trimethylhexane	5.07	127	**84**	3,4,5–trimethylheptane	5.59	164
36	1–ethyl–1–methylcyclopentane	4.36	122	**85**	Propylbenzene	3.85	159
37	*n*–octane	4.93	126	**86**	2,3–dimethyloctane	5.72	164
38	2,4,4–trimethylhexane	5.07	131	**87**	4–ethyloctane	5.85	168
39	2,4–dimethylheptane	5.2	134	**88**	5–methylnonane	5.85	165
40	2,2,3–trimethylhexane	5.07	134	**89**	4–methylnonane	5.85	165
41	4,4–dimethylheptane	5.2	135	**90**	1–ethyl–3–methylbenzene	3.97	161
42	3,3,5–trimethylcyclohexene	4.37	145	**91**	1–ethyl–4–methylbenzene	3.97	162
43	2,2,5,5–tetramethylhexane	5.46	137	**92**	3–ethyloctane	5.85	168
44	2,6–dimethylheptane	5.2	135	**93**	1,3,5–trimethylbenzene	4.09	165
45	*c*,*c*,*c*–1,3,5–trimethylcyclohexane	4.91	144	**94**	3–methylnonane	5.85	167
46	Propylcyclopentane	4.37	131	**95**	1–isopropyl–4–methylcyclohexane	5.32	169
47	1,3,5–trimethylcyclohexane	4.91	144	**96**	1–ethyl–2–methylbenzene	3.97	165
48	3,5,5–trimethylcyclohexene	4.37	145	**97**	2,2,4,6,6–pentamethylheptane	6.39	205
49	Ethylcyclohexane	4.4	132	**98**	*t*–butylbenzene	4.12	169

Table 1. (Continued)

No	Name	log*P*	BP	No	Name	log*P*	BP
99	1,2,4–trimethylbenzene	4.09	169	**133**	1,2–dimethyl–3–ethylbenzene	4.62	194
100	*n*–decane	5.98	174	**134**	1,2,4,5–tetramethylbenzene	4.74	197
101	Isobutylbenzene	4.25	173	**135**	(2–methylbutyl–)–benzene	4.78	205
102	*sec*–butylbenzene	4.25	174	**136**	1,2,3,5–tetramethylbenzene	4.74	198
103	3,7,7–trimethylbicyclo(4.1.0)–3–heptene	4.12	170	**137**	(3–methylbutyl–)–benzene	4.78	199
104	1–isopropyl–3–methylbenzene	4.37	175	**138**	1,2–diisopropylbenzene	5.3	204
105	1,2,3–trimethylbenzene	4.09	176	**139**	1,2,3,4–tetramethylbenzene	4.74	205
106	1–ethyl–2,5–dimethylbenzene	4.62	187	**140**	*n*–pentylbenzene	4.91	205
107	Dicyclopentadiene	3.44	175	**141**	1,4–diisopropylbenzene	5.3	203
108	Butylcyclohexane	5.46	181	**142**	1–t–butyl–3,5–dimethylbenzene	5.42	204
109	Indane (2,3–dihydro–1H–indene)	3.46	176	**143**	Naphthalene	3.32	218
110	1–isopropyl–2–methylbenzene	4.37	178	**144**	1–dodecene	6.5	213
111	1,3–diethylbenzene	4.5	181	**145**	1,3,5–triethylbenzene	5.68	215
112	1–propyl–4–methylbenzene	4.5	183	**146**	*n*–hexylbenzene	5.44	226
113	1,4–diethylbenzene	4.5	183	**147**	(1,1–diethylpropyl–)–benzene	5.71	243
114	Butylbenzene	4.38	183	**148**	2–methylnaphthalene	3.97	241
115	1–ethyl–3,5–dimethylbenzene	4.62	184	**149**	1–methylnaphthalene	3.97	245
116	4–methyldecane	6.38	185	**150**	Cyclohexylbenzene	4.91	235
117	1,2–diethylbenzene	4.5	183	**151**	1-*t*-butyl–3,4,5–trimethylbenzene	6.07	243
118	2–methyldecane	6.38	185	**152**	1,1,6–trimethyltetralin	5.7	247
119	Neopentylbenzene	4.65	186	**153**	*n*–heptylbenzene	5.97	245
120	1–propyl–2–methylbenzene	4.5	185	**154**	1,1'–biphenyl	4.03	254
121	3–methyldecane	6.38	185	**155**	2–ethylnaphthalene	4.49	258
122	1–isopropyl–4–methylbenzene	4.37	177	**156**	1–ethylnaphthalene	4.49	259
123	1–ethyl–2,4–dimethylbenzene	4.62	188	**157**	2,6–dimethylnaphthalene	4.61	262
124	(1,2–dimethylpropyl–)–benzene	4.65	188	**158**	2,3–dimethylnaphthalene	4.61	268
125	1–ethyl–3,4–dimethylbenzene	4.62	190	**159**	1,4–dimethylnaphthalene	4.61	268
126	1–t–butyl–3–methylbenzene	4.77	189	**160**	1,5–dimethylnaphthalene	4.61	265
127	(1–ethylpropyl–)–benzene	4.78	191	**161**	1,2–dimethylnaphthalene	4.61	266
128	1–undecene	5.97	193	**162**	*n*–octylbenzene	6.49	262
129	2–ethyl–1,3–dimethylbenzene	4.62	190	**163**	1,8–dimethylnaphthalene	4.61	270
130	*n*–undecane	6.51	196	**164**	Fluorene	4.23	293
131	1–ethyl–3–isopropylbenzene	4.9	192	**165**	2,5–dimethylheptane	5.2	136
132	*sec*–pentylbenzene	4.78	193	**166**	*p*–xylene	3.44	138

2.2.1 Data reduction

Initially, the TIs were transformed by the natural logarithm of the index plus one. Since the magnitude of some TIs is several orders greater than that of others, re–scaling is conducted to minimize the effect of scale. However, minimal values for some of the Molconn–Z parameters were much less than zero. These indices were logarithmically scaled on a case–by–case basis using the natural logarithm of the index plus *x*, where *x* was an integer large enough to make the minimal value of the index greater than zero. Next, correlation analysis was conducted on the indices. In all cases of a perfect correlation between several indices, only one of the indices was retained within the descriptor set. Additionally, a number of indices encoding features not present in the data set (having zero values for all compounds) were discarded.

2.2.2 Statistical analysis software

Two statistical software packages were used for the construction of similarity spaces used in this study. For the development of the arbitrary similarity space, SAS [52] was used to conduct a principal component analysis (PCA) on the transformed indices to minimize the intercorrelation of indices. This was done using the SAS procedure PRINCOMP. For the tailored spaces, an in–house ridge regression (RR) [53] program was used to select a small set of descriptors for the development of each of the spaces.

2.2.3 Construction of arbitrary similarity spaces

A traditional (arbitrary) molecular similarity space was constructed for the set of 166 JP–8 constituents using the principal components created using the SAS PRINCOMP procedure. Only PCs with eigenvalues greater than or equal to one have been retained for this study. A more detailed explanation of this approach has been provided in a previous study by Basak *et al.* [20]. These PCs were subsequently used as independent variables (in place of the TIs) to determine similarity scores in the Euclidean distance method described later. After the PCA, a correlation analysis was conducted on the PCs to determine which TIs were most highly correlated with each of the PCs. This allows for the creation of similarity spaces based on a small set of TIs (as has been done previously), and also provides some insight into the general nature of the principal components, *i.e.*, which aspects of molecular structure are explained by each of the PCs [6,54,55].

2.2.4 Construction of tailored similarity spaces

Two tailored similarity spaces were constructed for use in this study. One of the spaces was tailored specifically to log *P* and the other for BP. As was mentioned earlier, the RR method was used in the development of these spaces. RR is a method wherein modeling is conducted using the entire set of descriptors retained after the data reduction step as opposed to subset regression. This regression method is useful in cases where the descriptors are highly multicollinear and where the number of descriptors is substantially larger than the number of observations [56]. Conceptually, RR can be thought of as recasting the regression as one using the principal components of the predictor variables as new predictors. It differs in that in principal component regression the leading components are retained and used just as in ordinary least squares regression while the trailing components are dropped. RR retains all components, but weights each of them in accordance with the component's eigenvalue and the 'ridging constant' *k*. More details on the RR method can be found in some of our previous papers [32,57–58].

One of the by–products of the RR is a ranking of the contribution of the indices. The absolute values of this ranking score were used to select the descriptors for use in the development of tailored similarity spaces. Separate RR studies were conducted for log *P* and BP, resulting in a selection of optimal descriptors for use in constructing the tailored similarity spaces.

Table 2. Summary of the First Fifteen Principal Components Derived from a set of 222 Topological Indices Calculated for a Set of 166 JP–8 Constituents

PC	Eigenvalue	Proportion of Explained Variance	Cumulative Explained Variance	First Most Correlated TI		Second Most Correlated TI	
1	93.38	0.421	0.421	DN^2N_4	0.99349	DN^21_4	0.99136
2	45.84	0.206	0.627	Phia	–0.97965	ASN_2	0.96310
3	26.24	0.118	0.745	$^2\chi^b$	0.84243	$^2\chi^v$	0.83715
4	12.74	0.057	0.802	IC_3	0.74736	IC_4	0.73754
5	9.32	0.042	0.844	J^B	–0.64153	$^4\chi^b$	0.61874
6	6.31	0.029	0.873	$^{10}\chi$	0.55023	$^9\chi^v$	0.54724
7	4.75	0.021	0.894	SdsCH	0.53646	Shvin	0.52438
8	3.97	0.018	0.912	$^5\chi^b_C$	0.49541	$^5\chi^v_C$	0.49167
9	2.66	0.012	0.924	$^3\chi_{Ch}$	–0.50749	G_{MAX}	0.48248
10	2.36	0.011	0.935	$^9\chi^v_{Ch}$	–0.64096	$^9\chi_{Ch}$	–0.60901
11	1.91	0.009	0.944	J^B	0.35572	$^8\chi_{Ch}$	0.28873
12	1.68	0.008	0.952	Shvin	–0.39971	SdsCH	–0.39780
13	1.38	0.006	0.958	O_{ORB}	0.36812	O	0.31193
14	1.15	0.005	0.963	$^{10}\chi^v$	0.40648	SdsCH	0.26160
15	1.07	0.005	0.968	$^6\chi^b$	0.31114	$^6\chi^v$	0.30726

Table 3. Fifteen TIs Selected by RR for the 166 JP–8 Chemicals. Indices Common to both RR Sets are Indicated in Bold

PC	TIs from RR for log *P*	(t–value)	TIs from RR for BP	(t–value)
1	$^0\chi^b$	(16.47)	**ANN_5**	(16.77)
2	$^0\chi^v$	(16.42)	**ANN_3**	(16.10)
3	**Fw**	(14.77)	$AN1_3$	(15.81)
4	AZS_1	(14.22)	ANN_1	(15.51)
5	**W**	(14.03)	**W**	(15.30)
6	**ANS_3**	(14.00)	P_0	(15.07)
7	**AZS_3**	(13.31)	**ANS_3**	(14.74)
8	ANS_1	(12.19)	**I^W_D**	(14.58)
9	$^0\chi$	(11.71)	**DN^21_4**	(14.10)
10	ka_1	(11.42)	**AZS_3**	(13.93)
11	**I^W_D**	(11.31)	AZN_3	(13.88)
12	**ANN_3**	(11.31)	AZN_5	(13.13)
13	DN^2S_3	(11.29)	AZN_1	(12.81)
14	**ANN_5**	(11.14)	**Fw**	(12.65)
15	Q_V	(11.08)	DN^2N_3	(12.46)

2.3 Quantification of Intermolecular Similarity

Once the similarity spaces were constructed, it was possible to calculate similarity scores based on the intermolecular distances within the arbitrary and tailored molecular similarity spaces. Intermolecular similarity was measured using Euclidean distance (ED) within an *n*–dimensional space derived from TIs or PCs. The ED between two molecules, *i* and *j*, is defined as:

$$ED_{ij} = \left[\sum_{k=1}^{n} (D_{ik} - D_{jk})^2 \right]^{1/2} \quad (1)$$

where *n* is equal to the number dimensions (descriptors) used to define the similarity space, whether

those dimensions are derived from TIs or PCs. D_{ik} and D_{jk} are the data values of the k^{th} dimension for molecules *i* and *j*, respectively.

Once distances between all molecules within the molecular similarity space have been calculated, these distance "scores" can then be used for analog selection or in *K*NN–based property estimation. This type of quantifiable analog selection can be a powerful tool for finding chemicals that are similar to a chemical of interest, replacing the need for subjective assessment of molecular similarity. More often than not, we are interested in predicting a property of interest. In this case, *K*NN–based similarity offers an alternative to standard linear regression approaches that works well for large, diverse data sets.

*K*NN–based property estimation is carried out by selecting the *k*–nearest neighbors for each compound and using the average of the neighbor's properties as an estimate of the property of our chemical of interest. A number of similar chemicals (k = 1–10, 15, 20, 25) are selected and the property of interest is estimated based on the values of these nearest neighbors. For instance, in estimating the log *P* of the probe compound, the mean log *P* for the *k*–nearest neighbors was used as the estimate. *K*NN estimation was carried out for all chemicals in all three of the similarity spaces, resulting in a full cross–validation. Thus the correlation coefficients reported are the cross–validated correlation coefficients.

3 RESULTS AND DISCUSSION

The principal objective of this paper was to illustrate the utility of tailoring similarity spaces to a specific property as opposed to the standard method of constructing similarity spaces that are property independent. To this end, we used three spaces, *viz*., an arbitrary principal component space that would be used for the *K*NN–based estimation of both log *P* and BP, a topological index space based on the RR weighting of the indices for log *P*, and a topological index space based on the RR weighting of the indices for BP.

From the initial set of 369 topological indices, 222 were retained for inclusion in the PCA and RR procedures after data reduction. From this set of 222 indices, 15 PCs were extracted with eigenvalues greater than or equal to one, resulting in the construction of a 15–dimensional arbitrary similarity space. Table 2 presents a summary of the two TIs most–highly correlated with each of the 15 PCs. For the sake of consistency, it was determined that we would then use the fifteen TIs with the highest rankings from the RR procedure. Table 3 presents the TIs selected for use in developing the similarity spaces tailored for log *P* and BP.

The t–values, indicated in Table 3, are model coefficients extracted from the RR procedure and used to rank the TIs from most to least influential based on the absolute value of the regression coefficient. On close examination of tables 2 and 3 we find that none of the TIs selected by RR are

well represented in the PCs. Only one of the indices chosen for the tailored BP model, $DN^2 1_4$, shows up as the second most–correlated TI in PC_1. Otherwise, the tailored sets have little in common with the TIs selected by RR. Further analysis shows that, of the five TIs most correlated with each of the fifteen PCs, $DN^2 1_4$, is still the only TI shared in common between the arbitrary similarity space and either of the tailored spaces. A much higher degree of overlap exists between the two tailored sets. These sets share a total of seven of the fifteen TIs in common (indicated in bold face in Table 3).

Interestingly, beyond the seven shared indices, each of the tailored sets show a marked difference in the types of indices selected. The set developed for modeling log *P* is skewed towards zero–order chi indices, while the BP set shows a strong tendency towards the AZN triplets. While there is significant overlap between the two tailored sets of descriptors, it is encouraging to see that they show distinct differences as well. It is also encouraging to see the low–degree of overlap between the indices prevalent in the arbitrary set versus those present in the tailored sets. The arbitrary set should be a general characterization of the structural diversity within the data set and while this is useful for property estimation, there is no intrinsic link to any particular property. The tailored sets are geared towards the prediction of a specific property and, as such, should be geared more strongly towards defining the property of interest than simply characterizing the structural diversity of the structure space.

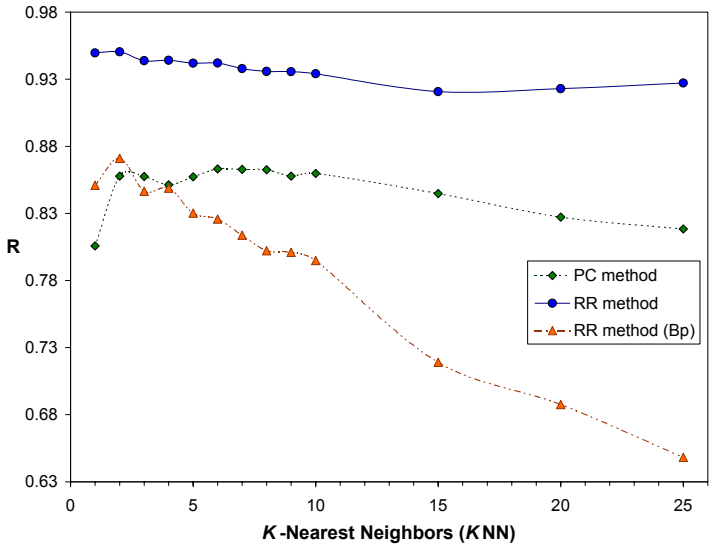

Figure 1. Plot of regression coefficient, *R*, for *K*NN-based estimation of log *P* in arbitrary and tailored similarity spaces at varying levels of *K* (*k* = 1–10, 15, 20 and 25).

Three Euclidean distance–based molecular similarity spaces were constructed from the PCs and TIs indicated in Tables 2 and 3: (*a*) an arbitrary molecular structure space using the fifteen PCs indicated in Table 2, (*b*) a space tailored for log *P* estimation based on the fifteen TIs presented in the second column of Table 3, and (*c*) a space tailored for BP estimation based on the TIs presented

in the third column of Table 3.

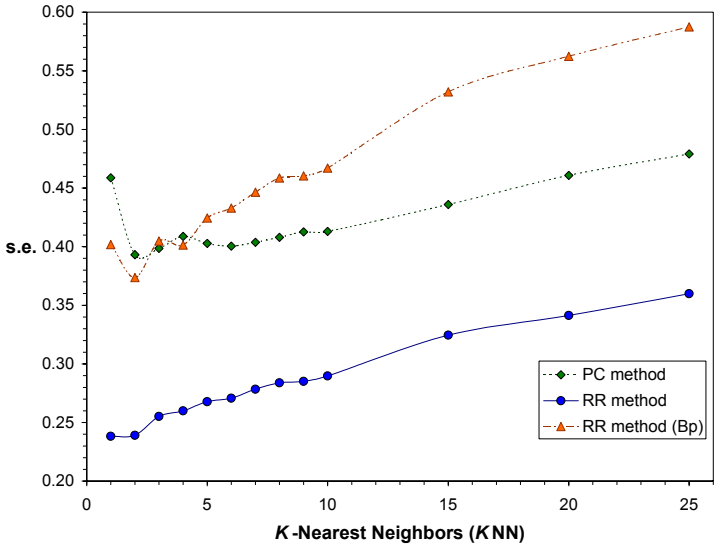

Figure 2. Plot of standard error, *s.e.*, for *K*NN-based estimation of log *P* in arbitrary and tailored similarity spaces at varying levels of *K* (*k* = 1–10, 15, 20 and 25).

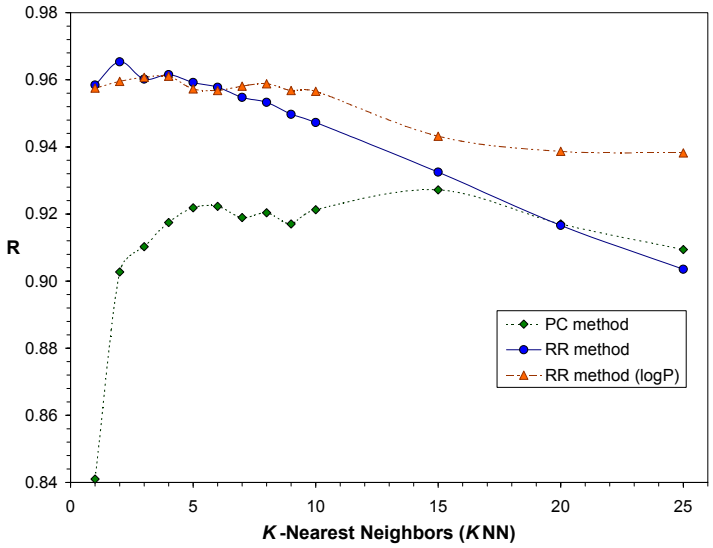

Figure 3. Plot of regression coefficient, *R*, for *K*NN-based estimation of BP in arbitrary and tailored similarity spaces at varying levels of *K* (*k* = 1–10, 15, 20 and 25).

Finally, *K*NN–based property estimation was carried out on the three similarity spaces. First we examined the ability of each of the three spaces to estimate log *P*. In part this was done to verify that the tailored spaces are indeed fitted to the property of interest rather than simply another nonspecific structure space. The results of this analysis are depicted in Figures 1 and 2. Figure 1 presents the correlation coefficients for log *P* estimation in each of the similarity spaces for *K* = 1–10, 15, 20 and 25. Likewise, Figure 2 presents the standard error of log *P* estimation for each of the similarity spaces. As can be seen from these figures, the space tailored to log *P* definitely out–performs both of the other spaces for the purposes of estimating log *P*. As might be expected, the

arbitrary structure space outperforms the BP tailored space in estimating log *P* except for when using the one and two nearest neighbors. So, for the purposes of log *P* prediction, our tailored similarity space meets our expectations in its performance versus the performance of other spaces.

The examination of these structure spaces for the estimation of BP was carried out in a manner identical to that for the estimation of log *P*. Each of the three similarity spaces was used in *K*NN–based estimation of BP for the complete set of 166 chemicals. These results are summarized in Figures 3 and 4. Figure 3 presents the correlation coefficients for BP estimation in each of the similarity spaces for *K* = 1–10, 15, 20 and 25. Likewise, Figure 4 presents the standard error of BP estimation for each of the similarity spaces. As can be seen from these figures, the space tailored to BP definitely out–performs the arbitrary structure space, though, somewhat surprisingly, the space tailored to log *P* performs about as well as the BP tailored space. The BP tailored space just slightly outperforms the log *P* tailored space through *K* = 1–6. However, at higher values of *K*, the log *P* space actually outperforms the BP tailored space for the estimation of BP. While this is interesting, not too much weight should be given to the model's performance at higher values for *K*. As was shown in a recent study [30], loss of data variance is a real concern at the higher values of *K*. Thus we ideally want a model that has a high correlation, *R*, and low standard error, *s.e.*, using a minimal number of neighbors. Taking this into consideration, the two tailored similarity spaces are still essentially identical with regards to the prediction of BP for this particular set of 166 JP–8 components.

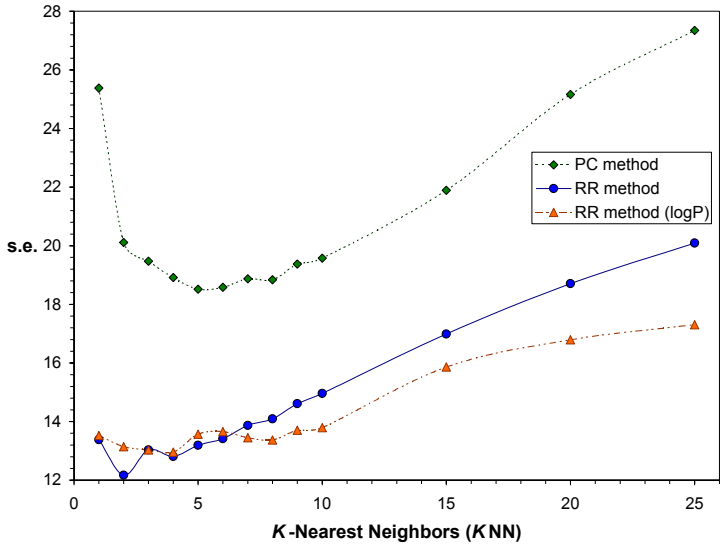

Figure 4. Plot of standard error, *s.e.*, for *K*NN-based estimation of BP in arbitrary and tailored similarity spaces at varying levels of *K* (*k* = 1–10, 15, 20 and 25).

It should be noted that while none of the molecular connectivity indices (chi indices) were selected by the RR method for modeling BP, they have been shown to be related to the modeling of normal boiling point in a number of studies [60–62]. Bearing this mind, we should not be terribly

surprised by the performance of the log *P* tailored space in the estimation of BP. After all, seven of the fifteen indices were shared in common between the two tailored sets, and then if we consider the chi indices as also related to BP, we now see that ten of the fifteen parameters in the log *P* set are also important for the prediction of BP.

4 CONCLUSIONS

As can be seen from the results presented in this study, tailored similarity spaces show definite promise in the development of property–specific similarity spaces, as opposed to standard structure–based similarity spaces. Further studies are needed to verify the general utility of this approach, specifically we need to examine the utility of spaces constructed from smaller "training sets" of chemicals when applied to large, diverse data sets. If these methods can be applied successfully to increase the predictive power of similarity measures for large, diverse data sets, this will become a powerful tool for both risk assessment and pharmaceutical design.

Acknowledgment

The authors acknowledge the financial support of this research by the Air Force Office of Scientific Research grant F496200210138. This is contribution number 318 from the Center for Water and the Environment of the Natural Resources Research Institute.

5 REFERENCES

[1] J. C. Arcos, Structure–Activity Relationships: Criteria for Predicting the Carcinogenic Activity of Chemical Compounds, *Environ*. *Sci*. *Technol*. **1987**, *21*, 743–745.
[2] C. M. Auer, J. V. Nabholz, and K. P. Baetcke, *Environ*. *Health Perspect*. **1990**, *87*, 183–197.
[3] D. M. Sanderson and C. G. Earnshaw, Computer Prediction of Possible Toxic Action from Chemical Structure: The DEREK System, *Human Experimental Toxicol.* **1991**, *10*, 261–273.
[4] J. Ashby and R. W. Tennant, Definitive Relationships among Chemical Structure, Carcinogenicity, and Mutagenicity for 301 Chemicals Tested by the U.S. NTP, *Mutat*. *Res*. **1991**, *257*, 229–306.
[5] M. Lajiness, Molecular Similarity–Based Methods for Selecting Compounds for Screening; in: *Computational Chemical Graph Theory*, Ed. D. H. Rouvray, Nova, New York, 1990, pp 299–316.
[6] S. C. Basak, D. Mills, B. D. Gute, A. T. Balaban, K. Basak, and G. D. Grunwald, Use of Mathematical Structural Invariants in Analyzing Combinatorial Libraries: A Case Study with *Psoralen* Derivatives, in: *Some Aspects of Mathematical Chemistry*, Eds. D. K. Sinha, S. C. Basak, R. K. Mohanty, and I. N. Basumallick, Visva–Bharati University, India, 2002, in press.
[7] M. Johnson, S. C. Basak, and G. Maggiora, A Characterization of Molecular Similarity Methods for a Property Prediction, *Math*. *Comput*. *Modelling* **1988**, *11*, 630–634.
[8] P. Willett and V. Winterman, A Comparison of Some Measures of Inter–Molecular Structural Similarity, *Quant*. *Struct*.*–Act*. *Relat*. **1986**, *5*, 18–25.
[9] P. Willett, J. M. Barnard, and G. M. Downs, Chemical Similarity Searching, *J*. *Chem*. *Inf*. *Comput*. *Sci*. **1998**, *38*, 983–996.
[10] R. E. Carhart, D. H. Smith, and R. Venkataraghavan, Atom Pairs as Molecular Features in Structure–Activity Studies: Definition and Applications, *J*. *Chem*. *Inf*. *Comput*. *Sci*. **1985**, *25*, 64–73.
[11] S. C. Basak and G. D. Grunwald, Molecular Similarity and Risk Assessment: Analog Selection and Property Estimation using Graph Invariants, *SAR QSAR Environ*. *Res*. **1994**, *2*, 289–307.
[12] S. C. Basak and G. D. Grunwald, Use of Topological Space and Property Space in Selecting Structural Analogs, *Mathl*. *Modelling Sci*. *Computing* **1994**, *4*, 464–469.
[13] S. C. Basak, S. Bertelsen, and G. Grunwald, Application of Graph Theoretical Parameters in Quantifying Molecular Similarity and Structure–Activity Studies, *J*. *Chem*. *Inf*. *Comput*. *Sci*. **1994**, *34*, 270–276.

[14] S. C. Basak, S. Bertelsen, and G. D. Grunwald, Use of Graph Theoretic Parameters in Risk Assessment of Chemicals, *Toxicol. Lett*. **1995**, *79*, 239–250.
[15] S. C. Basak, B. D. Gute, and G. D. Grunwald, Use of graph invariants in QMSA and predictive toxicology, in: *Discrete Mathematical Chemistry*, Eds. P. Hansen, P. Fowler and M. Zheng, DIMACS Series 51, American Mathematical Society, Providence, Rhode Island, 2000, pp 9–24.
[16] S. C. Basak, B. D. Gute, and G. D. Grunwald, Development and Application of Molecular Similarity Methods using Nonempirical Parameters, *Mathl. Modelling Sci. Computing* **2002**, in press.
[17] M. Randić, Similarity Methods of Interest in Chemistry; in: *Mathematical Methods in Contemporary Chemistry*, Ed. S. I. Kuchanov, Gordon and Breach, Amsterdam, 1996, pp 1–100.
[18] G. M. Downs, P. Willett, and W. Fisanick, Similarity Searching and Clustering of Chemical–Structure Databases using Molecular Property Data, *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 1094–1102.
[19] B. D. Gute, G. D. Grunwald, D. Mills, and S. C. Basak, Molecular Similarity Based Estimation of Properties: A Comparison of Structure Spaces and Property Spaces, *SAR QSAR Environ. Res*. **2001**, *11*, 363–382.
[20] S. C. Basak, V. R. Magnuson, G. J. Niemi, and R. R. Regal, Determining Structural Similarity of Chemicals using Graph–Theoretic Indices, *Discrete Appl. Math*. **1988**, *19*, 17–44.
[21] S. C. Basak and G. D. Grunwald, Estimation of Lipophilicity from Molecular Structural Similarity, *New J. Chem*. **1995**, *19*, 231–237.
[22] S. C. Basak and G. D. Grunwald, Molecular similarity and estimation of molecular properties, *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 366–372.
[23] S. C. Basak and G. D. Grunwald, Predicting Mutagenicity of Chemicals using Topological and Quantum Chemical Parameters: A Similarity Based Study, *Chemosphere* **1995**, *31*, 2529–2546.
[24] S. C. Basak and G. D. Grunwald, Tolerance Space and Molecular Similarity, *SAR QSAR Environ. Res*. **1995**, *3*, 265–277.
[25] S. C. Basak, B. D. Gute, and G. D. Grunwald, Estimation of Normal Boiling Points of Haloalkanes using Molecular Similarity, *Croat. Chim. Acta* **1996**, *69*, 1159–1173.
[26] S. C. Basak and B. D. Gute, Use of Graph Theoretic Parameters in Predicting Inhibition of Microsomal Hydroxylation of Anilines by Alcohols: A Molecular Similarity Approach, in *Proceedings of the International Congress on Hazardous Waste: Impact on Human and Ecological Health*, Eds. B. L. Johnson, C. Xintaras, and J. S. Andrews, Princeton Scientific Publishing Co, Inc, 1997, pp 492–504.
[27] S. C. Basak, B. D. Gute, and G. D. Grunwald, Characterization of the Molecular Similarity of Chemicals using Topological Invariants; in: *Advances in Molecular Similarity*, Eds. R. Carbo–Dorca and P. G. Mezey, JAI Press, Stanford, Connecticut, Vol. 2, 1998, pp. 171–185.
[28] S. C. Basak, G. D. Grunwald, G. E. Host, G. J. Niemi, and S. P. Bradbury, A Comparative Study of Molecular Similarity, Statistical and Neural Network Methods for Predicting Toxic Modes of Action of Chemicals, *Environ. Toxicol. Chem*. **1998**, *17*, 1056–1064.
[29] S. C. Basak, B. D. Gute, and G. D. Grunwald, Assessment of Mutagenicity of Chemicals from Theoretical Structural Parameters: A Hierarchical Approach, *SAR QSAR Environ. Res*. **1999**, *10*, 117–129.
[30] B. D. Gute and S. C. Basak, Molecular similarity–based estimation of properties: A comparison of three structure spaces, *J. Mol. Graphics Modelling* **2001**, *20*, 95–109.
[31] S. C. Basak, D. Mills, B. D. Gute, G. D. Grunwald, and A. T. Balaban, Applications of topological indices in predicting property/ bioactivity/ toxicity of chemicals; in: *Topology in Chemistry: Discrete Mathematics of Molecules*, Eds. D. H. Rouvray and R. B. King, Horwood Publishing Ltd., Chichester, United Kingdom, 2001.
[32] S. C. Basak, B. D. Gute, D. Mills, and D. M. Hawkins, Quantitative Molecular Similarity Methods in the Property/ Toxicity Estimation of Chemicals: A Comparison of Arbitrary versus Tailored Similarity Spaces, *J. Mol. Struct*. (*Theochem*) **2002**, accepted.
[33] C. L. Russom, E. B. Anderson, B. E. Greenwood, and A. Pilli, ASTER: An Integration of the AQUIRE Data Base and the QSAR System for use in Ecological Risk Assessments. *Sci. Total Environ*. **1991**, *109/110*, 667–670.
[34] AFOSR JP–8 Jet Fuel Workshop, University of Arizona, Tucson, Arizona, Jan. 11–12, 2000.
[35] S. C. Basak, D. K. Harriss, and V. R. Magnuson, POLLY 2.3: Copyright of the University of Minnesota, 1988.
[36] Molconn–Z v3.50, Hall and Associates Consulting, Quincy, MA, 2000.
[37] P. A. Filip, T. S. Balaban, and A. T. Balaban, 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.
[38] H. Wiener, Structural Determination of Paraffin Boiling Points, *J. Am. Chem. Soc*. **1947**, *69*, 17–20.
[39] M. Randić, On Characterization of Molecular Branching, *J. Am. Chem. Soc*. **1975**, *97*, 6609–6615.
[40] L. B. Kier and L. H. Hall, *Molecular Connectivity in Structure–Activity Analysis*, Research Studies Press, Letchworth, 1986.
[41] D. Bonchev and N. Trinajstić, Information Theory, Distance Matrix and Molecular Branching, *J. Chem. Phys*. **1977**, *67*, 4517–4533.
[42] C. Raychaudhury, S. K. Ray, J. J. Ghosh, A. B. Roy, and S. C. Basak, Discrimination of Isomeric Structures using

Information Theoretic Topological Indices, *J. Comput. Chem.* **1984**, *5*, 581–588.
[43] S. C. Basak, A. B. Roy, and J. J. Ghosh, Study of the Structure–Function Relationship of Pharmacological and Toxicological Agents using Information Theory; in: *Proceedings of the 2nd International Conference on Mathematical Modelling*, Eds. X. J. R. Avula, R. Bellman, Y. L. Luke, and A. K. Rigler, University of Missouri–Rolla, Rolla, Missouri, Vol. II, 1980, pp 851–856.
[44] S. C. Basak and V. R. Magnuson, Molecular Topology and Narcosis: A Quantitative Structure–Activity Relationship (QSAR) Study of Alcohols using Complementary Information Content (CIC), *Arzneim. Forsch.* **1983**, *33*, 501–503.
[45] A. B. Roy, S. C. Basak, D. K. Harriss, and V. R. Magnuson, Neighborhood Complexities and Symmetry of Chemical Graphs and Their Biological Applications; in: *Mathematical Modelling in Science and Technology*, Eds. X. J. R. Avula, R. E. Kalman, A. I. Lipais, and E. Y. Rodin, Pergamon Press, New York, 1984, pp 745–750.
[46] A. T. Balaban, Highly Discriminating Distance–Based Topological Index, *Chem. Phys. Lett.* **1982**, *89*, 399–404.
[47] A. T. Balaban, Topological Indices based on Topological Distances in Molecular Graphs, *Pure Appl. Chem.* **1983**, *55*, 199–206.
[48] A. T. Balaban, Chemical Graphs. Part 48. Topological Index J for Heteroatom–Containing Molecules taking into Account Periodicities of Element Properties, *MATCH* (*Commun. Math. Chem.*) **1986**, *21*, 115–122.
[49] L. B. Kier, A Shape Index from Molecular Graphs, *Quant. Struct.–Act. Relat.* **1985**, *4*, 109–116.
[50] L. B. Kier and L. H. Hall, The Kappa Indices for Modeling Molecular Shape and Flexibility; in: *Topological Indices and Related Descriptors in QSAR and QSPR*, Eds. J. Devillers and A. T. Balaban, Gordon and Breach Science Publishers, Amsterdam, 1999, 455–489.
[51] L. B. Kier and L. H. Hall, The Electrotopological State: Structure Modeling Molecular for QSAR and Database Analysis; in: *Topological Indices and Related Descriptors in QSAR and QSPR*, Eds. J. Devillers and A. T. Balaban, Gordon and Breach Science Publishers, Amsterdam, 1999, 491–562.
[52] SAS Institute Inc., in SAS/STAT User's Guide, Release 6.03 Edition, SAS Institute Inc., Cary, NC, 1988.
[53] A. E. Hoerl and R. W. Kennard, Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics* **1970**, *8*, 27–51.
[54] S. C. Basak, G. J. Niemi, and G. D. Veith, Optimal Characterization of Structure for Prediction of Properties, *J. Math. Chem.* **1991**, *7*, 243–272.
[55] S. C. Basak and B. D. Gute, Characterization of Molecular Structures using Topological Indices, *SAR QSAR Environ. Res.* **1997**, *7*, 1–21.
[56] D. Hawkins, S. Basak, and X. Shi, QSAR with Few Compounds and Many Features, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 663–670.
[57] S. C. Basak, D. M. Hawkins, and D. Mills, Predicting Blood:Air Partition Coefficient of Structurally Diverse Chemicals using Theoretical Molecular Descriptors; in: *Advances in Molecular Similarity*, Eds. R. Carbo–Dorca and P.G. Mezey, Vol. 5, 2002, in press.
[58] S. C. Basak, H. El–Masri, D. M. Hawkins, and D. Mills, Exposure Assessment of Volatile Organic Chemicals (VOCs): Predicting Blood:Air Partition Coefficients of Diverse Chemicals using Theoretical Descriptors, *J. Chem. Inf. Comput. Sci.* **2002**, submitted.
[59] S. C. Basak, D. Mills, D. M. Hawkins, and H. El–Masri, Prediction of Human Blood:Air Partition Coefficient: A Comparison of Structure–based and Property–based Methods, *Risk Analysis* **2002**, submitted.
[60] L. B. Kier and L. H. Hall, *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York, 1976.
[61] D. E. Needham, I.–C. Wei, and P. G. Seybold, Molecular Modeling of the Physical Properties of the Alkanes, *J. Am. Chem. Soc.* **1992**, *110*, 4186–4194.
[62] S. C. Basak, B. D. Gute, and G. D. Grunwald, A Comparative Study of Topological and Geometrical Parameters in Estimating Normal Boiling Point and Octanol/Water Partition Coefficient, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1054–1060.