

Relative Eccentric Distance Sum/Product Indices for QSAR/QSPR: Development, Evaluation, and Application

Monika Gupta,[†] Harish Jangra,[‡] P. V. Bharatam,[§] and A. K. Madan^{*,||}

[†]Faculty of Pharmaceutical Sciences, M. D. University, Rohtak 124 001, India

[‡]Department of Pharmacoinformatics, National Institute of Pharmaceutical Education and Research, Sahibzada Ajit Singh Nagar (Mohali) 160062, India

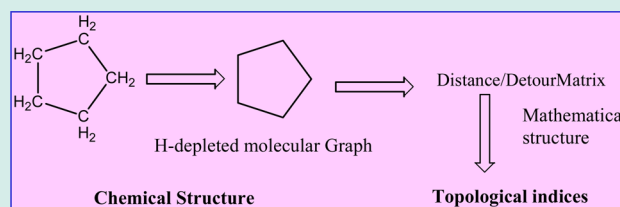
[§]Department of Medicinal Chemistry, National Institute of Pharmaceutical Education and Research, Sahibzada Ajit Singh Nagar (Mohali) 160062, India

^{||}Faculty of Pharmaceutical Sciences, Pandit Bhagwat Dayal Sharma University of Health Sciences, Rohtak 124 001, India

S Supporting Information

ABSTRACT: In the present study, five detour/distance matrix based molecular descriptors (MDs) termed as relative eccentric distance sum/product indices (denoted by $R_{\xi_1}^{eSV}$, $R_{\xi_2}^{eSV}$, $R_{\xi_3}^{eSV}$, $RP_{\xi_1}^{eSV}$, and $RP_{\xi_2}^{eSV}$), as well as their topochemical versions denoted by ($R_{\xi_1}^{ecSV}$, $R_{\xi_2}^{ecSV}$, $R_{\xi_3}^{ecSV}$, $RP_{\xi_1}^{ecSV}$, and $RP_{\xi_2}^{ecSV}$) have been conceptualized for exclusive use for molecules containing cyclic moieties. The said MDs exhibited exceptionally high discriminating power and high sensitivity toward branching/relative position of substituents in cyclic structures amalgamated with negligible degeneracy. Subsequently, the proposed MDs along with other MDs were successfully utilized for the development of models for the prediction of human glutaminyl cyclase (hQC) inhibitory activity using decision tree (DT), random forest (RF) and moving average analysis (MAA). A data set comprising of 45 analogues of substituted 3-(1H-imidazol-1-yl) propyl thiourea derivatives was used. DT identified proposed relative eccentric distance sum topochemical index-1 as the most important MD. High accuracy of prediction up to 96%, 93%, and 95% was observed in case of models derived from decision tree, random forest, and MAA, respectively. The statistical significance of proposed models was assessed through specificity, sensitivity, overall accuracy, Mathew's correlation coefficient (MCC), and intercorrelation analysis.

KEYWORDS: hQC inhibitors, 3-imidazolyl propyl thiourea, relative eccentric distance product indices and relative eccentric distance sum indices, combinatorial library, virtual screening



INTRODUCTION

Finding new drugs is a highly complex, expensive, and time-consuming task, as there is no single systematic way to automatically discover a drug even when the disease, targets, and molecular mechanisms of drug activity are well understood.¹ Many investigators and drug companies therefore resort to computer-aided drug design (CADD) technologies because of their high efficiency and versatility in the design of new drugs, thereby saving time and money.² One of the major goals of computer-aided drug design and discovery strategies is the identification of new lead chemical compounds.³ These approaches include (quantitative) structure–activity relationship [(Q)SAR] modeling techniques. A (Q)SAR is essentially a mathematical equation/model that is determined from a set of molecules with known activities using computational approaches. The exact form of the relationship between structure and activity can be determined using a variety of statistical methods and molecular descriptors and the resulting model is then used to predict the activity of new molecules.⁴

One major emphasis in the (Q)SAR methodology is the development of easily calculable parameters, which are available for any arbitrary structure. A large number of constitutional,

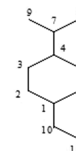
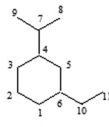
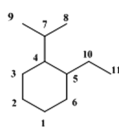
topological, geometric, electrostatic, and quantum chemical molecular descriptors have been introduced in theoretical chemistry with the objective of expressing chemical structures in a numerical form. Such structural descriptors can be used to model physical, chemical, or biological properties.⁵ It has been well recognized for some time that a success or a failure of a structure–property–activity study often critically depends on the selection of MDs. Hence, it is not surprising to see continual development of novel MDs.⁶

Molecular topological factors that taken into account the arrangements of atoms across the parent molecular skeleton, concepts of steric relations and molecular bulk, branchedness, and relationships among various nonbonded parts of the molecule would be useful to better understand relationships between molecular structure and their empirical properties.^{2,7} Consequently MDs based on molecular topology have emerged as molecular descriptors of choice in structure–activity/property relationship studies and rational drug design. They are in

Received: March 19, 2013

Revised: December 30, 2013

Published: January 31, 2014



Distance Matrix (D)

	1	2	3	4	5	6	7	8	9	10	11	E	S _i	S*E _i	
1	0	1	1	2	2	3	3	4	5	5	3	4	5	30	150
2	1	0	1	1	2	2	3	3	4	4	2	3	4	25	100
3	1	0	1	1	2	2	3	3	3	3	1	2	3	20	60
4	2	1	2	0	1	1	2	2	2	2	3	3	3	19	57
5	2	1	2	2	0	1	1	1	1	1	3	3	4	24	96
6	1	1	1	1	1	0	1	1	1	1	3	4	4	5	145
7	3	3	3	2	2	3	0	1	1	1	3	4	4	24	96
8	4	4	4	3	3	3	1	0	1	1	4	5	5	33	165
9	5	4	4	3	3	3	1	0	1	1	4	5	5	33	165
10	3	2	1	2	3	4	3	4	0	1	4	4	4	27	108
11	4	3	2	3	3	4	4	5	1	0	5	5	5	36	180

	1	2	3	4	5	6	7	8	9	10	11	E	S _i	S*E _i	
1	0	1	1	2	2	3	3	4	5	5	2	3	5	28	140
2	1	0	2	1	1	3	2	3	4	4	1	2	4	23	92
3	1	2	0	3	1	2	3	4	4	3	4	4	4	27	108
4	2	1	3	0	2	1	2	3	3	2	3	3	3	22	86
5	2	3	1	2	0	1	2	3	3	4	5	5	5	26	130
6	3	3	2	1	1	0	1	2	2	3	4	4	4	21	84
7	3	2	3	2	2	2	1	1	0	1	1	4	5	26	130
8	4	4	4	3	3	3	2	1	0	2	5	6	6	35	210
9	5	4	4	3	3	3	2	1	0	2	5	6	6	35	210
10	3	2	1	3	2	4	3	4	5	0	1	5	30	150	
11	4	3	2	4	3	5	4	5	6	1	0	6	39	234	

	1	2	3	4	5	6	7	8	9	10	11	E	S _i	S*E _i	
1	0	1	1	2	2	3	3	4	5	5	1	2	5	26	130
2	1	0	2	1	1	3	2	3	4	4	2	3	4	25	100
3	1	2	0	3	1	2	3	4	4	2	3	4	4	25	100
4	2	1	3	0	2	1	2	3	3	3	4	4	4	24	96
5	2	3	1	2	0	1	2	3	3	4	5	5	5	24	96
6	3	3	2	2	1	1	0	1	2	2	4	4	5	23	115
7	3	2	3	2	2	2	1	0	1	1	5	6	6	28	168
8	4	4	4	3	3	3	2	1	0	2	6	7	7	37	259
9	5	4	4	3	3	3	2	1	0	2	6	7	7	37	259
10	3	2	2	3	3	4	3	4	5	0	1	6	33	198	
11	4	3	3	4	4	5	6	7	7	1	0	7	42	294	

Detour Matrix (Δ)

	1	2	3	4	5	6	7	8	9	10	11	Δ _{ij}	α _i	α _i * Δ _{ij}
1	0	5	4	3	4	5	4	5	5	6	6	46	6	276
2	5	0	5	4	3	4	5	6	6	6	7	7	7	357
3	4	5	0	5	4	3	6	7	7	1	2	7	44	308
4	3	4	5	0	5	4	1	2	2	6	7	7	39	273
5	4	3	4	5	0	5	6	7	7	5	6	6	47	282
6	5	4	3	4	5	0	5	6	6	4	5	6	47	282
7	4	5	6	1	6	5	0	1	1	7	8	8	44	352
8	5	6	7	2	7	6	1	0	2	8	9	9	53	477
9	5	6	7	2	7	6	1	0	2	8	9	9	53	477
10	6	6	1	6	5	4	7	8	8	0	1	8	51	408
11	6	7	2	7	6	5	8	9	1	0	9	60	60	540

	1	2	3	4	5	6	7	8	9	10	11	Δ _{ij}	α _i	α _i * Δ _{ij}
1	0	5	5	4	4	3	4	5	6	7	7	48	7	336
2	5	0	4	5	3	4	5	6	6	1	2	6	41	246
3	5	4	0	3	5	4	5	6	6	5	6	6	49	294
4	4	5	3	0	4	5	6	7	7	6	7	7	54	378
5	4	3	5	4	0	5	6	7	7	4	5	7	50	350
6	3	4	4	5	5	0	1	2	2	5	6	6	37	222
7	4	5	5	6	6	1	0	1	6	7	7	42	294	
8	5	6	6	7	7	2	1	0	2	7	8	8	51	408
9	5	6	6	7	7	2	1	0	2	7	8	8	51	408
10	6	1	5	6	4	5	6	7	7	0	1	7	48	336
11	7	2	6	7	5	6	7	8	8	1	0	8	57	456

	1	2	3	4	5	6	7	8	9	10	11	Δ _{ij}	α _i	α _i * Δ _{ij}	
1	0	5	5	4	4	3	4	5	5	1	2	5	5	38	190
2	5	0	4	5	3	4	5	6	6	6	7	7	7	51	357
3	5	4	0	3	5	4	5	6	6	6	7	7	7	51	357
4	4	5	3	0	4	5	6	7	7	5	6	7	7	52	364
5	4	3	5	4	0	5	6	7	7	5	6	7	7	52	364
6	3	4	4	5	5	0	1	2	2	4	5	5	5	35	175
7	4	5	5	6	6	1	0	1	1	5	6	6	6	40	240
8	5	6	6	7	7	2	1	0	2	6	7	7	7	49	343
9	5	6	6	7	7	2	1	0	2	6	7	7	7	49	343
10	1	6	6	5	5	4	5	6	6	0	1	6	45	270	
11	2	7	7	6	6	5	6	7	7	1	0	7	54	378	

$$RP_{\xi_1}^{ecSV} = \prod_{i=1}^n \left(\frac{\sigma_i * \eta_i}{S_i * E_i} \right)^{1/2} = 643.12 = 162.88 = 42.31$$

$$RP_{\xi_2}^{ecSV} = \frac{1}{k_m} \prod_{i=1}^n \frac{\sigma_i * \eta_i}{S_i * E_i} \quad \text{Where } k_m=100 = 4136.02 = 265.28 = 17.9$$

$$R_{\xi_1}^{ecSV} = \sum_{i=1}^n \frac{\sigma_i * \eta_i}{S_i * E_i} = 37.295 = 29.195 = 24.43$$

$$R_{\xi_2}^{ecSV} = \frac{1}{k_m} \sum_{i=1}^n \left(\frac{\sigma_i * \eta_i}{S_i * E_i} \right)^2 \quad \text{Where } k_m=10 = 13.70 = 8.88 = 6.78$$

$$R_{\xi_3}^{ecSV} = \frac{1}{k_m} \sum_{i=1}^n \left(\frac{\sigma_i * \eta_i}{S_i * E_i} \right)^3 \quad \text{Where } k_m=100 = 5.37 = 3.24 = 2.19$$

Figure 1. Calculation of the values of the relative eccentric distance product indices ($RP_{\xi_1}^{ecSV}$ and $RP_{\xi_2}^{ecSV}$) and relative eccentric distance sum indices ($R_{\xi_1}^{ecSV}$, $R_{\xi_2}^{ecSV}$, and $R_{\xi_3}^{ecSV}$) for three isomers of ethyl isopropyl cyclohexane.























particular inescapable in the development of successful (Q)SAR models as well as in screening combinatorial libraries.⁸ The utilization of topological indices (TIs) in combinatorial chemistry⁹ has been extensively reviewed by Bereg.¹⁰ TIs are relatively simpler and facilitate rapid calculation. Moreover, TIs encode useful information about various aspects of molecular architecture (size, shape, branching, and cyclicality). A number of TIs have been proposed so far. The calculation of these TIs is well documented in the literature.¹¹

The purpose of defining a TI is to represent each chemical structure with a numerical value, keeping it as discriminatory as possible. TIs are sometimes criticized because the physicochemical meaning of a topological index is not explicit as compared with other parameters. A thorough and conclusive interpretation of TIs is indeed not simple, requiring the total dissection of the calculation formula, as well as some knowledge of underlying theory and property modeling.¹² TIs also exhibit considerable mutual correlation, which can be a major problem when performing structure–activity studies because the employed statistical methods may fail or give little meaningful correlation of biological, chemical, or physical properties of molecules.¹³ The interest in developing new TIs for organic molecules has revived in recent years, as TIs have found new applications in similarity and diversity assessment, database mining, and virtual screening of combinatorial libraries.^{14–16} TIs can be classified on the basis of the type of matrix used, such as adjacency, distance, adjacency-cum-distance, centrality, and information content.

Most of the well-known TIs are derived from distance matrices to characterize molecular graphs. In contrast, there exist only a handful of TIs based upon detour matrices, and so *there exists a vast potential in utilizing detour matrices for developing novel TIs.*¹⁷

Alzheimer's disease is the most common cause of dementia, representing around 50–80% of all cases.¹⁸ The modified amyloid hypothesis suggested that causative event of AD pathology is the deposition of amyloid fibril β ($A\beta$) leading to formation of the neurofibrillary tangles, loss of neurons, vascular damage, and consequently dementia. The major fraction of $A\beta$ is N-terminally truncated possessing a glutamine that can subsequently be cyclized into pyroglutamate (pE).¹⁹ The cyclization renders the peptide lysosomal proteases and aminopeptidases resistant,²⁰ more prone to aggregation, increases its hydrophobicity,²¹ and neurotoxicity.²² The enzyme that catalyzes this conversion of glutamine to pE is glutaminyl cyclase (QC).²³ An important event in the pathogenesis of AD patients' brains is the increased expression of QC in the earliest stages of pathology.²⁴ The application of inhibitors of QC as a new strategy for the treatment of AD has proven to be successful in different transgenic animal models.²⁵ Unfortunately, many recent clinical trials assessing the efficacy of novel therapeutic agents for the treatment of AD have failed, highlighting the continued importance of novel scaffolds in AD research. Therefore, QC inhibitors may achieve the goal of preventing AD development or progression.²⁶

Table 1. Index Values of Relative Eccentric Distance Product Indices (${}^{RP}\xi_1^{SV}$ and ${}^{RP}\xi_2^{SV}$) and Relative Eccentric Distance Sum Indices (${}^{R}\xi_1^{SV}$, ${}^{R}\xi_2^{SV}$, ${}^{R}\xi_3^{SV}$) for All Possible Cyclic Structures Containing Four and Five Vertices

Compound number	Structure	${}^{RP}\xi_1^{SV}$	${}^{RP}\xi_2^{SV}$	${}^{R}\xi_1^{SV}$	${}^{R}\xi_2^{SV}$	${}^{R}\xi_3^{SV}$
1		6.95	0.48	10.68	2.93	0.83
2		9	0.81	12	3.6	1.08
3		27	7.29	22.75	15.08	11.01
4		81	65.61	36	32.4	29.16
5		4.48	0.2	9.22	1.74	0.34
6		6.96	4.84	10.97	2.45	0.56
7		8.35	0.69	11.86	2.89	0.73
8		32	10.24	20	8	3.2
9		11.65	1.36	14.27	4.58	1.59
10		33.27	11.07	20.6	8.71	3.77
11		37.87	14.34	22.15	10.58	5.47
12		19.26	3.7	17.24	6.59	2.74
13		62.59	39.17	26.81	15.12	8.95
14		83.17	69.17	30.5	20.15	14.28
15		47.05	22.13	23.33	10.89	5.08
16		64.66	41.81	26.53	14.12	7.53
17		88.68	78.64	30.13	18.24	11.08
18		119.7	143.36	36.67	32.48	34.79
19		163.8	268.44	41.6	41.98	51.45
20		221.7	491.52	48.13	56.036	74.26
21		409.6	1677.72	60.8	84.99	128.1
22		1024	10485.76	80	128	204.8

In the present investigation, five highly discriminating detour/distance matrix based MDs termed as *relative eccentric distance sum indices* (denoted by ${}^{R}\xi_1^{SV}$, ${}^{R}\xi_2^{SV}$, ${}^{R}\xi_3^{SV}$) and *relative eccentric distance product indices* (denoted by ${}^{RP}\xi_1^{SV}$, ${}^{RP}\xi_2^{SV}$), as well as their topochemical versions (denoted by ${}^{RP}\xi_1^{cSV}$, ${}^{RP}\xi_2^{cSV}$, ${}^{R}\xi_1^{cSV}$, ${}^{R}\xi_2^{cSV}$, and ${}^{R}\xi_3^{cSV}$) have been conceptualized. The proposed MDs along with diverse 2D and 3D MDs were successfully utilized through random forest, decision tree, and moving average analysis to build suitable models for the prediction of hQC inhibitory activity of substituted 3-(1H-imidazol-1-yl) propyl thiourea derivatives.²⁷

RESULTS AND DISCUSSION

In the present study, five novel detour cum distance matrix based indices as well as their topochemical counterparts have been conceptualized. The topostructural versions of these MDs have

been calculated from a detour matrix (Δ) and distance matrix (D), whereas the topochemical versions have been calculated using chemical detour matrix (Δ_c) and chemical distance matrix (D_c).

Evaluation of Eccentric Distance Sum/Product Indices.

The eccentric distance sum is the summation of product of eccentricity and distance sum of each vertex in the hydrogen suppressed molecular graph.²⁸ The ratio of both product of maximum path distance and path eccentricity on one hand and the distance sum and eccentricity on the other hand augmented the sensitivity of proposed MDs.

Evaluation for the Sensitivity toward Relative Position of Substituents.

As observed from Figure 1, simple change in the position of ethyl group from ortho to either meta or para leads to steep change in index values of proposed MDs. In case of the relative eccentric distance product index 1 (${}^{RP}\xi_1^{SV}$), the value changes from 643.12 to 42.31 as ethyl substituent is simply

shifted from ortho to para position. The index value also changes from 643.12 to 162.88 as ethyl substituent is shifted from ortho to meta position. In case of relative *t* eccentric distance product index 2 ($RP_{\xi_2}^{SV}$) the index value changes from 4136.02 to 17.9 as ethyl substituent is shifted from ortho to para position. The index value also changes from 4136.02 to 265.28 as ethyl substituent is shifted from ortho to meta position. In case of relative eccentric distance sum index 1 ($R_{\xi_1}^{SV}$) the index value changes from 37.295 to 24.43 as ethyl substituent is shifted from ortho to para position. In case of relative eccentric distance sum index 2 ($R_{\xi_2}^{SV}$), the index value changes from 13.70 to 6.78 as ethyl substituent is shifted from ortho to para position. In case of the relative distance sum index 3 ($R_{\xi_3}^{SV}$), the value changes by two times from 5.37 to 2.19 as ethyl substituent is shifted from ortho to para position. The index value also changes from 5.37 to 3.24 as ethyl substituent is shifted from ortho to meta position. This major change in the index value without changing number of vertices reveals exceptionally high sensitivity of proposed indices.

Evaluation of Eccentric Distance Sum/Product Indices for the Discriminating Power. The ratio of the highest to lowest value for all possible structures with the same number of vertices is called as discriminating power. The discriminating power is one of important characteristics of a TI as mentioned previously. As observed from Tables 1 and 2, the ratio of the highest to lowest value for all possible structures containing five vertices for $RP_{\xi_1}^{SV}$, $RP_{\xi_2}^{SV}$, $R_{\xi_1}^{SV}$, $R_{\xi_2}^{SV}$, and $R_{\xi_3}^{SV}$ is 229, 52429, 9, 74,

Table 2. Comparison of the Discriminating Power and Degeneracy of Relative Eccentric Distance Product Indices ($RP_{\xi_1}^{SV}$, $RP_{\xi_2}^{SV}$) and Relative Eccentric Distance Sum Indices ($R_{\xi_1}^{SV}$, $R_{\xi_2}^{SV}$, $R_{\xi_3}^{SV}$) Using All Possible Cyclic Structures Containing Four and Five Vertices^a

	$RP_{\xi_1}^{SV}$	$RP_{\xi_2}^{SV}$	$R_{\xi_1}^{SV}$	$R_{\xi_2}^{SV}$	$R_{\xi_3}^{SV}$
for four vertices					
minimum value	6.95	0.48	10.68	2.93	0.83
maximum value	81	65.61	36	32.4	29.16
ratio	1:12	1:136	1:3	1:11	1:35
degeneracy	0/4	0/4	0/4	0/4	0/4
for five vertices					
minimum value	4.48	0.2	9.22	1.74	0.33
maximum value	1024	10485.76	80	128	205
ratio	1:229	1:52428	1:9	1:74	1:621
degeneracy	0/18	0/18	0/18	0/18	0/18

^aDegeneracy: Number of compounds having same values/total number of compounds with same number of vertices.

Table 3. Intercorrelation Matrix^a

	χ	ξ^c	M_1	M_2	W	D	$RP_{\xi_1}^{SV}$	$RP_{\xi_2}^{SV}$	$R_{\xi_1}^{SV}$	$R_{\xi_2}^{SV}$	$R_{\xi_3}^{SV}$
χ	1	0.91	0.49	0.52	0.78	0.37	0.35	0.24	0.37	0.28	0.24
ξ^c		1	0.49	0.57	0.71	0.2	0.34	0.23	0.34	0.26	0.23
M_1			1	0.97	-0.02	-0.3	0.88	0.72	0.91	0.88	0.82
M_2				1	0.002	-0.3	0.88	0.74	0.88	0.87	0.82
W					1	0.56	-0.17	-0.14	-0.22	-0.25	-0.23
D						1	-0.35	-0.34	-0.34	-0.4	-0.4
$RP_{\xi_1}^{SV}$							1	0.91	0.94	0.98	0.97
$RP_{\xi_2}^{SV}$								1	0.75	0.87	0.94
$R_{\xi_1}^{SV}$									1	0.95	0.88
$R_{\xi_2}^{SV}$										1	0.97
$R_{\xi_3}^{SV}$											1

^a χ = molecular connectivity index; ξ^c = eccentric connectivity index; M_1 = Zagreb index 1; M_2 = Zagreb index 2; W = Wiener's index; D = Balaban's index.

and 621, respectively, which is exceptionally high (Tables 1 and 2). The exceptionally high discriminating power of proposed indices renders them extremely sensitive toward minor changes in molecular structures. This extreme sensitivity toward branching as well relative position of substituents in cyclic structure and high discriminating power of proposed indices are clearly evident from respective values of all possible structures with five vertices. Further, to encode chemical information of a particular heteroatom involved in a molecular structure one can easily resort to topochemical versions of proposed indices (details in Supporting Information).

Out of a total of ten proposed indices $R_{\xi_1}^{SV}$, $R_{\xi_2}^{SV}$, $R_{\xi_3}^{SV}$, $RP_{\xi_2}^{SV}$, $RP_{\xi_1}^{SV}$, and $RP_{\xi_2}^{SV}$ belong to *third generation*, $RP_{\xi_1}^{SV}$, $RP_{\xi_2}^{SV}$, and $R_{\xi_3}^{SV}$ belong to *fourth generation*, and $R_{\xi_3}^{SV}$ belongs to *fifth generation* as per the criteria specified by Dureja and Madan.²⁹

Evaluation of Eccentric Distance Sum/Product Indices for the Degeneracy. The measure of the ability of an index to differentiate between the relative positions of atom in a molecule is termed as degeneracy. All the proposed MDs did not exhibit any degeneracy for all possible cyclic structures with four and five vertices (Table 2). The low degeneracy indicates the enhanced capability of proposed MDs to differentiate and demonstrate slight variations in the molecular structure, which clearly reveals the remote chance of different structures having the same index value.

Intercorrelation Analysis. Intercorrelation analysis of the proposed MDs with other well-known and widely used MDs revealed that these are not correlated with Wiener's index, molecular connectivity index, eccentric connectivity index and Balaban's index. However, these proposed MDs are weakly correlated with Zagreb indices M_1 and M_2 (Table 3) as per the criteria specified by Trinajstić et al.³⁰

The proposed MDs possess certain distinct advantages over existing ones. First, only few existing MDs are derived from detour matrices. The proposed MDs involve the use of a detour matrix and are based upon relative values of longest distance and shortest distance between various atoms in a hydrogen suppressed molecular structure. This approach provides valuable information with regard to shape factor and is of utmost importance in molecules containing cyclic moieties.

Second, though a very large number of MDs of diverse nature have been reported in literature, only few MDs belong to fourth and fifth generations. Three of the descriptors proposed here belong to fourth generation whereas one belongs to fifth generation. Third, topochemical versions of the proposed MDs exhibit negligible degeneracy. Fourthly, very high index values of

complex chemical structures may result in numerous problems. Previously many researchers resorted to either logarithmic or square root approach so as to keep index values to be within reasonable limits. But, such an approach also results in steep reduction in discriminating power. To overcome this problem, a recently reported approach [i.e., dividing the index values by a constant factor (k_m)] has been used. This method has a distinct edge as it reduces the index values of complex chemical structures to be within reasonable limits without compromising with the discriminating power.³¹ Use of relative values of longest distance and shortest distance between various atoms also helps in keeping index values of complex organic structures to be within limits.

Finally, the proposed MDs are easily interpretable because of their simplicity as compared to other MDs and hence can be easily utilized for development of lead molecules through reverse engineering as well as virtual screening.

Utilization of Eccentric Distance Sum/Product Indices.

In this study the novel as well as existing MDs have been utilized to develop suitable models for the prediction of hQC inhibitory activity of substituted 3-(1*H*-imidazol-1-yl) propyl thiourea derivatives²⁷ (Figure 2 and Table 6) using decision tree (DT), random forest, and MAA.

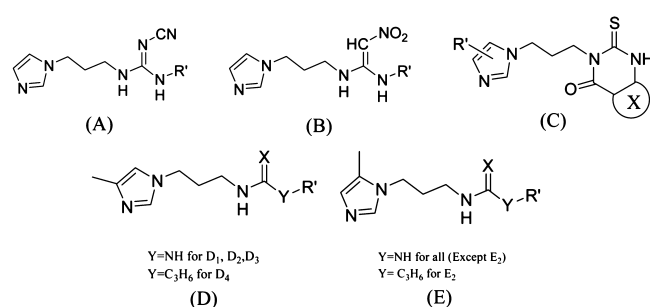


Figure 2. Basic structures and arbitrary atom numbering scheme for the substituted 3-(1*H*-imidazol-1-yl) propyl thiourea derivatives.²⁷

Models Based upon Decision Tree (DT) and Random Forests (RF). DT was built from a set of 46 MDs including the proposed relative eccentric distance product topochemical index 1 (A1) and relative eccentric distance sum topochemical index 1 (A2) enlisted in Table 4. The MDs at root node is most important and the importance of MDs decreases as the length of tree increases.

The classification of substituted substituted 3-(1*H*-imidazol-1-yl) propyl thiourea derivatives as inactive and active with respect to hQC inhibitory activity using a single tree, based on relative eccentric distance sum topochemical index 1 (A2), Ghose crippen molar refractivity (A13) and molecular connectivity topochemical index, (A40) (Figure 3). The decision tree identified the relative eccentric distance sum topochemical index 1 (A2) as the most important index. The decision tree classified the analogues with an accuracy of 95.5%. The specificity and sensitivity of the training set were of the order of 96.4% and 94.1% (Table 5). In 10-fold cross-validation, 84.4% of substituted 3-imidazolyl propyl thiourea analogues were correctly classified with regard to the said biological activity. The specificity and sensitivity of cross validated set were found to be 85.7% and 82.3% respectively and value of MCC for training and cross validated set was found to be 0.91 and 0.7, respectively (Table 5).

The random forests were grown with 46 MDs (Table 4). The RF classified substituted 3-imidazolyl propyl thiourea derivatives

Table 4. List of Molecular Descriptors Employed for the Study

code	descriptor ^a
A1	relative eccentric distance product topochemical index 1, $\xi_1^{RP, EC^{SV}}$
A2	relative eccentric distance sum topochemical index 1, $\xi_1^{R, EC^{SV}}$
A3	superpendentic index, \int_C^p
A4	total information index of atomic composition, IAC
A5	first Zagreb index by valence vertex degrees, ZM1 V
A6	Schultz MTI by valence vertex degrees, SMTIV
A7	Gutman MTI by valence vertex degrees, GMTIV
A8	reciprocal distance Wiener type index, RDSUM
A9	maximal electrotopological negative variation, MAXDN
A10	sum of Kier Hall electrotopological states, Ss
A11	Wiener type index from polarizability weighted distance matrix, Whetp
A12	molecular electrotopological variation, DELS
A13	Ghose crippen molar refractivity, AMR
A14	leading eigenvalue from polarizability weighted distance matrix, Eig1p
A15	E-state topological parameter, TIE
A16	Randic type eigen vector based index from Vander waals weighed distance matrix, VRv1
A17	Randic type eigen vector based index from electronegativity weighed distance matrix, VRe1
A18	Randic type eigen vector based index from polarizability weighed distance matrix, VRp1
A19	eccentric connectivity topochemical index, ξ_c^c
A20	A total size index/unweighted, A_u
A21	1st component size directional WHIM index, L1e
A22	superaugmented eccentric connectivity distance sum topochemical index 2, ξ_{c2}^{SED, EC^c}
A23	Wiener's topochemical index, Wc
A24	augmented eccentric connectivity topochemical index 1, ξ_1^{Ac, EC^c}
A25	V total size index/weighted by atomic masses, V_m
A26	V total size index/weighted by atomic electrotopological states, V_s
A27	average vertex distance degree, VDA
A28	A total size index/weighted by atomic masses, A_m
A29	molecular weight, MW
A30	Broto–Moreau autocorrelation $-\log_8$ /weighted by atomic masses, ATSM
A31	Geary autocorrelation $-\log_8$ /weighted by atomic masses, GATSM
A32	Eigen vector coefficient sum from electronegativity, VEe1
A33	leverage weighted total index/weighted by atomic masses, HATSM
A34	total information content on distance equality, IDET
A35	H autocorrelation of lag 1/weighted by atomic Vander Walls volume, H1v
A36	weighted by atomic Vander Walls volume, HTv
A37	H total index/weighted by atomic masses, HTm
A38	H autocorrelation of lag 0/weighted by atomic masses, H0m
A39	Kier flexibility index, PHI
A40	Molecular connectivity topochemical index, χ^A
A41	3D Wiener index, W3D
A42	3D Balaban index, J3D
A43	Balaban type index from electronegativity weighted distance matrix, J_{hete}
A44	total information content index (neighborhood symmetry of 1 order), TIC1
A45	valence connectivity index chi 0, χ_{0v}
A46	Kier symmetry index, S_{0K}

^aThe majority of Dragon descriptors have been defined in textbooks by Todeschini and Consonni,³² Karelson,³³ and Devillers and Balaban.³⁴

as inactive and active with an accuracy of 93.3% with respect to hQC inhibitory activity. The out-of-bag (OOB) estimate of error was found to be only 6.7%. The specificity and sensitivity was of the order of 92.8% and 92%, respectively, and the value of MCC was found to be 0.86 (Table 5).

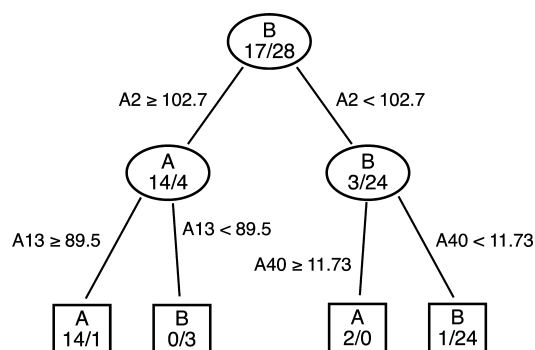


Figure 3. Decision tree for distinguishing active analogue (A) from inactive analogue (B). (A2) Relative eccentric distance sum topochemical index-1 ($R_{\zeta_1}^{ecSV}$). (A13) Ghose crippen molar refractivity (AMR). (A40) Molecular connectivity topochemical index, (χ^A).

High values of MCC simply indicate robustness of the proposed DT and RF based models for hQC inhibitory activity. Decision tree selected A2, relative eccentric distance sum topochemical index 1, as the most important descriptor demonstrating its significance in (Q)SAR/QSPR studies.

Models Based upon Moving Average Analysis. Using a single descriptor at a time, five independent MAA based models using relative eccentric distance product topochemical index 1 (A1), relative eccentric distance sum topochemical index 1 (A2), Ghose crippen molar refractivity (A13), A total size index/unweighted (A20), and first component size directional WHIM index (A21) were developed for predicting the hQC inhibitory activities. The proposed models have been illustrated in Table 7. The overall accuracy of prediction hQC inhibitory activity varied from 86.8% to >95%. Existence of a transitional range in a model is ideal because it clearly indicates a gradual change in the biological activity. The average IC_{50} (Table 7 and Figure 4) for active range in all the models for hQC inhibitory activity varied from 41.8 to 56.1 nM. Extremely low values of average IC_{50} indicate high potency of the active ranges in the proposed models.

The specificity and sensitivity for all the MAA based models was found to be >85% and >75% respectively and the value of MCC was found to be >0.72 (details in Supporting Information). *High values of MCC indicate robustness of the proposed MAA based models for hQC inhibitory activity.* Moreover, intercorrelation analysis (Supporting Information) revealed that none of the pairs of indices A1, A2, A13, A20, and A21 are correlated with each other.

These models can be used to reduce huge compound libraries to a handful of compounds for ultimate synthesis and biological screening in a cost-effective manner. Very few compounds

predicted to be active by all the models should constitute a group of compounds for synthesis and biological screening.

CONCLUSION

The proposed *relative eccentric distance sum indices* and *relative eccentric distance product indices* exhibited exceptionally high discriminating power amalgamated with negligible degeneracy. Proposed MDs were also found to be highly sensitive toward both the presence and the relative positions of heteroatoms. Moreover, these indices were found to be noncorrelating with important topological descriptors. These qualities ensure their utility in drug design, quantitative structure activity/property relationships, combinatorial library design, isomer discrimination, and similarity/dissimilarity studies.

Subsequently the proposed MDs along with other MDs were successfully employed for development of numerous models for development of models for the prediction of hQC inhibitory activity of substituted 3-imidazolyl propyl thiourea derivatives through DT, random forest and MAA. The proposed MD A2 was identified as the most important descriptors by the decision tree. The models exhibited high degree of predictability with regard to hQC inhibitory activity using decision tree, random forest and moving average analysis. The accuracy of prediction of single descriptor based models using DT, RF, and MAA was found to be 96%, 93%, and 95%, respectively. *High values of MCC indicate robustness of the proposed DT, RF, and MAA based models for hQC inhibitory activity.* High accuracy of prediction of proposed models offers vast potential for providing lead structures for the development of potent therapeutic agents as hQC inhibitors for the treatment of Alzheimer's disease.

METHODOLOGY

The values of $R_{\zeta_m}^{ecSV}$ and $RP_{\zeta_m}^{ecSV}$ were calculated for all possible cyclic structures with four and five vertices using an in-house computer program (Figure 1 and Table 1).

Calculation of Topological Indices. *Relative Eccentric Distance Sum Index.* Relative eccentric distance sum index, denoted by $R_{\zeta_m}^{ecSV}$, may be defined as the summation of ratio of the product of maximum path sum and path eccentricity and the product of distance sum and eccentricity of each vertex in a hydrogen suppressed molecular graph having n vertices. It can be expressed as per the following:

$$R_{\zeta_m}^{ecSV} = \frac{1}{k_m} \sum_{i=1}^n \left(\frac{\sigma_i \times \eta_i}{S_i \times E_i} \right)^m \quad (1)$$

where σ_i is the maximum path sum and η_i is the path eccentricity (both obtained through detour matrix), S_i is distance sum, and E_i is the eccentricity (both obtained through distance matrix) of

Table 5. Confusion Matrix for hQC Inhibitory Activity Using Models Based on Decision Tree and Random Forest

model	description	nature of ranges	number of compound predicted		specificity (%)	sensitivity (%)	overall accuracy of prediction (%)	OOB error (%)	MCC
			active	inactive					
decision tree	training set	active	16	1	96.4	94.1	95.5		0.91
		inactive	1	27					
	cross validated set	active	14	3	85.7	82.3	84.4		0.7
		inactive	4	24					
random forest	active	16	1	92.8	92	93.3	6.7	0.86	
	inactive	2	26						

Table 6. Relationship of A1, A2, A13, A15, and A27 with hQC Inhibitory Activity^a


Cd. No.	R'	X	A1	A2	A13	A20	A21	hQC inhibitory activity					Reported ²⁷
								predicted using MAA based models					
								A1	A2	A13	A20	A21	
A ₁	-Me	-	14.48	21.77	46.86	3.36	4.92	-	-	-	-	-	-
A ₂	-CH ₂ C ₃ H ₆	-	29.34	26.51	58.8	5.71	7.03	-	-	-	-	±	-
A ₃	-Ph	-	215.76	35.05	66.64	8.84	4.7	-	-	-	±	-	-
A ₄	4-BrPh	-	47.25	31.44	74.26	8.84	4.7	-	-	-	±	-	-
A ₅	4-CF ₃ Ph	-	94.22	36.21	72.61	11.5	4.77	-	-	-	±	-	-
A ₆	4- <i>i</i> -PrPh	-	100.22	35.47	80.83	14.7	5.21	-	-	-	-	-	-
A ₇	4-OMePh	-	93.96	34.34	71.75	13.5	5.35	-	-	-	±	-	-
A ₈	3, 4-di-OMePh	-	253.73	39.02	76.86	15.1	6.55	-	±	-	-	-	-
A ₉	3,4-O(CH ₂) ₂ OPh	-	44411.14	61.65	72.38	14.4	5.92	±	-	-	-	-	-
B ₁	-Me	-	15.33	22.85	63.43	5.89	4.78	-	-	-	-	-	-
B ₂	-C ₆ H ₁₂	-	235.14	36.17	84.44	10	4.52	-	-	-	±	-	-
B ₃	-Ph	-	235.14	36.17	83.2	9.85	4.55	-	-	-	±	-	-
B ₄	4-ClPh	-	82.71	33.94	88.01	9.85	4.53	-	-	-	±	-	-
B ₅	4-CF ₃ Ph	-	101.78	37.35	89.18	12.7	4.55	-	-	-	±	-	-
B ₆	5-Naphthyl	-	102427.09	65.31	99.65	13	5.89	±	-	+	±	-	-
B ₇	4-OMePh	-	102.75	35.51	88.32	12	5.82	-	-	-	±	-	-
B ₈	3, 4-di-OMePh	-	278.31	40.18	93.43	15.4	6.41	±	±	±	-	-	-
B ₉	3,4-O(CH ₂) ₂ OPh	-	55005.37	63.04	88.94	13	6.28	±	-	-	±	-	-
C ₁	-	-Ph	776486.94	80.95	76.55	6.27	5.29	±	-	-	-	-	-
C ₂	-		1624441.76	87.14	78.28	6.2	5.29	±	-	-	-	-	-
C ₃	-		527208307.14	140	89.41	7.32	7.38	-	-	-	-	±	-
C ₄	-		125954216.76	126.3	84.81	6.29	6.8	-	-	-	-	-	-
C ₅	-		7692804268.59	166	98.61	9.54	8.23	-	-	+	±	+	-
C ₆	4-Me	-Ph	589988.78	77.22	81.7	7.1	6.15	±	-	-	-	-	-
C ₇	5-Me	-Ph	1858950.63	86.67	80.99	6.51	4.98	-	-	-	-	-	+
C ₈	5-Me		1343469119.21	147.1	93.85	7.8	7.19	-	-	±	±	±	+
C ₉	5-Me		19987222910.71	173.6	103.1	9.97	8.14	-	-	+	±	+	+
C ₁₀	5-Me		56356822.56	105.4	106.1	9.02	8.09	-	-	+	±	+	+
C ₁₁	5-Me		113708198.38	111.2	106.6	10.6	6.93	-	-	+	+	±	+
D ₁	3, 4-di-OMePh	S	140.45	36.1	95.97	9.69	12.2	-	-	+	±	+	+

Table 6. continued

Cd. No.	R'	X	A1	A2	A13	A20	A21	hQC inhibitory activity					Reported ²⁷
								predicted using MAA based models					
								A1	A2	A13	A20	A21	
D ₂	3,4-O(CH ₂) ₂ OPh	-NCN	264.84	39.9	102.8	11.6	11.1	±	±	+	+	+	+
D ₃	3,4-O(CH ₂) ₂ OPh	-CHNO ₂	29203.24	59.59	77.49	9.59	11.2	±	-	-	±	+	-
D ₄	3,4-O(CH ₂) ₂ OPh	S	34794.24565	60.89	94.06	9.6	10.7	±	-	±	±	+	-
E ₁	3,4-di-OMePh	S	463.78	40.22	96	10.3	10.9	+	+	+	+	+	+
E ₂	3,4-di-OMePh	S	1045.81	44.75	102.9	11.9	9.96	+	+	+	±	+	+
E ₃	3,4-di-MePh	-NCN	511.70	40.54	81.87	10.7	9.03	+	+	-	+	+	+
E ₄	3,4-di-ClPh	-NCN	345.20	39.24	81.4	7.66	7.21	±	±	-	-	±	-
E ₅	-biphenyl	-NCN	1790.82	48.37	96.92	13.4	10.8	+	+	+	±	+	+
E ₆	3,4-O(CH ₂) ₂ OPh	-NCN	88428.55	65.47	77.52	9.98	9.99	±	-	-	±	+	-
E ₇	2,3,4-tri-OMePh	-NCN	1237.65	46.78	87.12	14	10.4	+	+	-	±	+	+
E ₈	-C ₆ H ₁₂	-CHNO ₂	617.17	40.31	89.59	7.81	7.21	+	+	±	±	±	+
E ₉	4-ClPh	-CHNO ₂	196.65	37.57	93.16	7.51	6.92	-	±	±	-	±	+
E ₁₀	4-CF ₃ Ph	-CHNO ₂	267.06	41.23	94.33	8.66	8.66	±	+	+	±	+	+
E ₁₁	5-naphthyl	-CHNO ₂	64983.38	63.91	104.8	11.2	7.69	±	-	+	±	+	+
E ₁₂	3,4-O(CH ₂) ₂ OPh	-CHNO ₂	109757.63	66.87	94.09	10.2	9.59	±	-	+	+	+	+

^aNote: (+) active compound, (-) inactive compound, and (±) compound in the transitional range.

Table 7. Proposed MAA-Based Models for the Prediction of hQC Inhibitory Activity^a

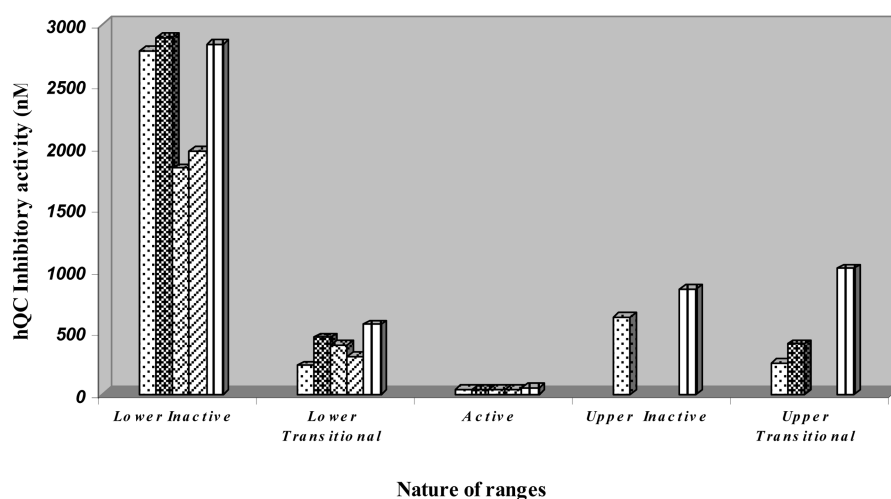
index	nature of range	index value	total compounds in the range	numbers compounds predicted correctly	overall accuracy of prediction (%)	hQC inhibitory activity average IC ₅₀ (nM) ^b	classification (%)
A1	lower inactive	<264.84	16	14	87.88	2784.29	73.33
	lower transitional	264.84 to <463.78	4	NA		234.5	
	active	463.78 to <29203.24	6	6		43.48	
	upper inactive	29203.24 to <1858951	11	9		626.66	
	upper transitional	≥1858951	8	NA		254.63	
A2	inactive	<37.57	14	13	95.23	2893.85	46.66
	lower transitional	37.57 to <40.22	5	NA		463	
	active	40.22 to <59.59	7	7		42.13	
	upper transitional	≥59.59	19	NA		409.79	
A13	inactive	<89.58	27	24	92.11	1834.29	88.89
	transitional	89.58 to <94.09	5	NA		402.4	
	active	≥94.09	13	11		41.81	
A20	lower inactive	<7.798	11	9	90	2839.22	44.4
	lower transitional	7.79 to <10.18	15	NA		568.93	
	active	10.18 to <11.51	5	5		56.06	
	upper transitional	11.51 to <14.4	10	NA		1021.96	
	upper inactive	≥14.4	4	4		855	
A21	inactive	<6.92	22	21	86.84	1971.9	84.44
	transitional	6.92 to <7.68	7	NA		308.28	
	active	≥7.68	16	12		43.33	

^aNA: Not applicable. ^bAverage IC₅₀ values are based upon correctly predicted analogues in the particular range.

vertex i , m is equal to 1, 2, and 3 for $R_{\xi_1}^{RSV}$, $R_{\xi_2}^{RSV}$, and $R_{\xi_3}^{RSV}$. The value of k_m is equal to 1, 10, and 10^2 for k_1 , k_2 , and k_3 , respectively, and n is the number of vertices in a hydrogen suppressed graph G.

The detour distance $\Delta(i, j|G)$ between the vertices i and j of G is the length of the longest path having maximum number of edges separating i and j . The σ_i is the maximum path sum that may be defined as the sum of length of longest path between vertex i and all other vertices in a hydrogen suppressed graph G.

The path eccentricity η_i of vertex i , in graph G is the length of longest path having maximum number of edges separating i and vertex j that is farthest from i , ($\eta_i = \max \Delta(i, j), j|G$). The distance $d(i, j)$ between vertices i and j means the length of a simple path which joins the vertices i and j in graph G and contains minimum number of edges. The S_i is distance sum that may be defined as the sum of length of shortest path between vertex i and all other vertices in graph G. E_i is the eccentricity also referred to as



- A1-Relative adjacent eccentric distance product index-1
- ▨ A2- Relative adjacent eccentric distance sum index-1
- ▩ A13- Ghose crippen molar refractivity
- ▧ A20- A total size index/unweighted
- ▤ A21- 1st component size directional WHIM index

Figure 4. Average IC₅₀ (µM) values of substituted 3-(1H-imidazol-1-yl) propyl thiourea derivatives for hQC inhibitory activity in various ranges of MAA-based models.

separation of a vertex i in a graph G is the distance from i to the vertex farthest from i in G , that is, $E_i = \max d(i, j)$.¹¹

Similarly, topochemical version of the aforementioned index termed as relative eccentric distance sum topochemical index, denoted by $R_{\xi_m}^{ecSV}$, may be defined as the summation of ratio of the product of the maximum chemical path sum and chemical path eccentricity and the product of chemical distance sum and chemical eccentricity of each vertex in a hydrogen suppressed molecular graph having n vertices. It can be expressed as per the following:

$$R_{\xi_m}^{ecSV} = \frac{1}{k_m} \sum_{i=1}^n \left(\frac{\sigma_{ic} \times \eta_{ic}}{S_{ic} \times E_{ic}} \right)^m \quad (2)$$

where σ_{ic} is the maximum chemical path sum, η_{ic} is chemical path eccentricity (both obtained through chemical detour matrix),³¹ and S_{ic} is chemical distance sum, E_{ic} is the chemical eccentricity (both obtained through chemical distance matrix) of vertex i , m is equal to 1, 2, and 3 for $R_{\xi_1}^{ecSV}$, $R_{\xi_2}^{ecSV}$, and $R_{\xi_3}^{ecSV}$. The value of k_m is equal to 1, 10, and 10^2 for k_1 , k_2 , and k_3 , respectively, and n is the number of vertices in a hydrogen suppressed graph G .

Relative Eccentric Distance Product Index. Relative distance product, denoted by $RP_{\xi_m}^{ecSV}$, may be defined as the product of the ratio of product of the maximum path sum and path eccentricity and the product of distance sum and eccentricity of each vertex in a hydrogen suppressed molecular graph having n vertices. It can be expressed as per the following:

$$RP_{\xi_m}^{ecSV} = \frac{1}{k_m} \prod_{i=1}^n \left(\frac{\sigma_i \times \eta_i}{S_i \times E_i} \right)^x \quad (3)$$

where σ_i is the maximum path sum, η_i is the path eccentricity, S_i is distance sum, and E_i is the eccentricity of vertex i and n is the number of vertices in a hydrogen suppressed graph G . The values

of x , k_m , and m are equal to 1/2, 1, and 1 for $RP_{\xi_1}^{ecSV}$, while for $RP_{\xi_2}^{ecSV}$ the values for x , k_m , and m are equal to 1, 100, and 2, respectively.

Similarly topochemical version of the above-mentioned index termed as relative eccentric distance product topochemical index ($RP_{\xi_m}^{ecSV}$) may be defined as the product of ratio of product of the maximum chemical path sum and chemical path eccentricity and the product of chemical distance sum and chemical eccentricity of each vertex in a hydrogen suppressed molecular graph having n vertices. It can be expressed as per the following:

$$RP_{\xi_m}^{ecSV} = \frac{1}{k_m} \prod_{i=1}^n \left(\frac{\sigma_{ic} \times \eta_{ic}}{S_{ic} \times E_{ic}} \right)^x \quad (4)$$

where σ_{ic} is the maximum chemical path sum, η_{ic} is chemical path eccentricity (both obtained through chemical detour matrix), S_{ic} is chemical distance sum, E_{ic} is the chemical eccentricity (both obtained through chemical distance matrix) of vertex i , and n is the number of vertices in a hydrogen suppressed graph G . The values of x , k_m , and m are equal to 1/2, 1, and 1 for $RP_{\xi_1}^{ecSV}$, while for $RP_{\xi_2}^{ecSV}$, the values for x , k_m , and m are equal to 1, 100, and 2, respectively.

Relative eccentric distance sum indices and relative eccentric distance product indices can be easily calculated from the detour matrix (Δ) and distance matrix (D). Calculation of relative eccentric distance sum indices and relative eccentric distance product indices for three isomers of ethyl isopropyl cyclohexane has been exemplified in Figure (Figure 1). Relative eccentric distance sum indices and relative eccentric distance product indices were evaluated for discriminating power, degeneracy, intercorrelation with existing MDs, and sensitivity toward branching, as well as relative position of substituents in cyclic structures. The discriminating power and degeneracy of the relative eccentric distance sum index and relative eccentric distance product index were investigated using all possible structures with four and five vertices (Table1). However, each chemical structure contained one

nitrogen atom as heteroatom in case of topochemical indices (Supporting Information).

The sensitivity of the proposed indices toward branching as well as relative position of substituents in cyclic structures was evaluated using three isomers of ethyl isopropyl cyclohexane (Figure 1).

The intercorrelation of relative eccentric distance/product indices with other well-known indices, such as Wiener's index, Balaban's index (D), Randić's molecular connectivity index, eccentric connectivity index, and Zagreb indices ($M1$ and $M2$), was investigated (Table 3). This intercorrelation was determined with respect to index values of all possible structures containing four and five vertices.

Molecular Descriptors (MDs). Topochemical version of all the five relative eccentric distance sum topochemical indices and relative eccentric distance product topochemical indices (denoted by $R_{S_1}^{ecSV}$, $R_{S_2}^{ecSV}$, $R_{S_3}^{ecSV}$, $RP_{S_1}^{ecSV}$, and $RP_{S_2}^{ecSV}$), along with 450 other 2D and 3D MDs of diverse nature, were used to hold the structural properties of the compounds from all aspects. All computational work was performed on E-Dragon version 1.0. The values of other MDs which are not the part of Dragon were computed separately using an in-house computer program. The descriptors used in the study include topostructural, topochemical, topological charge indices, constitutional, physicochemical, walk and path counts, information based indices, and variety of 3D descriptors. Most of these descriptors have been reviewed in the various textbooks dealing with molecular descriptors.^{32–34} The additional descriptors calculated from an in house program included augmented eccentric connectivity topochemical index,³⁵ and supraaugmented eccentric connectivity distance sum topochemical index 2.³⁶ The processing of the MDs was done by removing invariable (constant) columns and cross correlated descriptors (with $r > 0.97$). The said exclusion method was used to reduce the collinearity and correlation between descriptors.

Finally, 46 MDs (Table 4) were shortlisted from a large pool of MDs on the basis of noncorrelating nature and classification ability and subsequently employed for further analysis by DT and RF. The five MDs employed for models based upon MAA included the newly proposed A1, relative eccentric distance product index 1, A2, relative eccentric distance sum index 1, A13, Ghose crippen molar refractivity,^{32,37} A20, a total size index/unweighted,^{32–34,38} and A21, first component size directional WHIM index.^{32–34,38}

Data Set. All the 45 substituted 3-(1*H*-imidazol-1-yl) propyl thiourea derivatives reported by Buchholz et al. were selected as a data set for the purpose of present study.²⁷ The basic structures for the said derivatives are shown in Figure 2 and various substituents are enlisted in Table 6.

Classification Techniques. Decision Tree. Decision trees provide a simple and powerful method for building models for prediction of structure–activity relationships. This popular machine learning method is also known as recursive partitioning.³⁹ By partitioning the data into disjoint groups, decision trees produce nonlinear models that are interpretable which is a highly valuable property of any statistical machine learning method when applied to (Q)SAR studies.⁴⁰ The decision tree involves rules to split each node in the tree into subsets, stopping rules which determine when a node cannot be split further and rules for pruning or simplifying the initial tree.⁴¹ In present study, DTs were grown to recognize the importance of MDs. In a decision tree, the molecules at each parent node are categorized as per the descriptor value, into two child nodes. The prediction for

molecule reaching a given terminal node is obtained through a majority vote of molecules reaching the same terminal node in the training set. The tree with lowest value of error in cross validation is selected as an optimal tree. In present study, R program (version 2.1.0) along with the RPART library was used to grow decision trees separately for hQC inhibitory activity.

Random Forest. Random forest is a method for classification and regression that was introduced by Breiman and Cutler.⁴² This technique is simply based upon an ensemble of upruned decision trees, through which the prediction of a continuous variable is provided as an average of the predictions of all the trees. From the training data of n molecules, a bootstrap sample is drawn. For each bootstrap sample, a tree is grown with the modification that each node, choose the best split among a randomly selected subset of all the descriptors used in study. The tree is grown to the maximum size (i.e., until no further splits are possible) and not pruned back. These steps are repeated until a sufficiently large number of such trees are grown.⁴³ Random Forest includes a method for assessing the importance of descriptors for the model. When each MD is replaced by random noise, then the resulting deterioration in the model quality serves as a measure of descriptor importance. The model can be validated by assessing the change in mean-square-error for the out-of-bag.³⁹

In the present study, the RFs were grown separately for hQC inhibitory activity with the R program (version 2.1.0) using the random forest library.

Moving Average Analysis. Moving average analysis of correctly predicted compounds is the basis of development of single MD based model.⁴⁴ For the selection and evaluation of range specific features, exclusive activity ranges were discovered from the frequency distribution of therapeutic response level. This was accomplished by plotting the relationship between index values and hQC inhibitory activity and then identifying the active range by analyzing the resultant data by maximization of the moving average with respect to active compounds (<35% inactive, 35–65% transitional and $\geq 65\%$ active). The data set comprised of variable degree of hQC inhibitory activity. Since no reference compound was reported in the data set, therefore, compounds having reported IC_{50} values of ≤ 100 nM were arbitrarily considered to be active [and labeled as "A" ($N = 17$)] while those possessing IC_{50} values > 100 nM [and labeled as "B" ($N = 28$)] were treated to be inactive for the purpose of present study. The hQC inhibitory activity assigned to each compound using proposed models was subsequently compared with the reported biological activity. The accuracy of classification for each range in the proposed models as well as the overall accuracy of prediction of various models was calculated. The average IC_{50} values for each range of proposed models were also calculated.

Data Analysis. The validation of the DT based models and self-consistency test were performed by 10-fold cross validation (CV) method. For classification models the sensitivity and specificity values were calculated.^{45,46} Sensitivity and specificity represent the classification accuracies for the active and inactive compounds involved in the data set.^{45,47} The randomness of model was also determined by calculating Mathew's correlation coefficient (MCC). The MCC values ranging between -1 to $+1$ indicate the robustness of model.⁴⁸ MCC accounts for both sensitivity and specificity and it is generally used as a balanced measure in dealing with data imbalance situation.⁴⁵ The degree of correlation can be appraised by correlation coefficient r . Pairs of MDs with $r \geq 0.97$ are considered to be highly intercorrelated, those with $0.90 \leq r \leq 0.97$ are considered appreciably correlated, those with $0.50 \leq r \leq 0.89$ are weakly correlated and finally pairs

of MDs with low r values (<0.50) are not intercorrelated.³⁰ The intercorrelation between MDs finally utilized for developing MAA based models, that is, A1, A2, A13, A20, and A21 was also investigated. Results are summarized in Tables 5–7 and Figures 3 and 4.

■ ASSOCIATED CONTENT

■ Supporting Information

Methodology and calculation of index values for cyclic structures containing one heteroatom, discriminating power and degeneracy of proposed descriptors, confusion matrix and intercorrelation matrix of MDs used in MAA based models. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: madan_ak@yahoo.com.

Notes

The authors declare no competing financial interest.

■ REFERENCES

- (1) Jordheim, L.; Galmarini, C. M.; Dumontet, C. Drug resistance to cytotoxic nucleoside analogues. *Curr. Drug Targets* **2003**, *4*, 443–460.
- (2) Gálvez, J.; García-Doménech, R. On the contribution of molecular topology to drug design and discovery. *Curr. Comput.-Aided Drug Des.* **2010**, *6*, 252–268.
- (3) Balaban, A. T.; Basak, S. C.; Beteringhe, A.; Mills, D.; Supuran, C. T. QSAR study using topological indices for inhibition of carbonic anhydrase II by sulfanilamides and Schiff bases. *Mol. Diversity* **2004**, *8*, 401–412.
- (4) Selassie, C. D.; Mekapati, S. B.; Verma, R. P. QSAR: Then and now. *Curr. Topics Med. Chem.* **2002**, *2*, 1357–1379.
- (5) Carhart, R. E.; Smith, D. H.; Venkataraghvan, R. Atom pairs as molecular features in structure–activity studies: Definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
- (6) Randić, M.; Pompe, M. The variable connectivity index $1\sigma_f$ versus the traditional molecular descriptors: A comparative study of $1\sigma_f$ against descriptors of CODESSA. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 631–638.
- (7) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure Activity Analysis*; John Wiley: London, 1986.
- (8) Randić, M. Graph valence shells as molecular descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 627–630.
- (9) Gozalbes, R.; Doucet, J. P.; Derouin, F. Application of topological descriptors in QSAR and drug design: history and new trends. *Curr. Drug Targets: Infect. Disord.* **2002**, *2*, 93–102.
- (10) Măndoiu, I. I.; Zelikovsky, A.; Bereg, S. Topological indices in combinatorial chemistry. *Bioinf. Algorithms* **2007**, DOI: 10.1002/9780470253441.ch19.
- (11) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.
- (12) Maran, U.; Sild, S.; Tulp, I.; Takkis, K.; Moosus, M. Molecular descriptors from two-dimensional chemical structure. In *Silico Toxicology: Principles and Applications*; Cronin, M. T. D., Madden, J. C., Eds.; Royal Society of Chemistry: London, 2010; pp 148–187.
- (13) Hollas, B. An analysis of the redundancy of graph invariants used in Chemoinformatics. *Discuss. Appl. Math.* **2006**, *154*, 2484–2498.
- (14) Grassy, G.; Calas, B.; Yasri, A.; Lahana, R.; Woo, J.; Iyer, S.; Kaczorek, M.; Floch, R.; Buelow, R. Computer-assisted rational design of immunosuppressive compounds. *Nat. Biotechnol.* **1998**, *16*, 748–752.
- (15) Ivanciuc, O.; Taraviras, S. L.; Cabrol-Bass, D. Quasi-orthogonal basis sets of molecular graph descriptors as a chemical diversity measure. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 126–134.
- (16) Taraviras, S. L.; Ivanciuc, O.; Cabrol-Bass, D. Identification of groupings of graph theoretical molecular descriptors using a hybrid cluster analysis approach. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1128–1146.
- (17) Eduardo, A.; Castro, A.; Matías Tueros, A.; Toropov, A. A. Maximum topological distances based indices as molecular descriptors for QSPR 2—Application to aromatic hydrocarbons. *Comput. Chem.* **2000**, *24*, 571–576.
- (18) Abbott, A. Dementia: A problem for our age. *Nature* **2011**, *475*, S2–S4.
- (19) Schilling, S.; Wasternack, C.; Demuth, H. U. Glutamyl cyclases from animals and plants: a case of functionally convergent protein evolution. *Biol. Chem.* **2008**, *389*, 983–991.
- (20) De Kimpe, L.; Van Haastert, E. S.; Kaminari, A.; Zwart, R.; Rutjes, H.; Hoozemans, J. J.; Scheper, W. Intracellular accumulation of aggregated pyroglutamate amyloid beta: convergence of aging and Ab pathology at the lysosome. *Age (Dordrecht, Neth.)* **2013**, *35* (3), 673–687, DOI: 10.1007/s11357-012-9403-0.
- (21) Huang, K. F.; Liu, Y. L.; Cheng, W. J.; Ko, T. P.; Wang, A. H. J. Crystal structures of human glutamyl cyclase, an enzyme responsible for protein N-terminal pyroglutamate formation. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (37), 13117–13122.
- (22) He, W.; Barrow, C. J. The A β 3-pyroglutamyl and 11-pyroglutamyl peptides found in senile plaque have greater beta-sheet forming and aggregation propensities in vitro than full-length A β . *Biochemistry* **1999**, *38*, 10871–10877.
- (23) Schlenzig, D.; Manhart, S.; Cinar, Y.; Kleinschmidt, M.; Hause, G.; Willbold, D.; Funke, S. A.; Schilling, S.; Demuth, H. U. Pyroglutamate formation influences solubility and amyloidogenicity of amyloid peptides. *Biochemistry* **2009**, *48*, 7072–7078.
- (24) De Kimpe, L.; Bennis, A.; Zwart, R.; van Haastert, E. S.; Hoozemans, J. J. M.; Scheper, W. Disturbed Ca²⁺ homeostasis increases glutamyl cyclase expression: Connecting two early pathogenic events in Alzheimer's disease in vitro. *PLoS One* **2012**, *7* (9), No. e44674, DOI: 10.1371/journal.pone.0044674.
- (25) Schilling, S.; Zeitschel, U.; Hoffmann, T.; Heiser, U.; Francke, M.; Kehlen, A.; Holzer, M.; Hutter-Paier, B.; Prokesch, M.; Windisch, M.; Jagla, W.; Schlenzig, D.; Lindner, C.; Rudolph, T.; Reuter, G.; Cynis, H.; Montag, D.; Demuth, H. U.; Rossner, S. Glutamyl cyclase inhibition attenuates pyroglutamate Abeta and Alzheimers disease-like pathology. *Nat. Med.* **2008**, *14*, 1106–1111.
- (26) Long, J. M.; Lahiri, D. K. Current drug targets for modulating Alzheimer's amyloid precursor protein: role of specific micro-RNA species. *Curr. Med. Chem.* **2011**, *18*, 3314–3321.
- (27) Buchholz, M.; Hamann, A.; Aust, S.; Brandt, W.; Bohme, L. T.; Hoffmann, L.; Schilling, S.; Demuth, H. U.; Heiser, U. Inhibitors for human glutamyl cyclase by structure based design and bioisosteric replacement. *J. Med. Chem.* **2009**, *52*, 7069–7080.
- (28) Gupta, S.; Singh, M.; Madan, A. K. Eccentric distance sum: A novel graph invariant for predicting biological and physical properties. *J. Math. Anal. Appl.* **2002**, *275*, 386–401.
- (29) Dureja, H.; Madan, A. K. *Distance in Molecular Graphs: Applications*; Gutman, I., Furtula, B., Eds.; University of Kragujevac: Kragujevac, Serbia, 2012; pp 55–80.
- (30) Trinajstić, N.; Nikolic, S.; Basak, S. C.; Lukovits, I. Distance indices and their hypercounterparts: Intercorrelation and use in the structure-property modeling. *SAR QSAR Environ. Res.* **2001**, *12*, 31–54.
- (31) Marwaha, R. K.; Jangra, H.; Das, K. C.; Bharatam, P. V.; Madan, A. K. Fourth generation detour matrix-based topological indices for QSAR/QSPR—Part-1: Development and evaluation. *Int. J. Comput. Biol. Drug Des.* **2012**, *5*, 335–360.
- (32) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley VCH Verlag GmbH: Weinheim, Federal Republic of Germany, 2009; pp 239.
- (33) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley-Interscience: New York, 2000.
- (34) Devillers, J.; Balaban, A. T. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach: Amsterdam, the Netherlands, 1999.
- (35) Bajaj, S.; Sambhi, S. S.; Madan, A. K. Model for prediction of anti-HIV activity of 2-pyridinone derivatives using novel topological descriptors. *QSAR Comb. Sci.* **2006**, *25*, 813–823.

(36) Gupta, M.; Gupta, S.; Dureja, H.; Madan, A. K. Superaugmented eccentric distance sum connectivity indices: novel highly discriminating topological descriptors for QSAR/QSPR. *Chem. Biol. Drug Des.* **2012**, *79* (1), 38–52, DOI: 10.1111/j.1747-0285.2011.01264.x.

(37) Ghose, A. K.; Crippen, G. M. Atomic Physicochemical Parameters for Three-Dimensional-Structure-Directed Quantitative Structure–Activity Relationships. 2. Modeling Dispersive and Hydrophobic Interactions. *J. Comput. Chem.* **1987**, *27*, 21–35.

(38) Todeschini, R.; Gramatica, P.; Provenzani, R.; Mareno, E. Weighted holistic invariant molecular descriptors. Part 2. Theory development and application on modelling physicochemical properties of polyaromatic hydrocarbons. *Chemom. Intell. Lab. Syst.* **1995**, *17* (2), 221–229.

(39) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Chapman & Hall/CRC: Boca Raton, FL, 1984.

(40) Bruce, C. L.; Melville, J. L.; Pickett, S. D.; Hirst, J. D. Contemporary QSAR classifiers compared. *J. Chem. Inf. Model.* **2007**, *47*, 219–227.

(41) Seyagha, M.; Mazouza, E. L. M.; Schmitzerb, A.; Villeminc, D.; Jarida, A.; Cherqaoui, D. Classification structure–activity relationship study of reverse transcriptase inhibitors. *Lett. Drug Des. Discovery* **2011**, *8*, 585–595.

(42) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.

(43) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.

(44) Gupta, S.; Singh, M.; Madan, A. K. Predicting anti-HIV activity: Computational approach using a novel topological descriptor. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 671–678.

(45) Han, L.; Wang, Y.; Bryant, S. H. Developing and validating predictive decision tree models from mining chemical structural fingerprints and high throughput screening data in pubchem. *BMC Bioinf.* **2008**, *9*, 401–408, DOI: 10.1186/1471-2105-9-401.

(46) Roy, K.; Mitra, I. On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. *Comb. Chem. High Throughput Screening* **2011**, *14*, 450–474.

(47) Lamanna, C.; Bellini, M.; Padova, A.; Westerberg, G.; Maccari, L. Straight forward recursive partitioning model for discarding insoluble compounds in the drug discovery process. *J. Med. Chem.* **2008**, *51*, 2891–2897.

(48) Baldi, P.; Bruank, S.; Chauvin, Y.; Andersen, C. A. F.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinf.* **2000**, *16*, 412–424.