

# Classification of Some Active HIV-1 Protease Inhibitors and Their Inactive Analogues Using Some Uncorrelated Three-Dimensional Molecular Descriptors and a Fuzzy c-Means Algorithm

Thy-Hou Lin,\* Ging-Ming Wang, and Yao-Hua Hsu

Department of Life Science, National Tsing Hua, University, Hsinchu, Taiwan 30043, R.O.C.

Received May 7, 2002

A fuzzy c-means algorithm was used to classify some 3D convex hull descriptors computed for 345 active HIV-1 protease inhibitors collected from literature and 437 inactive analogues searched from the MDL/ISIS database. The number of descriptors used to represent each compound was from 4 to 8, and they were uncorrelated using the principal component analysis. These uncorrelated descriptors were then divided into two groups and classified by the fuzzy c-means algorithm. The classification produced a clear-cut switch in membership functions computed for each uncorrelated descriptor at the group boundary. Compounds with nonswitching membership functions computed were treated as outliers, and they were counted for estimating the accuracy of the classification. The averaged accuracy of classification for the active inhibitor set was about 80% which was better than that directly classified by a linear discriminant function on the original 3D convex hull descriptors. The whole classification scheme was also applied to several sets of some conventional descriptors computed for each compound, but the averaged accuracy was around 58%. Further classification using some 3D convex hull descriptors searched from comparing the distribution of these descriptors was performed on a new data set composed of 289 outliers-deducted active inhibitors and 63 outliers identified from the inactive analogues through previous classification. This final classification identified 19 inactive analogues which were similar in structural and topological features to those of some highly active inhibitors classified together with them.

## INTRODUCTION

To derive a relationship between some molecular properties and molecular descriptors computed for some molecules chosen, one usually relies on the use of mathematic techniques such as multiple linear regression<sup>1</sup> or partial least squares regression.<sup>2</sup> One always needs to compromise efficiency, generality, ease of interpretation, and ease of automatic perception in choosing a set of molecular descriptors to construct a meaningful relationship.<sup>3</sup> Before establishing such a relationship, computed molecular descriptors are often subjected to some prior analyses such as principal components analysis,<sup>4</sup> clustering analysis,<sup>5</sup> or analysis to assess differences in information content.<sup>6</sup> These prior analyses are performed either to reduce the dimension of the original descriptor sets or to select an effective subset out of the original one. However, it has been shown that this approach can introduce bias and can lead to artificially high “goodness of fit” parameters in linear regression analyses.<sup>7</sup> Therefore, different molecular descriptors are developed and employed to account for difference in molecular features or properties.<sup>8</sup> For examples, topological descriptors have been successfully used in classifying chemical structures and in predicting physiochemical and biological properties,<sup>9–12</sup> while electronic or chemical properties have been estimated using electronic descriptors.<sup>13–15</sup> However, it has been shown that molecular geometric or physicochem-

ical properties can be represented with a combination of a variety of descriptors.<sup>16</sup>

In this work, we have used some novel 3D descriptors computed as the summed molecular path lengths among any three convex hull vertices for classification of 345 active HIV-1 protease inhibitors<sup>17–24</sup> from their 437 inactive analogues searched from the MDL/ISIS database.<sup>25</sup> The exposed functional groups on the vertices of a convex hull of each structure generated through molecular dynamics and energy minimization were identified, and the molecular path lengths among any three of these vertices were summed and treated as a 3D convex hull descriptor as described previously.<sup>26</sup> Each compound was represented by a set of these 3D convex hull descriptors computed. A principal components analysis was conducted next to eliminate the correlation among these descriptors and to reduce the dimension of the descriptors to two, namely, the first and the second principal components of the original descriptor set. The set of principal components was then classified using a fuzzy c-means algorithm.<sup>27,28</sup> The fuzzy c-means algorithm generalizes the hard c-means algorithm to allow a point to be partially assigned to multiple clusters. Therefore, it produces a soft partition for a given data set. The fuzzy c-means algorithm seeks a good partition by searching for prototypes and their corresponding membership functions such that the corresponding objective function is minimized.<sup>28</sup> A switch in membership functions or a boundary in the distribution of membership functions can be seen when applying the algorithm to two well separated groups. If the classification

\* Corresponding author phone: 886-3-574-2759; fax: 886-3-572-1746; e-mail: thlin@life.nthu.edu.tw.

**Table 1.** Structures of the 345 Active HIV-1 Protease Inhibitors Collected from the Literature<sup>17–24</sup> and Classified in the Study

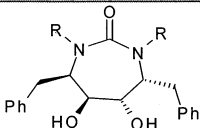
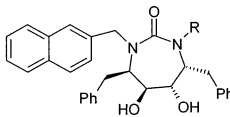
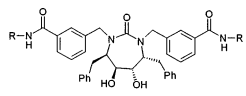
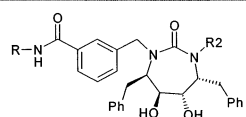
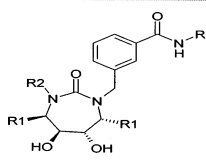
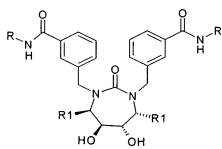
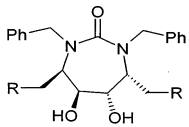
1 <sup>23</sup> (A)	Compound ID number				pKi			
	1 <sub>(147)</sub> <sup>a</sup>	2 <sub>(137)</sub>	3 <sub>(119)</sub>	4 <sub>(90)</sub>	5.24	7.00	8.10	8.85
	6 <sub>(108)</sub>	7 <sub>(142)</sub>	8 <sub>(143)</sub>	9 <sub>(144)</sub>	8.34	6.59	6.10	5.96
	11 <sub>(132)</sub>	12 <sub>(124)</sub>	13 <sub>(115)</sub>	14 <sub>(128)</sub>	7.31	7.92	8.15	7.52
	16 <sub>(130)</sub>	17 <sub>(109)</sub>	18 <sub>(117)</sub>	19 <sub>(97)</sub>	7.44	8.28	8.14	8.74
	21 <sub>(125)</sub>	22 <sub>(99)</sub>	23 <sub>(89)</sub>	24 <sub>(107)</sub>	7.66	8.68	8.89	8.37
	27 <sub>(104)</sub>	28 <sub>(139)</sub>	29 <sub>(122)</sub>	30 <sub>(136)</sub>	8.52	6.84	8.01	7.05
	26 <sub>(146)</sub>	51 <sub>(52)</sub>	32 <sub>(65)</sub>	52 <sub>(50)</sub>	5.40	9.85	9.51	9.92
	33 <sub>(129)</sub>	34 <sub>(105)</sub>	35 <sub>(91)</sub>	36 <sub>(141)</sub>	7.47	8.52	8.85	6.62
	39 <sub>(92)</sub>	40 <sub>(127)</sub>	41 <sub>(116)</sub>	42 <sub>(112)</sub>	8.85	7.57	8.15	8.24
	45 <sub>(145)</sub>	46 <sub>(96)</sub>	47 <sub>(140)</sub>	48 <sub>(103)</sub>	5.73	8.80	6.80	8.55
	38 <sub>(110)</sub>	53 <sub>(51)</sub>	44 <sub>(133)</sub>	54 <sub>(62)</sub>	8.28	9.92	7.29	9.55
	49 <sub>(72)</sub>	43 <sub>(126)</sub>	37 <sub>(83)</sub>	50 <sub>(67)</sub>	9.38	7.66	9.05	9.47
	31 <sub>(135)</sub>	25 <sub>(131)</sub>	20 <sub>(134)</sub>	15 <sub>(138)</sub>	7.07	7.43	7.22	6.96
	10 <sub>(148)</sub>	5 <sub>(95)</sub>			5.11	8.80		
2 <sup>23</sup> (A)	Compound ID number				pKi			
	55 <sub>(87)</sub>	56 <sub>(77)</sub>	57 <sub>(93)</sub>	65 <sub>(85)</sub>	8.96	9.22	8.85	9.00
	58 <sub>(94)</sub>	59 <sub>(63)</sub>	60 <sub>(100)</sub>	64 <sub>(84)</sub>	8.82	9.55	8.64	9.03
	61 <sub>(111)</sub>	62 <sub>(114)</sub>	66 <sub>(66)</sub>	63 <sub>(106)</sub>	8.28	8.16	9.48	8.44
3 <sup>23</sup> (A)	Compound ID number				pKi			
	67 <sub>(25)</sub>	68 <sub>(9)</sub> <sup>*b</sup>	69 <sub>(10)</sub> <sup>*</sup>	83 <sub>(29)</sub>	10.41	10.74	10.70	10.37
	70 <sub>(31)</sub>	71 <sub>(39)</sub>	72 <sub>(57)</sub>	84 <sub>(60)</sub>	10.35	10.18	9.68	9.59
	73 <sub>(76)</sub>	74 <sub>(68)</sub>	75 <sub>(73)</sub>	90 <sub>(59)</sub>	9.24	9.44	9.37	9.61
	76 <sub>(101)</sub>	77 <sub>(80)</sub>	78 <sub>(58)</sub>	89 <sub>(2)</sub> <sup>*</sup>	8.62	9.13	9.68	10.92
	79 <sub>(37)</sub>	80 <sub>(74)</sub>	81 <sub>(70)</sub>	82 <sub>(64)</sub>	10.20	9.37	9.39	9.54
	85 <sub>(17)</sub>	86 <sub>(1)</sub> <sup>*</sup>	87 <sub>(11)</sub>	88 <sub>(7)</sub> <sup>*</sup>	10.57	10.96	10.70	10.80
	91 <sub>(22)</sub>	92 <sub>(49)</sub>	93 <sub>(44)</sub>	95 <sub>(54)</sub>	10.46	9.94	10.07	9.82
	96 <sub>(38)</sub>	97 <sub>(55)</sub>	98 <sub>(46)</sub>	99 <sub>(18)</sub> <sup>*</sup>	10.19	9.74	9.96	10.57
	100 <sub>(15)</sub> <sup>*</sup>	101 <sub>(4)</sub> <sup>*</sup>	102 <sub>(5)</sub> <sup>*</sup>	103 <sub>(13)</sub> <sup>*</sup>	10.60	10.85	10.85	10.62
4 <sup>23</sup> (A)	Compound ID number				pKi			
	104 <sub>(102)</sub>	105 <sub>(98)</sub>	106 <sub>(78)</sub>	107 <sub>(82)</sub>	8.60	8.72	9.15	9.07
	108 <sub>(24)</sub>	109 <sub>(40)</sub> <sup>*</sup>	110 <sub>(36)</sub> <sup>*</sup>	111 <sub>(32)</sub> <sup>*</sup>	10.42	10.16	10.28	10.33
	112 <sub>(42)</sub>	113 <sub>(45)</sub>	114 <sub>(71)</sub>	115 <sub>(8)</sub> <sup>*</sup>	10.12	10.02	9.39	10.80
	116 <sub>(12)</sub> <sup>*</sup>	117 <sub>(3)</sub> <sup>*</sup>			10.64	10.92		
5 <sup>24</sup> (A)	Compound ID number				pKi			
	XV655 <sub>(41)</sub>	SD143 <sub>(21)</sub> <sup>*</sup>	SD152 <sub>(16)</sub> <sup>*</sup>	SD145 <sub>(23)</sub> <sup>*</sup>	10.13	10.48	10.60	10.43
6 <sup>24</sup> (A)	Compound ID number				pKi			
	XV638 <sub>(19)</sub> <sup>*</sup>	XV652 <sub>(6)</sub> <sup>*</sup>	SD146 <sub>(14)</sub> <sup>*</sup>		10.57	10.85	10.62	
7 <sup>19</sup> (A)	Compound ID number				pKi			
	6 <sub>(75)</sub>	19 <sub>(123)</sub>	20 <sub>(120)</sub>	21 <sub>(113)</sub>	9.36	8.01	8.10	8.19
	22 <sub>(86)</sub>	23 <sub>(88)</sub>	24 <sub>(79)</sub>		9.00	8.92	9.15	

Table 1 (Continued)

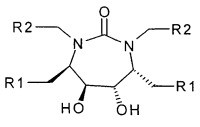
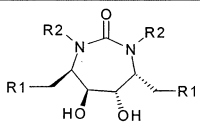
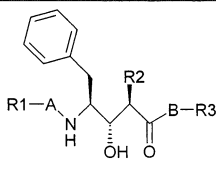
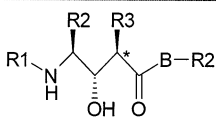
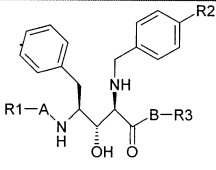
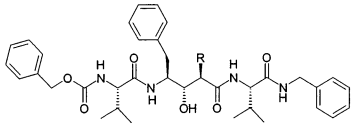
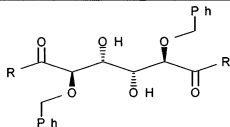
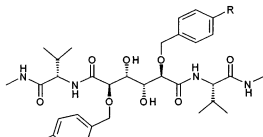
8 <sup>19</sup> (A)	Compound ID number				pKi			
	13 <sub>(26)</sub> *	15 <sub>(27)</sub> *	17 <sub>(47)</sub>	35 <sub>(69)</sub>	10.40	10.40	9.96	9.40
	18 <sub>(53)</sub>	25 <sub>(61)</sub>	26 <sub>(28)</sub> *	36 <sub>(43)</sub>	9.85	9.57	10.40	10.10
	27 <sub>(34)</sub> *	28 <sub>(20)</sub> *	30 <sub>(33)</sub> *	33 <sub>(30)</sub> *	10.30	10.52	10.32	10.37
	31 <sub>(48)</sub>	32 <sub>(56)</sub>			9.96	9.70		
9 <sup>17</sup> (A)	Compound ID number				pKi			
	2 <sub>(118)</sub>	3 <sub>(81)</sub>	6 <sub>(121)</sub>	7 <sub>(35)</sub> *	8.11	9.10	8.04	10.29
10 <sup>20</sup> (B)	Compound ID number				pKi			
	282215 <sub>(160)</sub> *	282310 <sub>(266)</sub>	282311 <sub>(195)</sub>	282312 <sub>(305)</sub>	8.21	7.36	7.96	5.92
	282314 <sub>(278)</sub>	282327 <sub>(233)</sub>	282329 <sub>(164)</sub> *	282349 <sub>(263)</sub>	7.14	7.64	8.16	7.38
	282350 <sub>(200)</sub>	282351 <sub>(206)</sub>	282365 <sub>(171)</sub>	282366 <sub>(224)</sub>	7.92	7.87	8.08	7.77
	283010 <sub>(193)</sub>	283011 <sub>(190)</sub>	283012 <sub>(152)</sub> *	283013 <sub>(184)</sub>	7.98	7.99	8.38	8.00
	283043 <sub>(216)</sub>	283045 <sub>(217)</sub>	283046 <sub>(211)</sub>	283047 <sub>(310)</sub>	7.82	7.81	7.84	5.80
	283051 <sub>(296)</sub>	283052 <sub>(261)</sub>	283053 <sub>(178)</sub>	283054 <sub>(312)</sub>	6.30	7.44	8.02	5.74
	282382 <sub>(256)</sub>	282388 <sub>(292)</sub>	282389 <sub>(299)</sub>	282985 <sub>(208)</sub>	7.47	6.67	6.13	7.85
	283004 <sub>(161)</sub> *	283005 <sub>(168)</sub>	282390 <sub>(267)</sub>	282391 <sub>(275)</sub>	8.21	8.12	7.35	7.19
	282392 <sub>(257)</sub>	282978 <sub>(246)</sub>	282979 <sub>(198)</sub>	282981 <sub>(172)</sub>	7.47	7.57	7.96	8.08
	282395 <sub>(192)</sub>	282396 <sub>(288)</sub>	282412 <sub>(196)</sub>	282946 <sub>(197)</sub>	7.98	6.80	7.96	7.96
	282967 <sub>(314)</sub>	282969 <sub>(289)</sub>	282423 <sub>(258)</sub>	282429 <sub>(230)</sub>	5.64	6.74	7.47	7.70
	282450 <sub>(287)</sub>	282939 <sub>(204)</sub>	282943 <sub>(282)</sub>	282944 <sub>(253)</sub>	6.82	7.89	6.97	7.48
	282453 <sub>(189)</sub>	282455 <sub>(248)</sub>	282456 <sub>(173)</sub>	282870 <sub>(150)</sub> *	7.99	7.54	8.07	8.47
	282915 <sub>(239)</sub>	282916 <sub>(228)</sub>	282457 <sub>(177)</sub>	282479 <sub>(274)</sub>	7.62	7.74	8.04	7.21
	282480 <sub>(241)</sub>	282833 <sub>(295)</sub>	282834 <sub>(297)</sub>	282835 <sub>(207)</sub>	7.59	6.34	6.19	7.85
	282518 <sub>(165)</sub> *	282529 <sub>(244)</sub>	282539 <sub>(265)</sub>	282826 <sub>(235)</sub>	8.15	7.57	7.37	7.64
	282828 <sub>(231)</sub>	282832 <sub>(318)</sub>	282540 <sub>(183)</sub>	282542 <sub>(167)</sub>	7.68	5.57	8.00	8.13
	283055 <sub>(191)</sub>	282823 <sub>(304)</sub>	282824 <sub>(249)</sub>	282825 <sub>(252)</sub>	7.99	5.96	7.54	7.52
	282547 <sub>(238)</sub>	282558 <sub>(225)</sub>	282632 <sub>(251)</sub>	282658 <sub>(317)</sub>	7.62	7.77	7.52	5.60
	282659 <sub>(301)</sub>	282664 <sub>(162)</sub> *	282666 <sub>(182)</sub>	282700 <sub>(271)</sub>	6.10	8.19	8.01	7.23
	282701 <sub>(174)</sub>	282807 <sub>(308)</sub>	282808 <sub>(311)</sub>	282822 <sub>(194)</sub>	8.05	5.85	5.77	7.97
	282779 <sub>(234)</sub>	282796 <sub>(215)</sub>	282797 <sub>(219)</sub>	282752 <sub>(280)</sub>	7.64	7.82	7.80	7.04
	282753 <sub>(203)</sub>	282756 <sub>(214)</sub>	282713 <sub>(245)</sub>	282714 <sub>(213)</sub>	7.89	7.82	7.57	7.82
	282747 <sub>(218)</sub>	282748 <sub>(302)</sub>	282749 <sub>(285)</sub>	282751 <sub>(294)</sub>	7.80	6.07	6.86	6.38
11 <sup>20</sup> (B)	Compound ID number				pKi			
	282730 <sub>(226)</sub>	282364 <sub>(286)</sub>	283194 <sub>(175)</sub> *	283253 <sub>(202)</sub>	7.77	6.85	8.05	7.90
	282665 <sub>(179)</sub>	283186 <sub>(321)</sub>	283489 <sub>(247)</sub>		8.02	5.43	7.55	
12 <sup>20</sup> (B)	Compound ID number				pKi			
	283134 <sub>(154)</sub> *	283143 <sub>(157)</sub> *	283573 <sub>(303)</sub>	283580 <sub>(291)</sub>	8.34	8.22	6.03	6.68
	283206 <sub>(220)</sub>	283209 <sub>(180)</sub>	283567 <sub>(313)</sub>	283568 <sub>(201)</sub>	7.80	8.02	5.72	7.92
	283239 <sub>(242)</sub>	283245 <sub>(283)</sub>	283569 <sub>(155)</sub> *	283560 <sub>(319)</sub>	7.59	6.89	8.25	5.57
	283249 <sub>(309)</sub>	283254 <sub>(221)</sub>	283549 <sub>(149)</sub> *	283550 <sub>(156)</sub>	5.82	7.80	8.48	8.24
	283260 <sub>(185)</sub>	283261 <sub>(209)</sub>	283532 <sub>(188)</sub>	283537 <sub>(176)</sub>	8.00	7.85	8.00	8.05
	283262 <sub>(232)</sub>	283263 <sub>(205)</sub>	283522 <sub>(254)</sub>	283521 <sub>(277)</sub>	7.68	7.89	7.48	7.15
	283265 <sub>(181)</sub>	283266 <sub>(276)</sub>	283519 <sub>(229)</sub>	283520 <sub>(300)</sub>	8.02	7.16	7.72	6.11
	283267 <sub>(284)</sub>	283321 <sub>(307)</sub>	283336 <sub>(279)</sub>	283497 <sub>(316)</sub>	6.89	5.89	7.10	5.60
	283498 <sub>(270)</sub>	283516 <sub>(151)</sub> *	283337 <sub>(306)</sub>	283341 <sub>(199)</sub>	7.27	8.39	5.92	7.96
	283342 <sub>(240)</sub>	283478 <sub>(315)</sub>	283481 <sub>(322)</sub>	283490 <sub>(158)</sub> *	7.62	5.60	5.34	8.22
	283343 <sub>(268)</sub>	283353 <sub>(222)</sub>	283356 <sub>(281)</sub>	283451 <sub>(210)</sub>	7.33	7.80	7.04	7.85
	283471 <sub>(153)</sub> *	283472 <sub>(170)</sub> *	283364 <sub>(293)</sub>	283365 <sub>(269)</sub>	8.37	8.11	6.64	7.29
	283366 <sub>(169)</sub> *	283436 <sub>(250)</sub>	283440 <sub>(262)</sub>	283441 <sub>(259)</sub>	8.11	7.54	7.43	7.47
	283374 <sub>(212)</sub>	283406 <sub>(298)</sub>	283407 <sub>(264)</sub>	283410 <sub>(223)</sub>	7.82	6.17	7.38	7.80
	283412 <sub>(186)</sub>	283434 <sub>(320)</sub>			8.00	5.52		

Table 1 (Continued)

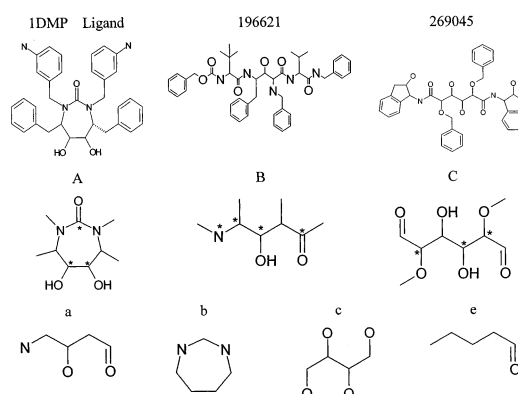
13 <sup>21</sup> (B) <sup>a</sup>	Compound ID number				pKi			
	6 <sub>(159)</sub> *	7 <sub>(187)</sub>	8 <sub>(163)</sub> *	9 <sub>(260)</sub>	8.21	8.00	8.16	7.47
	10 <sub>(236)</sub>	11 <sub>(255)</sub>	12 <sub>(243)</sub>	13 <sub>(272)</sub>	7.64	7.48	7.59	7.23
	14 <sub>(273)</sub>	15 <sub>(290)</sub>	16 <sub>(166)</sub>	17 <sub>(227)</sub>	7.21	6.70	8.13	7.77
	18 <sub>(237)</sub>				7.64			
14 <sup>18</sup> (C)	Compound ID number				pKi			
	28 <sub>(331)</sub> *	37 <sub>(342)</sub> *	38 <sub>(323)</sub> *	47 <sub>(345)</sub> *	9.40	6.86	5.30	5.70
	39 <sub>(343)</sub> *	40 <sub>(344)</sub> *	41 <sub>(339)</sub> *	48 <sub>(335)</sub> *	6.30	6.18	8.64	9.05
	42 <sub>(341)</sub> *	43 <sub>(326)</sub> *	46 <sub>(324)</sub> *		7.09	9.70	5.15	
15 <sup>22</sup> (C)	Compound ID number				pKi			
	2 <sub>(330)</sub> *	3 <sub>(327)</sub> *	4 <sub>(334)</sub> *	5 <sub>(338)</sub> *	9.40	9.52	9.15	8.85
	6 <sub>(336)</sub> *	7 <sub>(337)</sub> *	8 <sub>(340)</sub> *	9 <sub>(328)</sub> *	8.92	8.92	8.42	9.52
	10 <sub>(332)</sub> *	11 <sub>(333)</sub> *	12 <sub>(325)</sub> *	13 <sub>(329)</sub> *	9.22	9.22	10.05	9.52

<sup>a</sup> The inhibitors in each series were arranged in order of pKi and designated with the number parenthesized and subscripted. The arranged inhibitors were classified using the linear discriminant function,<sup>29</sup> and the results were shown in Table 5a,b. <sup>b</sup> These marked inhibitors were selected for structural alignment and CoMFA analysis using the SYBYL 6.7 program,<sup>32</sup> and the results were presented in Figure 2a–c.

for the two groups is not perfect, there will be no switching in membership functions for some members on each side of the boundary. These members are recognized as outliers in the classification process. On the other hand, all the membership functions will be collapsed on each other or there will be no boundary between if the two are too close to be separated. The two principal components computed for the active or inactive set of HIV-1 protease inhibitors were treated as the two separated groups and classified by the fuzzy c-means algorithm.<sup>28</sup> We found that a boundary was present between the two principal components classified, and there were also outliers on each side of the boundary. However, no apparent boundary was observed for the classified principal components computed from some conventional molecular descriptors for the same sets of compounds using the same algorithm. The primary classification identified some outliers from the inactive set which were grouped together with the outliers deducted active one for further classification. The difference in distribution of frequency of each compound being identified as an outlier in these two groups was examined to find a set of new descriptors for further classification. A comparison of classification results using the aforementioned scheme with those of the same sets of multiple descriptors directly classified by a linear discriminant function<sup>29</sup> was also presented.

## METHODS

The compound original identification number (ID) and the corresponding activity expressed in pKi of the 345 active HIV-1 protease inhibitors were collected from literature<sup>17–24</sup> and listed in Table 1. These compounds were divided into three series, namely, A (148 compounds), B (174 compounds), and C (23 compounds) according to the basic structure units shown in Figure 1. The ligand ((4*R*,5*S*,6*S*,7*R*)-1,3-bis(3-aminobenzyl)-4,7-dibenzyl-5,6-dihydroxyhexahydro-1,3-diazepin-2-one) of X-ray structure 1dmp<sup>30</sup> (Figure



**Figure 1.** Structures on the first line are the three active HIV-1 protease inhibitors used to construct the structures for all the 345 active inhibitors listed in Table 1. The three basic structure units used to divide all the active inhibitors into A, B, and C series are shown on the second line, while the four query structures a, b, c, and e used to search the MDL/ISIS<sup>25</sup> database to obtain the 437 inactive analogues are shown on the third line. The correspondence atoms used in aligning the structures and performing the CoMFA analysis are also marked with a “\*” symbol on the three basic structure units on the second line.

1) was used as a template to construct series A structures. The templates for construction of series B and C structures were Extreg 196621 (N-[2-(benzylamino)-(4*S*)-(benzyl-oxycarbonyl-L-tert-leucylamino)-3-hydroxy-5-phenylpentanoyl]-L-valine benzylamide and Extreg 269045 ((2*R*,3*R*,4*R*,5*R*)-2,5-di(benzyloxy)-3,4-dihydroxy-N,N'-bis[(2*R*)-hydroxy-indan-(1*S*)-yl]hexanediamide) (Figure 1) and both were searched from the MDL/ISIS database.<sup>25</sup> For series A structures, the get\_near\_res module of DOCK 4.0 program<sup>31</sup> was used to extract protein atoms that were within a distance of 10 Å of any ligand atom for energy minimization accompanying with the ligand using the SYBYL6.7 program.<sup>32</sup> A distance-dependent dielectric constant ( $\epsilon = 32r$ ) where  $r$  was the distance was used, and the nonbonded cutoff was set at 8 Å. The ligand depleted X-ray structure 4hvp<sup>33</sup> was used as the receptor for docking both series B and C

**Table 2.** 16 Conventional Descriptors Described by Xu and Stevenson<sup>34</sup>

designation	description
n-H	number of non-H atoms
N&O	number of N and O atoms
N&1-M	number of N atoms with at least one H atom
MLipo	molecular lipophilicity, number of carbon atoms as the terminal group
OH	number of hydroxyl group
Hdon	number of H-bond donor
Haccp	number of H-bond acceptor
Caps	number of caps
2-deg	number of 2-degree chain atom, acyclic atom connected with 2 non-H atoms
3-deg	number of 3-degree chain atom, acyclic atom connected with 3 non-H atoms
2-dgc	number of 2-degree cyclic atom, cyclic atom connected with 2 non-H atoms
3-dgc	number of 2-degree cyclic atom, cyclic atom connected with 2 non-H atoms
n-Hpol	number of non-H polar bonds, polarity
n-Hrot	number of non-H rotating bonds, flexibility
amide	number of amide linkage
MCD	molecular cyclized degree: number of ring atoms/total number of atoms

structures into it. Both series of structures were further energy minimized using the same protocol as that described for series A. The validity of structures thus constructed for each series was checked by aligning some selected structures against the template in each series using the SYBYL FIT module.<sup>32</sup> The goodness of each alignment was examined using the SYBYL CoMFA (comparative molecular field analysis) module<sup>32</sup> and the default settings within the program. The structures of 437 inactive analogues of HIV-1 protease inhibitors were searched from the MDL/ISIS database<sup>25</sup> using the query structures a, b, c, and e depicted in Figure 1. Computation of a convex hull for each of the 782 structures generated was then performed. Each 3D convex hull descriptor was computed as the sum of molecular path lengths among any three vertex atoms of the convex hull identified, namely, a sum of path lengths between 1–2, 1–3, and 2–3 as described previously. The 16 conventional descriptors described by Xu and Stevenson<sup>34</sup> and listed in Table 2 were also computed for each structure. The computation of these descriptors was performed using a program developed in-house.

To classify the active and inactive HIV-1 protease inhibitors, each compound was represented by 4 to 8 3D convex hull or conventional descriptors. The frequency of appearance of each 3D convex hull or 16 conventional descriptors computed for each structure was counted and sorted. Then a desired number of descriptors were randomly picked from the sorted descriptors to form a set of multiple descriptors. The first and the second principal components of each set of multiple descriptors were computed from the eigenvalues of the covariance matrix of each descriptor set. These two principal components were separated into two groups as for active or inactive HIV-1 protease inhibitor set. The center of mass for each group was computed and treated as a prototype for initiation of classification by the fuzzy c-means algorithm.<sup>28</sup>

The objective function  $J_m$  of the fuzzy c-means algorithm<sup>28</sup> was extended from the hard c-means one by incorporating a

fuzzy membership function  $\mu_{C_i}$  in it as follows

$$J_m(P, V) = \sum_{i=1}^k \sum_{\chi_k \in X} (\mu_{C_i}(\chi_k))^m \|\chi_k - v_i\|^2 \quad (1)$$

where  $V$  was a vector of cluster centers to be identified,  $P$  was a fuzzy partition of the data set  $X$  formed by  $C_1, C_2, \dots, C_k$ , and  $\|\cdot\|$  was the Euclidean distance between the two vectors  $\chi_k$  and  $v_i$ . The parameter  $m$  was a weight that determines the degree to which partial members of a cluster affect the clustering result. A good partition could be found by searching for prototypes  $v_i$  that minimize the objective function  $J_m$ . A constrained fuzzy partition  $\{C_1, C_2, \dots, C_k\}$  could be a local minimum of the objective function  $J_m$  only if the following conditions were satisfied:

$$\mu_{C_i}(x) = \frac{1}{\sum_{j=1}^k \left( \frac{\|x - v_i\|^2}{\|x - v_j\|^2} \right)^{1/m-1}} \quad 1 \leq i \leq k, x \in X \quad (2)$$

$$v_i = \frac{\sum_{x \in X} (\mu_{C_i}(x))^m \times x}{\sum_{x \in X} (\mu_{C_i}(x))^m} \quad 1 \leq i \leq k \quad (3)$$

Therefore, the fuzzy c-means algorithm<sup>28</sup> updated the prototypes and the membership function iteratively using eqs 2 and 3 until a convergence criterion was reached, namely, until  $\sum_{i=1}^k \|\nu_i^{\text{previous}} - v_i\| \leq \epsilon$  where  $\epsilon$  was a threshold for the convergence criteria. However, Xie and Beni<sup>35</sup> introduced a validity measure that considers both the compactness of clusters as well as the separation between clusters. Intuitively, the more compact the clusters were and the further the separation between clusters, the more desirable a partition. To achieve this, the Xie-Beni validity index<sup>35</sup> (denoted as  $\nu_{XB}$ ) was calculated as follows

$$\nu_{XB} = \left( \frac{\sum \sigma_i}{n} \right) \times \frac{1}{d_{\min}^2} \quad (4)$$

where  $\sigma_i$  was the variation of cluster  $C_i$  defined as

$$\sigma_i = \sum_j \mu_{C_i}(x_j) \|x_j - v_i\|^2 \quad (5)$$

where  $n$  was the cardinality of the data set and  $d_{\min}$  was the shortest distance between cluster centers defined as

$$d_{\min} = \min_{\substack{i,j \\ i \neq j}} \|v_i - v_j\| \quad (6)$$

The first term in eq 4 was a measure of noncompactness, and the second term was a measure of nonseparation. Hence, the product of the two reflected the degree that clusters in a soft partition were not compact and not well separated. Obviously, the lower the cluster index  $\nu_{XB}$ , the better the soft partition was.<sup>35</sup> Therefore, the fuzzy c-means algorithm<sup>28</sup> used includes the following two steps: (i) initializes prototype  $V = \{\nu_{XB}^1, \nu_{XB}^2\}$ , where  $\nu_{XB}^1$  represents the centers of



mass for the two principal components of the active compound set and  $v_{XB}^2$  represents the centers of mass for the two principal components of the inactive compound set; (ii) computes membership functions using eq 2 and updates the prototype,  $v_{XB}^i$  in  $V$  using eq 3 until  $\sum_{i=1}^C ||v_{XB}^{i, previous} - v_{XB}^i|| \leq \epsilon$ . The parameter  $m$  in eq 2 or 3 was set as 2 throughout the entire computation process. The best classification result was searched from thousands of data matrices randomly generated for a desired number of descriptors chosen. A linear discriminant program described by James<sup>29</sup> was also used to directly classify some multiple descriptor sets computed for the two inhibitor sets. By assuming the multivariate normality within each descriptor set and an equal covariance for each set, the linear discriminant program was broken down into three parts: (i) the calculation of estimates of the set means and common covariance matrix, (ii) the calculation of the linear discriminant function coefficients, and (iii) the classification of the descriptor sets using the linear discriminant functions.<sup>29</sup> The linear discriminant function<sup>29</sup> could be written as

$$\int_i(x) = \sum_{k=1}^n C_{ki} + C_{0i} \quad (7)$$

where the  $C_{ki}$  were the elements of the vector  $C_i$  and  $C_{0i}$  was a constant and both were defined as follows

$$C_i = \sum^{-1} \mu_i \quad (8)$$

and

$$C_{0i} = -\frac{1}{2} \mu_i' \sum^{-1} \mu_i \quad (9)$$

where  $\mu_i$  was the set mean vector and  $\sum$  was the set covariance matrix.<sup>29</sup>

## RESULTS AND DISCUSSION

The 3D convex hull descriptors have been successfully used in classification of 73 HIV-1 protease inhibitors to two groups which agreed with the visual classification result using the activity data as the classification criterion.<sup>26</sup> A large number of new inhibitors have been designed, synthesized, and assayed due to the growing problem of drug resistance.<sup>17–24</sup> The structure features of the 345 active HIV-1 protease inhibitors<sup>17–24</sup> listed in Table 1 are very diversified since it includes 11 L-mannaric acid derivatives, 12  $C_2$ -symmetric  $P_1/P_1'$  derivatives, 13 2-heterosubstituted 4-amino-3-hydroxy-5- phenylpentanoic acid derivatives, 92 2-heterosubstituted statine derivatives, and 217 cyclic urea derivatives.<sup>17–24</sup> However, the 345 active inhibitors can be divided into A (148 compounds), B (174 compounds), and C (23 compounds) series according to the three basic structure units shown in Figure 1. The range of activity in pKi in each series varies from 5.11 to 10.96 for A, 5.34 to 8.48 for B, and 5.15 to 10.05 for C (Table 1). Therefore, most of the active compounds were in series A or C while the medium active ones were in series B (Table 1). To test the validity of the structures generated, we have selected 37, 19, or 23 structures out of series A, B, or C for structure alignment using the SYBYL FIT module.<sup>32</sup> The pKi ranges of these selected compounds were 10.16–10.96, 8.05–8.48, and 5.15–10.05,

**Table 3.** Eight 3D Convex Hull and 16 Conventional Descriptors Computed for the 20 Most Active HIV-1 Protease Inhibitors

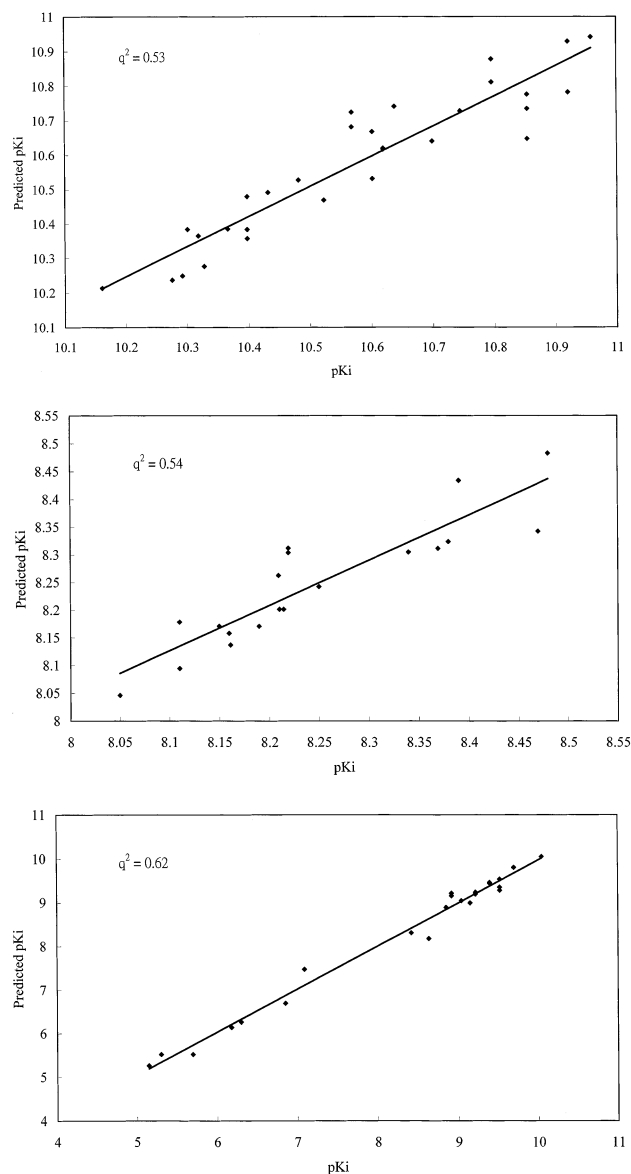
structure ID		3 D convex hull descriptors							
3-086	16	60	36	51	12	20	52	54	
3-089	4	16	57	52	32	47	31	39	
4-117	18	33	24	25	32	14	38	27	
3-101	50	38	6	39	36	47	28	52	
3-102	41	44	20	48	25	27	36	16	
6-652	32	52	39	25	42	36	22	41	
3-088	31	14	24	48	30	8	28	60	
4-115	37	34	16	14	31	22	35	41	
3-068	43	14	42	37	35	12	32	27	
3-069	46	39	31	40	28	4	37	38	
3-087	27	4	22	50	34	45	49	38	
4-116	37	25	39	40	16	15	32	27	
3-103	20	38	31	16	45	8	44	32	
6-146	26	46	51	58	43	22	30	24	
3-100	28	4	13	24	42	37	16	47	
5-152	16	32	29	39	15	28	26	27	
3-085	36	46	33	56	43	53	4	39	
3-099	45	12	29	42	10	22	14	40	
6-638	46	45	38	20	22	4	14	33	
8-28	34	35	43	42	22	6	30	24	

structure ID		16 conventional descriptors							
3-086	13	9	58	2	19	0	4	60	
3-089	11	19	10	6	3	0	1	2	
4-117	3	46	22	0	2	4	0	17	
3-101	11	0	3	5	2	23	24	10	
3-102	23	54	10	0	6	0	2	11	
6-652	62	1	13	0	2	9	8	23	
3-088	4	56	62	4	13	0	2	19	
4-115	7	10	60	0	2	8	5	18	
3-068	46	11	6	0	2	8	3	0	
3-069	3	8	0	9	46	4	4	0	
3-087	10	19	11	0	6	60	2	1	
4-116	11	17	2	4	0	24	10	0	
3-103	62	16	13	3	2	10	30	0	
6-146	0	62	7	74	2	9	8	0	
3-100	6	23	11	60	24	5	6	0	
5-152	7	53	7	14	1	43	0	8	
3-085	9	0	2	26	2	4	1	10	
3-099	11	6	8	23	3	5	54	0	
6-638	54	1	11	26	9	0	8	6	
8-28	1	1	6	19	12	2	10	0	

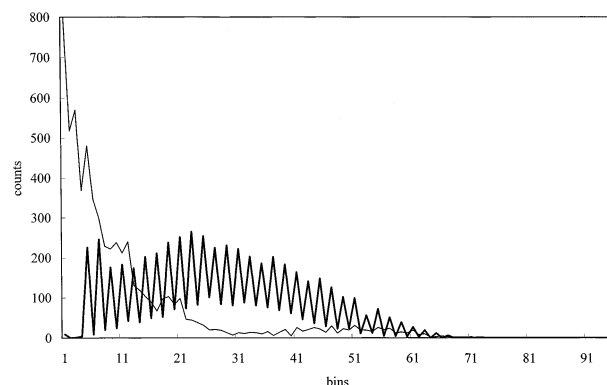
respectively (Table 1). The correspondence atoms used for aligning the structures by the SYBYL FIT module<sup>32</sup> were chosen from the three basic structure units used for dividing the structures (Figure 1). The SYBYL atomic types and ID number of these correspondence atoms were C2(1), C3(4), C3(5); C3(2), N3(4), C2(9), C3(1); and C3(20), C3(15), C3(13) for series A, B, and C, respectively (Figure 1). The root-mean-square values obtained after superposition of a pair of structures were from 0.001 to 0.044. The CoMFA analysis results for all the three sets of aligned structures were presented as plots of predicted pKi against measured pKi as shown in Figure 2a–c and on which the corresponding cross-validated  $q^2$  computed was 0.53, 0.54, and 0.62, respectively.

To account for the great structure diversity, a set of 3D convex hull or conventional descriptors was used to represent each structure. A comparison of a set of eight 3D convex hull or 16 conventional descriptors computed for the 20 most active inhibitors was given in Table 3. Note that the descriptors in each set were in random order since they were randomly picked from a frequency sorted list (for the 3D convex hull ones) or from the list given in Table 2 (for the 16 conventional ones). The information content retained by the 3D convex hull or the 16 conventional descriptors was



**Figure 2.** (a) The predicted pKi is plotted against the measured pKi for the 37 series A inhibitors selected (Table 1). (b) The predicted pKi is plotted against the measured pKi for the 19 series B inhibitors selected (Table 1). (c) The predicted pKi is plotted against the measured pKi for the 23 series C inhibitors selected (Table 1).

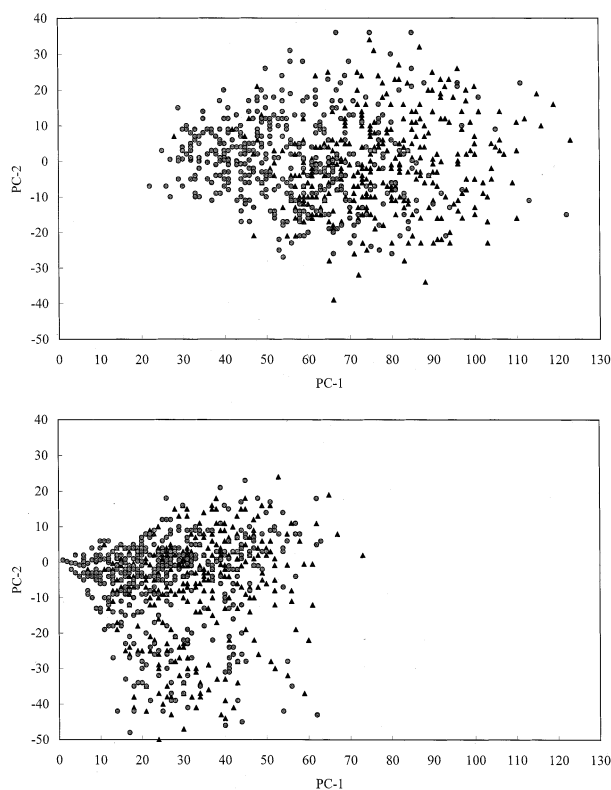
compared by computation of the Shannon entropy (SE) as that described by Godden and Bajorath<sup>36</sup> for each of them. To perform the comparison, we have generated 100 sets of descriptors and each containing 8 random descriptors for each compound. As revealed by the larger SE value computed, the information content retained by the 3D convex hull descriptors was slightly better than that retained by the 16 conventional ones (Figure 3). Since each inhibitor or inactive analogue was represented with 4 to 8 descriptors and the total number of compounds classified was 782, the dimensions of the data matrices classified vary from  $4 \times 782$  to  $8 \times 782$ . The results of a principal components analysis on the two  $8 \times 782$  matrices listed partly in Table 3 were presented in Figure 4a,b as plots of the second principal components (PC-2) versus the first ones (PC-1). To perform further classification using the fuzzy c-means algorithm, the two principal components were divided into two groups for the 345 active or 437 inactive inhibitors, respectively. In both



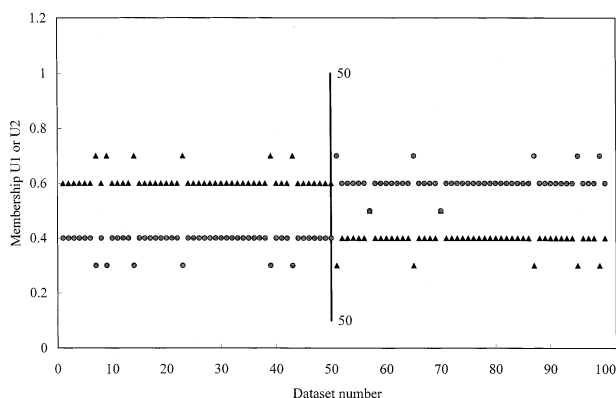
**Figure 3.** Computation of the Shannon entropy (SE) for the 100 sets of 3D convex hull and conventional descriptors for each compound being represented with 8 random descriptors. The distribution of the 3D convex hull descriptors was represented with a thick line, while that of the 16 conventional ones was represented with a thin line. The SE value computed for the 3D convex hull descriptors was 5.45, while that computed for the 16 conventional ones was 4.80.

cases, the two principal components account for more than 90% of the total variance of the original data set. These plots show that the two principal components of the 3D convex hull descriptors were more elliptically scattered than those of the 16 conventional ones. No apparent separation can be seen between the two groups at the stage of classification using neither type of the descriptors. However, at the stage of classification each of the original  $8 \times 782$  data matrices was transformed to two uncorrelated  $2 \times 782$  ones.

To test the feasibility of the fuzzy c-means algorithm,<sup>28</sup> we have applied it to the classification of a known data set, the Fisher's iris data set.<sup>37</sup> The data set consists of three  $4 \times 50$  data matrices obtained from measuring sepal and petal length and width of 50 plants of the three iris species, namely, iris setosa, iris versicolor, and iris virginica.<sup>37</sup> After transformation by principal components analysis, the data set of iris setosa was well separated from those of the iris versicolor and iris virginica (data not shown here). Therefore, the fuzzy c-means algorithm was used to perform further classification on the uncorrelated data sets of iris versicolor and iris virginica. The fuzzy membership functions U1 and U2 computed from eq 2 were used to designate the classification results of the original values of PC-1 and PC-2. As presented in Figure 5, the classification causes an evident switching in values of U1 and U2 right at the group boundary between the two data sets. The classification result was perfect since there was no outlier or no preliminary switching in fuzzy membership functions on both sides of the boundary. The classification results by the fuzzy c-means algorithm<sup>28</sup> on the two principal components of the 3D convex hull descriptors and the 16 conventional descriptors (Figure 4a,b) were presented in Figure 6a,b, respectively. There is an apparent boundary at the compound number 345 at which most of the U1's and U2's begin to switch. The values of most of the U1's or U2's of the active inhibitor set were around 0.9 or 0.1, while those of the inactive inhibitor set were around 0.1 or 0.9 (Figure 6a). A member in the active inhibitor set was identified as an outlier if its U1 or U2 was less than or greater than 0.5. The number of outliers identified for the active inhibitor set was 64 or the classification accuracy was 81% (Figure 6a or Table 4). The similar classification results using the 16 conventional descriptors

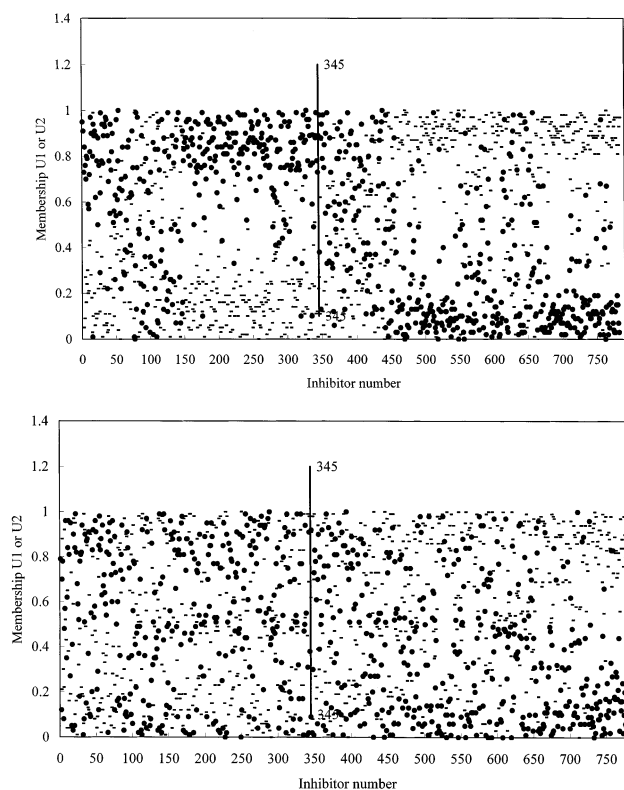


**Figure 4.** (a) The second principal component PC-2 is plotted against the first principal component PC-1 for the  $8 \times 782$  data matrix of 3D convex hull descriptors partially listed in Table 3 and uncorrelated with the principal components analysis. The two PC values are divided for the active inhibitors (represented with circles) and the inactive analogues (represented with triangles) for plotting. (b) The second principal component PC-2 is plotted against the first principal component PC-1 for the  $8 \times 782$  data matrix of the 16 conventional descriptors partially listed in Table 3 and uncorrelated with the principal components analysis. The two PC values are divided for the active inhibitors (represented with circles) and the inactive analogues (represented with triangles) for plotting.



**Figure 5.** The fuzzy c-means classification on the uncorrelated data sets of iris versicolor and iris virginica by the principal components analysis. The membership functions computed for each data set are represented with U1 (triangles) and U2 (circles). A line is drawn at the number of data set at 50 to highlight the boundary between the two iris plants.

for the two inhibitor sets were presented in Figure 6b. Except that no clear-cut boundary in values of U1's or U2's was observed, the number of outliers identified for the active inhibitor set was 140 or the classification accuracy was 59% for the case (Figure 6b or Table 4). A comparison for values of the Xie-Beni validity index<sup>35</sup> searched from the best



**Figure 6.** (a) The fuzzy c-means classification on the uncorrelated  $8 \times 782$  data matrix (Figure 3a) of the 3D convex hull descriptors partially listed in Table 3. The membership functions computed for each inhibitor are represented with U1 (circles) and U2 (dashes). A line is drawn at the inhibitor number 345 to highlight the boundary between the active and inactive inhibitor sets. (b) The fuzzy c-means classification on the uncorrelated  $8 \times 782$  data matrix (Figure 3b) of the 16 conventional descriptors partially listed in Table 3. The membership functions computed for each inhibitor are represented with U1 (circles) and U2 (dashes). A line is drawn at the inhibitor number 345 to highlight the boundary between the active and inactive inhibitor sets.

**Table 4.** Comparison of Number of Outliers Identified, Averaged Misclassified Frequency, and the Xie-Beni Validity Index<sup>35</sup> Computed for the 345 Active Inhibitors with Each Being Represented with Various Numbers of 3D Convex Hull or 16 Conventional Descriptors

no. of descriptors	3D convex hull descriptor			16 conventional descriptors		
	no. of outliers in active set	av mis-classified freq	Xie-Beni validity	no. of outliers in active set	av mis-classified freq	Xie-Beni validity
8	64 <sup>a</sup>	131 <sup>b</sup>	153 <sup>b</sup>	140 <sup>a</sup>	188 <sup>b</sup>	329 <sup>b</sup>
7	64	132	153	142	193	332
6	65	136	158	145	195	341
5	79	140	184	145	204	338
4	84	146	186	147	216	345

<sup>a</sup> The best number of outliers in the active set was searched from classification of thousands of multivariate data matrices generated for each number of descriptors used. <sup>b</sup> These were averaged from the classification of thousands of multivariate data matrices generated as described in *a*.

classification result for various numbers of the 3D convex hull descriptors with that for the 16 conventional ones used was given in Table 4. The corresponding number of outliers identified in the active inhibitor set for each descriptor set was also listed in the table. Since the Xie-Beni validity index<sup>35</sup> was searched from minimization of an objective





**Table 6.** Distribution of Frequency of the Convex Hull Vertex Atoms Involving in the Computation of Various 3D Convex Hull Descriptors Counted for the (a) 56 and (b) 289 Active Inhibitors Frequently Misclassified by the Fuzzy c-Means Algorithm

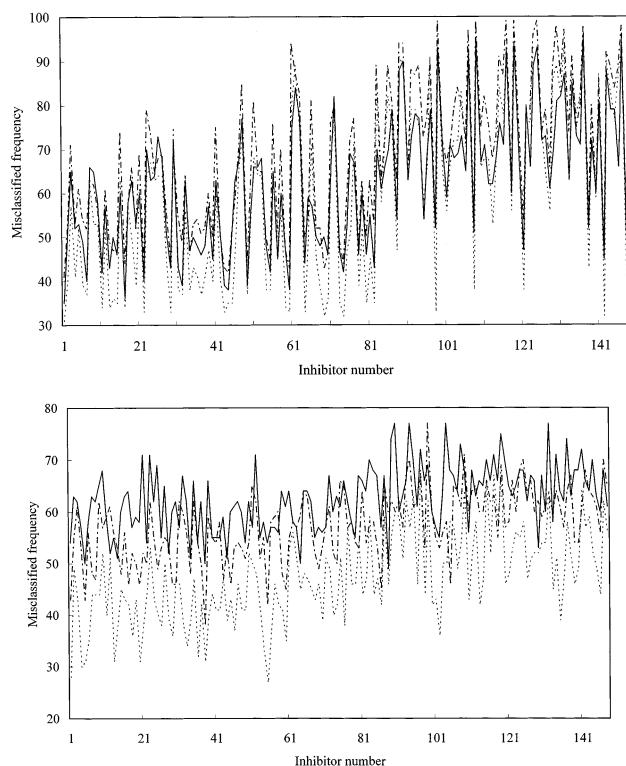
	O.3	C.ar	C.3	O.2	H	N.3	I	Cl	Br	C.2	F	N.ar	C.1	N.am	N.2	S.3	S.2	N.1
(a) 56 Active Inhibitors																		
4	1	60	2	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	1	10	8	1	0	0	1	0	1	3	1	0	0	1	0	0	0	0
7	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	1	14	4	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	3	9	3	0	0	0	0	0	0	0	0	0	3	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	2	4	3	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0
13	8	0	4	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	19	39	3	3	0	0	0	0	0	2	0	0	0	0	0	0	0	0
15	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	17	20	4	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
17	6	3	11	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
18	25	17	12	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0
19	5	6	4	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	11	30	10	4	1	0	0	2	0	3	1	0	1	0	0	0	0	0
21	8	4	10	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	4	25	8	3	0	0	0	0	0	8	0	1	0	2	0	0	0	0
23	11	14	10	2	0	0	0	0	0	1	0	0	1	0	0	0	0	0
24	10	34	11	4	1	1	0	0	0	9	0	1	0	1	0	0	0	0
25	15	26	8	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
26	9	30	13	0	0	0	1	1	0	6	0	0	0	0	0	0	0	0
27	10	19	7	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0
28	9	32	10	2	0	0	0	2	1	5	1	0	0	1	0	0	0	0
29	7	25	11	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
30	6	19	18	3	2	0	0	0	0	5	0	0	0	1	0	0	0	0
31	7	42	12	0	1	0	0	0	0	2	0	1	1	0	0	0	0	0
32	3	37	21	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
33	7	24	9	1	0	0	0	2	0	0	2	0	0	0	0	0	0	0
34	2	20	14	0	0	2	0	2	1	4	0	0	0	0	0	0	0	0
35	1	41	2	1	1	0	0	1	0	0	0	1	0	0	0	0	0	0
36	1	24	7	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0
37	0	19	3	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
38	0	12	3	1	1	0	0	0	0	4	0	0	0	0	0	0	0	0
39	1	17	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
40	0	7	5	0	2	0	0	0	0	3	3	0	0	1	0	0	0	0
41	2	7	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
42	0	0	4	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0
43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
49	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
51	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
52	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
53	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
61	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
62	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
64	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
66	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
67	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
68	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
71	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
73	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(b) 289 Active Inhibitors																		
4	1	96	2	0	0	0	0	0	0	122	0	2	0	0	2	2	1	0
5	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	2	0	0
6	9	47	46	4	2	2	0	0	0	23	3	0	0	4	1	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

**Table 6** (Continued)

	O.3	C.ar	C.3	O.2	H	N.3	I	Cl	Br	C.2	F	N.ar	C.1	N.am	N.2	S.3	S.2	N.1
(b) 289 Active Inhibitors																		
8	7	24	21	7	2	2	0	0	0	6	0	1	0	1	1	0	0	0
9	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	4	18	12	5	1	0	0	1	0	4	1	1	0	4	0	0	0	0
11	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
12	4	19	25	7	6	2	0	0	0	2	1	1	0	0	0	1	1	0
13	3	3	4	2	1	0	0	0	0	3	0	0	0	0	2	0	0	0
14	12	35	14	5	3	0	0	0	0	7	0	0	0	0	1	1	0	0
15	0	7	0	2	0	0	0	0	0	2	0	0	0	0	4	0	0	0
16	33	48	17	5	2	0	0	0	0	0	0	2	0	0	1	0	0	0
17	2	9	2	3	0	0	0	0	0	0	0	0	0	1	1	0	0	0
18	38	49	51	4	1	1	0	0	1	55	1	0	0	0	0	0	0	0
19	0	6	5	1	0	0	0	0	0	1	1	0	0	1	3	0	0	0
20	7	39	78	4	1	0	0	0	0	78	0	0	0	0	0	0	0	0
21	2	14	8	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0
22	20	50	90	9	3	0	0	0	1	69	0	0	0	2	2	0	0	0
23	2	8	5	0	0	0	0	1	0	8	0	0	0	0	0	0	0	0
24	17	87	51	13	2	0	0	0	0	92	0	0	0	0	1	1	0	0
25	16	32	2	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0
26	30	83	85	10	4	1	0	2	1	69	0	1	0	1	1	3	0	0
27	18	48	2	0	1	0	0	0	0	9	0	0	0	0	0	0	0	0
28	33	88	72	17	11	0	0	7	0	54	0	1	0	6	1	1	0	0
29	13	26	7	7	1	0	0	0	1	4	0	0	0	1	0	0	0	0
30	10	69	50	13	4	2	0	1	0	92	3	0	0	2	2	1	0	0
31	18	37	10	11	3	0	0	1	0	7	1	0	0	0	4	1	0	0
32	13	70	87	15	0	1	0	0	0	77	5	2	0	2	2	1	1	0
33	15	66	21	5	5	0	0	0	0	8	2	1	0	0	3	0	0	0
34	11	77	80	9	3	0	0	0	1	66	1	1	0	2	3	1	0	0
35	17	73	15	9	7	1	0	1	0	9	1	1	0	0	0	1	0	0
36	18	103	90	12	10	0	0	3	1	78	2	2	0	3	2	5	0	1
37	25	94	20	9	3	0	0	0	0	11	3	0	0	1	3	2	0	0
38	11	79	91	14	4	0	0	4	0	72	0	2	0	5	1	0	2	0
39	11	63	29	3	2	0	0	0	0	4	2	0	0	1	1	0	1	0
40	15	50	78	7	2	0	0	2	3	72	2	0	0	2	1	0	0	0
41	9	45	8	2	5	1	0	1	0	8	1	0	0	0	3	4	0	0
42	8	52	50	10	8	0	0	1	1	73	2	0	0	0	2	0	0	0
43	6	28	17	2	3	1	0	1	0	11	0	1	0	0	1	1	0	0
44	3	50	53	7	5	0	0	1	2	69	0	1	0	7	1	1	1	0
45	3	29	11	2	1	0	0	0	0	4	0	0	0	0	1	0	0	0
46	4	38	40	7	4	0	0	2	0	66	2	0	0	2	2	1	0	0
47	2	30	8	0	1	0	0	0	0	9	0	1	0	0	0	0	0	0
48	7	42	50	3	1	1	0	4	0	80	1	3	0	2	2	2	0	0
49	3	39	2	2	0	0	0	1	0	12	0	0	0	0	1	0	0	0
50	7	42	43	3	2	0	0	2	1	73	0	2	0	3	1	1	0	0
51	4	10	2	0	1	0	0	1	1	1	1	3	0	0	0	0	0	0
52	6	40	28	2	3	1	0	1	1	78	1	0	0	1	1	2	0	0
53	2	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
54	7	32	10	1	1	0	0	1	0	64	0	1	0	1	1	1	0	0
55	0	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
56	5	40	16	0	1	0	0	1	1	85	1	2	0	0	1	0	0	0
57	0	2	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0
58	7	37	10	0	5	0	0	1	1	67	0	1	0	0	0	0	0	0
60	1	20	18	1	3	0	0	1	2	25	0	1	0	0	0	0	0	0
61	1	3	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
62	4	17	10	4	1	0	0	1	0	20	0	0	0	0	0	0	0	0
64	2	8	9	1	0	0	0	0	0	13	0	0	0	0	0	0	0	0
65	0	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
66	3	4	6	0	2	0	0	0	0	6	0	0	0	0	0	0	0	0
67	0	2	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0
68	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
70	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
71	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
73	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

the frequency of vertex atom C.2 involving the computation of a 3D convex hull descriptor since such a frequency in the 289 set was much larger than that in the 56 one. However, most of the 3D convex hull descriptors computed involving vertex atoms O.3, C.ar, and C.3 for both sets of the active inhibitors (Table 6a,b). To further demonstrate that the active inhibitors can be effectively represented by the 3D convex hull descriptors, the misclassified frequencies counted were

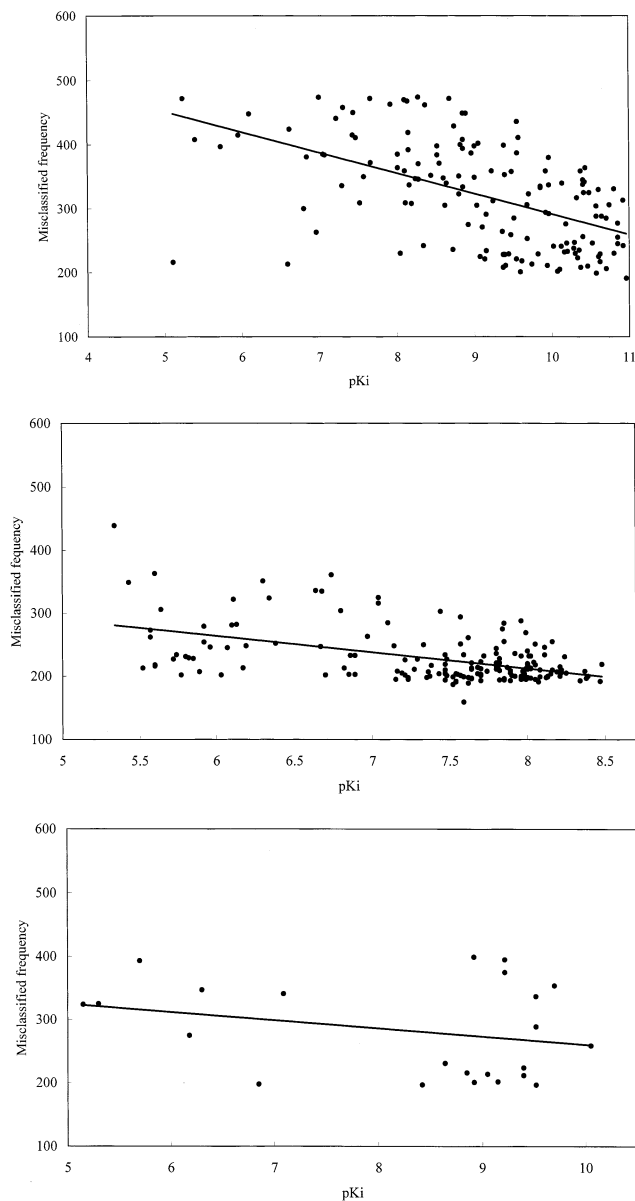
grouped for series A, B, and C inhibitors and are plotted against values of the corresponding p*K*<sub>i</sub> (Table 1) for each inhibitor series as shown in Figure 8a–c, respectively. The overall data trend on each plot is guided by a linear regression line. In general, these plots show that inhibitors of higher activity were less likely to be misclassified since the misclassified frequency reduces with the increase of p*K*<sub>i</sub>. However, such a trend for series B or C inhibitors was less



**Figure 7.** (a) The frequency of being misclassified by the fuzzy c-means algorithm for each of the 148 active inhibitors of series A (Table 1) represented with 4 (solid line), 6 (dot line), and 8 (dash dot line) 3D convex hull descriptors. (b) The frequency of being misclassified by the fuzzy c-means algorithm for each of the 148 active inhibitors of series A (Table 1) represented with 4 (solid line), 6 (dot line), and 8 (dash dot line) conventional descriptors.

evident than that for series A ones since most of the inhibitors in the formers were less active than those in the latter (Table 1).

Since there was an apparent switch in membership functions computed at the group boundary, the rule that U1 or U2 must be greater or less than 0.5 was used to identify the outliers in the inactive inhibitor sets (Figure 6a). The number of outliers identified for the set is 63. Although these were the inactive analogues searched from the MDL/ISIS database,<sup>25</sup> they were classified by the fuzzy c-means algorithm<sup>28</sup> as having the same structure feature as the majority members (289 inhibitors) of the active set. The distribution of frequency of each 3D convex hull descriptor counted for each compound in these two sets was compared in Figure 9. To easily examine the difference between these two distributions, each frequency counted was normalized with the largest one searched from each set of the descriptors computed for each inhibitor set. Apparently, the distribution of the 3D convex hull descriptors computed for both the inhibitor sets was nearly the same except that between descriptors 10 to 19 or 52 to 62 (Figure 9). These 20 descriptors were then used to classify a new compound set formed by the 289 active plus 63 inactive inhibitors. We first selected 3 to 8 descriptors randomly out of the 20 ones to form a new descriptor set. The numbers of inactive inhibitors and active ones in the new compound set which bear the new descriptor set were counted and the ratio between them were computed. The best classification result was identified as the one which gave the smallest ratio computed for all the new descriptor sets used. The clas-



**Figure 8.** (a) The frequency of being misclassified by the fuzzy c-means algorithm for each of the active inhibitors of series A (Table 1) is plotted against the corresponding pKi. (b) The frequency of being misclassified by the fuzzy c-means algorithm for each of the active inhibitors of series B (Table 1) is plotted against the corresponding pKi. (c) The frequency of being misclassified by the fuzzy c-means algorithm for each of the active inhibitors of series C (Table 1) is plotted against the corresponding pKi.

sification produced two results for which the ratios computed were the same but the corresponding new descriptor sets used were different as listed in Table 7. While there was a difference in the number and content of active inhibitors classified in each of the results, the ratio computed for each was 0.13 (Table 7). Although some of the highly active inhibitors of series A or C were also identified, most of the active inhibitors identified in these final classifications were from series B (Tables 1 and 7). The structures of 19 inactive inhibitors accompanying with the structures of the three most active inhibitors of each series (Tables 1 and 7) identified from the final classification were shown in Figure 10. It appeared that most of the inactive inhibitors contain nearly the same structural or topological features as those appearing on the three highly active ones. This was in accord with the

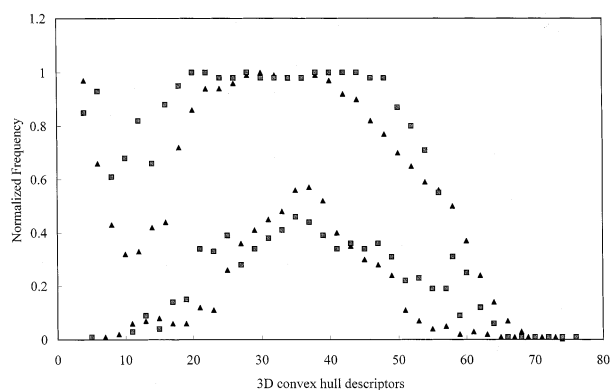


**Table 7.** Two Best Classification Results Obtained from Further Classification Using Several Sets of 3D Convex Hull Descriptors Formed through Examining the Distribution of Frequency (Figure 9) of Each 3D Convex Hull Descriptor Counted for the 289 Correctly Classified Active and 63 Frequently Misclassified Inactive Analogues (a: Descriptor Set, b: Inhibitor ID)

54	56	58 <sup>a</sup>							
3-86 <sup>b</sup>	3-89	3-87	3-103	6-146	3-85	3-91	3-83	3-93	3-97
3-90	3-84	3-82	2-66	3-81	12-283549	10-282870	12-283516	10-283012	12-283471
12-283134	12-283569	12-283550	12-283143	12-283490	13-6	10-282215	10-283004	10-282664	10-282518
13-16	10-282542	10-283005	10-282365	10-282981	10-282456	10-282701	12-283537	10-282457	11-282665
12-283209	12-283265	10-282666	10-282540	10-283013	12-283260	12-283412	13-7	12-283532	10-283011
10-283055	10-282395	10-283010	10-282822	10-282311	10-282412	12-283341	10-282350	11-283253	10-282753
10-282939	12-283263	10-282351	10-282985	12-283451	12-283374	10-282714	10-282756	10-282796	10-283043
10-283045	10-282747	10-282797	12-283206	12-283254	12-283353	12-283410	10-282366	10-282558	11-282730
13-17	10-282916	12-283519	10-282429	10-282828	12-283262	10-282779	10-282826	13-18	10-282547
10-282915	12-283342	10-282480	12-283239	13-12	10-282713	11-283489	10-282455	10-282824	12-283436
10-282632	10-282825	10-282944	10-282392	10-282423	12-283441	13-9	12-283440	10-282349	12-283407
10-282539	10-282310	10-282390	12-283343	12-283365	12-283498	10-282700	13-13	13-14	10-282479
10-282391	12-283266	12-283521	12-283267	11-282364	10-282450	13-15	12-283406	12-283573	12-283321
10-282807	12-283249	10-282808	12-283567	12-283478	12-283497	12-283434	14-38	15-9	15-13
15-10	15-4	15-5	A4	a5	a10	a57	a112	a122	a123
a124	a125	A126	A127	a128	a129	a130	a131	a139	c19
c96	c184								

52	54	56	58 <sup>a</sup>						
3-86 <sup>b</sup>	3-89	3-91	3-83	3-93	3-97	3-90	3-82	2-66	3-81
12-283549	10-282870	12-283516	10-283012	12-283471	12-283134	12-283569	12-283550	12-283143	12-283490
13-6	10-282215	10-283004	10-282664	10-282518	13-16	10-282542	10-283005	10-282365	10-282981
10-282456	10-282701	12-283537	10-282457	11-282665	12-283209	12-283265	10-282666	10-282540	10-283013
12-283260	12-283412	13-7	12-283532	10-283011	10-283055	10-282395	10-283010	10-282822	10-282311
10-282412	12-283341	10-282350	11-283253	10-282753	10-282939	12-283263	10-282351	10-282985	12-283451
12-283374	10-282714	10-282756	10-282796	10-283043	10-283045	10-282747	10-282797	12-283206	12-283254
12-283353	12-283410	10-282366	10-282558	11-282730	13-17	10-282916	12-283519	10-282429	10-282828
12-283262	10-282779	10-282826	13-18	10-282547	10-282915	12-283342	10-282480	12-283239	13-12
10-282713	11-283489	10-282455	10-282824	12-283436	10-282632	10-282825	10-282944	10-282392	10-282423
12-283441	13-9	12-283440	10-282349	12-283407	10-282539	10-282310	10-282390	12-283343	12-283365
12-283498	10-282700	13-13	13-14	10-282479	10-282391	12-283266	12-283521	12-283267	11-282364
10-282450	13-15	12-283406	12-283573	12-283321	10-282807	12-283249	10-282808	12-283567	12-283478
12-283497	12-283434	14-38	15-9	15-13	15-10	15-4	15-5	a4	a5
a10	A57	A112	A122	a123	A124	a125	a126	a127	a128
a129	a130	A131	A139	c19	C96	c184			

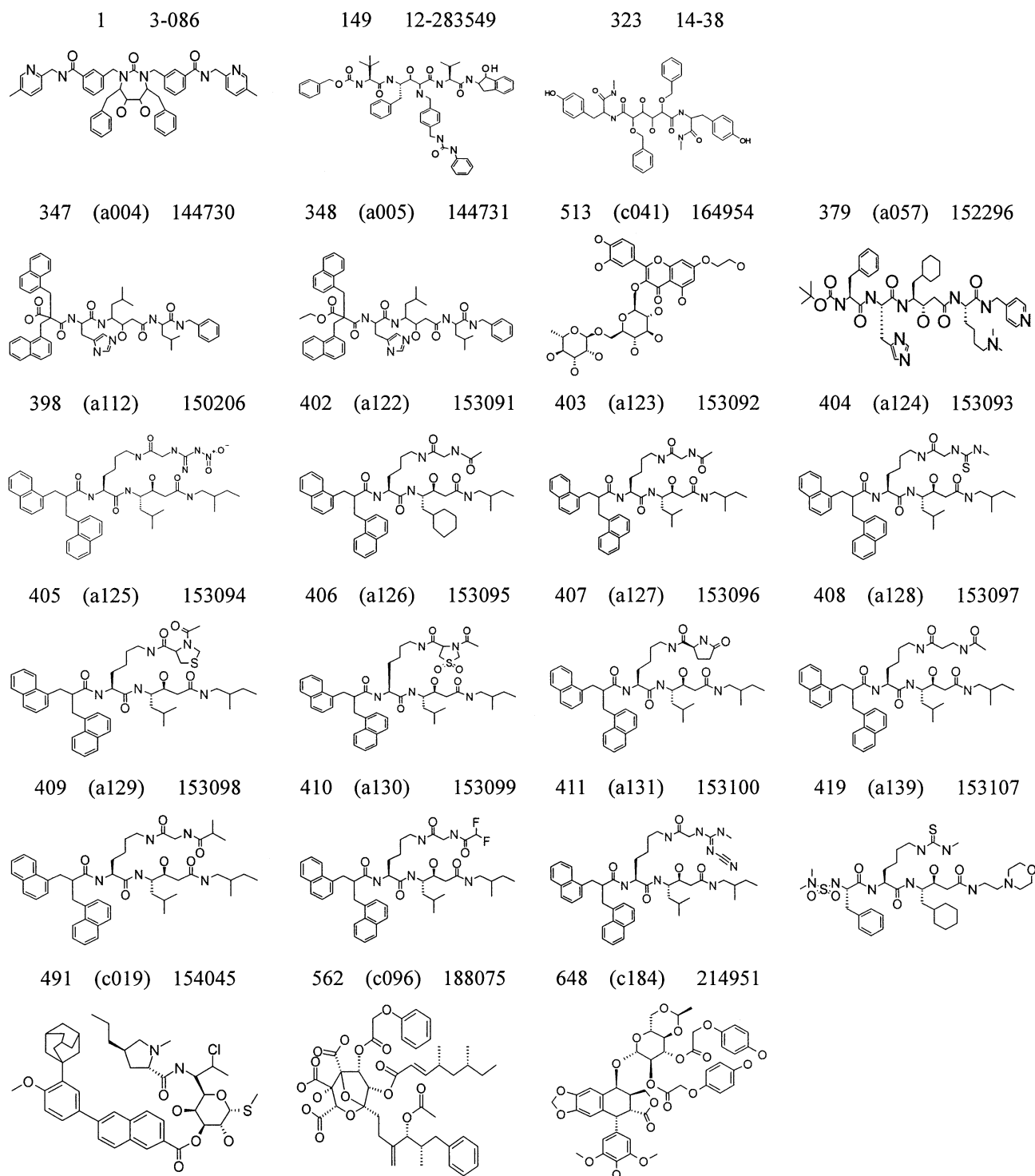
**Figure 9.** The distribution of normalized frequency of each 3D convex hull descriptor counted for descriptor values from 4 to 77 for the 289 correctly classified active inhibitors (squares) and 63 misclassified inactive analogues (triangles).

fact that the 3D convex hull descriptors used to classify them were a mixed type of structural (convex hull vertices) and topological (molecular path lengths) descriptors.

## CONCLUSION

The number of 3D convex hull or conventional descriptors used in the classification process was set as from 4 to 8 because no obvious improvement on the classification result was obtained by using more numbers of descriptors. A 3D convex hull descriptor computed for a molecule can be

envisaged as a descriptor that can combine the most exterior (the convex hull vertices) and interior (the molecular path lengths among them) natures of the molecule. Therefore, it is possible to classify a set of molecules of greater structural similarity as we have presented here. The fundamental difference between the 3D convex hull and the 16 conventional descriptors used lies in the fact that better information content is retained by the former than by the latter. The fuzzy c-means algorithm<sup>28</sup> used here aims to identify compact, well-separated clusters. Informally, a compact cluster has a "ball-like" shape. Deviation of clusters from the compactness will cause the classification result to deteriorate. A canonical analysis<sup>38</sup> has been performed for the  $8 \times 782$  data matrix of the 3D convex hull descriptors to generate two canonical vectors. However, the fuzzy c-means algorithm<sup>28</sup> fails to classify the clusters since all the data points projected by the canonical vectors tend to line up along a straight line (data not shown here). Unlike the classification using the linear discriminant function,<sup>29</sup> the fuzzy c-means algorithm<sup>28</sup> does not count on the distribution or covariance of the data set but only on the distance between any two points in a cluster and that between two clusters in different clusters. This would simplify the classification process and reduce the error as we have shown here. The classification using the fuzzy c-means algorithm<sup>28</sup> will be successful as long as the clusters are compact. To extend the fuzzy c-means algorithm, the Xie-Beni validity function<sup>35</sup> used for computing the distance between two cluster centers may be replaced



**Figure 10.** Structures of the 19 inactive analogues finally classified using the two sets of 3D convex hull descriptors listed in Table 7 are drawn with structures of the three most active inhibitors (Table 1) classified accompanying with them. The compound ID and the number in order of p*K*<sub>i</sub> arranged in each series (Table 1) for each active inhibitor are also shown. The compound ID, the series ID (Figure 1), and the compound number used in this study for each inactive analogue searched from the MDL/ISIS database<sup>25</sup> are also given.

with that for computing the ellipsoidal shape of clusters.<sup>39</sup> The fuzzy c-means algorithm<sup>28</sup> also imposes a sum-to-one constraint on the computation of cluster membership which may be relaxed using the probabilistic clustering algorithm developed by Krishnapuram and Keller.<sup>40</sup> However, it is also necessary to compromise the complexity of algorithm and the generality of data set to achieve a meaningful classification result. Since there were two initial prototypes (active or inactive inhibitor sets) chosen for classification, the

parameter *m* was set as 2 to limit the degree to which partial members of a cluster affect the classification result. The fuzzy c-means algorithm<sup>28</sup> is useful in applications where clusters touch or overlap since each data point can be classified to all clusters with some degree of membership. However, one should use an objective criterion to determine how good a partition is generated by the algorithm. Since the number of clusters was fixed as two, the only problem we encountered was whether there exists a natural grouping of the data set.

This was detected using the Xie-Beni validity function<sup>35</sup> since it was able to group those points that were too close to be separated in clusters together.

### ACKNOWLEDGMENT

This work is supported in part by a grant from the National Science Council, ROC (NSC90-2320-B007-001).

### REFERENCES AND NOTES

- (1) Tabachnick, B. G.; Fidell, L. S. *Using Multivariate Statistics*; Harper and Row: Philadelphia, 1983.
- (2) Dunn, W. J., III; Wold, S. A structure-carcinogenicity study of 4-nitroquinoline 1-oxides using the SIMCA method of pattern recognition. *J. Med. Chem.* **1978**, *21*, 1001–1007.
- (3) Bath, P. A.; Poirrette, A. R.; Willett, P.; Allen, F. H. Similarity searching in fields of three-dimensional chemical structures: comparison of fragment-based messages of shape similarity. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 141–147.
- (4) Viswanadhan, V. N.; Mueller, G. A.; Basak, S. C.; Weinstein, J. N. Comparison of a neural net-based QSAR algorithm (PCANN) with hologram- and multiple linear regression-based QSAR approaches: application to 1,4-dihydropyridine-based calcium channel antagonists. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 505–511.
- (5) Murcia-Soler, M.; Pérez-Giménez, F.; Nalda-Molina, R.; Salabert-Salvador, M. T.; Garcia-March, F. J.; Cercós-del-Pozo, R. A.; Garrigues, T. M. QSAR analysis of hypoglycemic agents using the topological indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1345–1354.
- (6) Godden, J. W.; Bajorath, J. Differential Shannon entropy as a sensitive measure of differences in database variability of molecular descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1060–1066.
- (7) Topliss, J. G.; Edwards, R. P. Chance factors in studies of quantitative structure–activity relationships. *J. Med. Chem.* **1979**, *22*, 1238–1244.
- (8) Randic, M.; Basak, S. C. A new descriptor for structure–property and structure–activity correlations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 650–656.
- (9) Randic, M.; Zupan, J. On interpretation of well-known topological indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 550–560.
- (10) Dury, L.; Latour, T.; Leher, L.; Barberis, F.; Vercauteren, D. P. A new graph descriptor for molecules containing cycles. application as screening criterion for searching molecular structures within large databases of organic compounds. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1437–1445.
- (11) Estrada, E.; Perdomo-López, I.; Torres-Labandeira, J. J. Combination of 2D-, 3D-connectivity and quantum chemical descriptors in QSPR. Complexation of  $\alpha$ - and  $\beta$ -cyclodextrin with benzene derivatives. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1561–1568.
- (12) Estrada, E.; Molina, E.; Perdomo-López, I. Can 3D structural parameters be predicted from 2D (topological) molecular descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1015–1021.
- (13) Famini, G. R.; Wilson, L. Y. Using theoretical descriptors in linear salvation energy relationships. *Theor. Comput. Chem.* **1994**, *1*, 213–241.
- (14) Suzuki, T.; Ide, K.; Ishida, M.; Shapiro, S. Classification of environmental extrogens by physicochemical properties using principal component analysis and hierarchical cluster analysis. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 718–726.
- (15) Baker, J. An algorithm for the location of transition states. *J. Comput. Chem.* **1986**, *7*, 385–395.
- (16) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley-Interscience: New York, 2000.
- (17) Hultén, J.; Bonham, N. M.; Nilroth, U.; Hansson, T.; Zuccarello, G.; Bouzide, A.; Åqvist, J.; Classon, B.; Danielson, U. H.; Karlén, A.; Kvarnström, I.; Samuelsson, B.; Hallberg, A. Cyclic HIV-1 protease inhibitors derived from mannitol: synthesis, inhibitory potencies, and computational predictions of binding affinities. *J. Med. Chem.* **1997**, *40*, 885–897.
- (18) Alterman, M.; Björnsen, M.; Mühlman, A.; Classon, B.; Kvarnström, I.; Danielson, H.; Markgren, P. O.; Nilroth, U.; Unge, T.; Hallberg, A.; Samuelsson, B. Design and synthesis of new potent C2-symmetric HIV-1 protease inhibitors. Use of L-mannaric acid as a peptidomimetic scaffold. *J. Med. Chem.* **1998**, *41*, 3782–3792.
- (19) Nugiel, D. A.; Jacobs, K.; Cornelius, L.; Chang, C.; Jadhav, P. K.; Holler, E. R.; Klabe, R. M.; Bachelier, L. T.; Cordova, B.; Garber, S.; Reid, C.; Logue, K. A.; Gorey-Feret, L. J.; Lam, G. N.; Erickson-Viitanen, S.; Seitz, S. P. Improved P1/P1' substituents for cyclic urea based HIV-1 protease inhibitors: synthesis, structure–activity relationship, and X-ray crystal structure analysis. *J. Med. Chem.* **1997**, *40*, 1465–1474.
- (20) Kroemer, R. T.; Ettmayer, P.; Hecht, P. 3D-quantitative structure–activity relationship of human immunodeficiency virus type-1 proteinase inhibitors: comparative molecular field analysis of 2-hetero-substituted statine derivatives-implications for the design of novel inhibitors. *J. Med. Chem.* **1995**, *38*, 4917–4928.
- (21) Scholz, D.; Billich, A.; Charpiot, B.; Ettmayer, P.; Lehr, P.; Rosenwirth, B.; Schreiner, E.; Gstach, H. Inhibitors of HIV-1 proteinase containing 2-heterosubstituted 4-amino-3-hydroxy-5-phenylpentanoic acid: synthesis, enzyme inhibition, and antiviral activity. *J. Med. Chem.* **1994**, *37*, 3079–3089.
- (22) Alterman, M.; Andersson, H. O.; Garg, N.; Ahlsén, G.; Lövgren, L.; Classon, C.; Danielson, U. H.; Kvarnström, I.; Vrang, L.; Unge, T.; Samuelsson, B.; Hallberg, A. Design and fast synthesis of C-terminal duplicated potent C2-symmetric P1/P1'-modified HIV-1 protease inhibitors. *J. Med. Chem.* **1999**, *42*, 3835–3884.
- (23) Debnath, A. K. Three-dimensional quantitative structure–activity relationship study on cyclic urea derivatives as HIV-1 protease inhibitors: application of comparative molecular field analysis. *J. Med. Chem.* **1999**, *42*, 249–259.
- (24) Prabhakar K.; Jadhav, P. A.; Woerner, F. J.; Chang, C. H.; Garber, S. S.; Anton, E. D.; Bachelier, L. T. Cyclic urea amides: HIV-1 protease inhibitors with low nanomolar potency against both wild type and protease inhibitor resistant mutants of HIV. *J. Med. Chem.* **1997**, *40*, 181–191.
- (25) The MDL/ISIS database is installed at the National Center for High Performance Computing, Taiwan, ROC. The URL is <http://saturn.nchc.gov.tw:9091/cds>.
- (26) Lin, T. H.; Yu, Y. S.; Chen, H. J. Classification of some active compounds and their inactive analogues using two three-dimensional molecular descriptors derived from computation of three-dimensional convex hulls for structures theoretically generated for them. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1210–1221.
- (27) Selim, S. Z.; Kamel, M. S. On the mathematical and numerical properties of the fuzzy c-means algorithm. *Fuzzy Sets Systems* **1992**, *49*, 181–191.
- (28) Yen, J.; Langari, R. *Fuzzy Logic Intelligence, Control, and Information*; Prentice Hall: New Jersey, 1999.
- (29) James, M. *Classification Algorithms*; Collins: London, 1985.
- (30) Lam, P. Y.; Ru, Y.; Jadhav, P. K.; Alrich, P. E.; DeLucca, G. V.; Eyeremann, C. J.; Chang, C. H.; Emmett, G.; Holler, E. R.; Daneker, W. F.; Li, L.; Confalone, P. N.; McHugh, R. J.; Han, Q.; Li, R.; Markwalder, J. A.; Seitz, S. P.; Sharpe, T. R.; Bachelier, L. T.; Rayner, M. M.; Klabe, R. M.; Shum, L.; Winslow, D. L.; Kornhauser, D. M.; Hodge, C. N. Cyclic HIV protease inhibitors: synthesis, conformational analysis, P2/P2' structure–activity relationship and molecular recognition of cyclic ureas. *J. Med. Chem.* **1996**, *39*, 3514–3525.
- (31) Kuntz, I. D.; Meng, E. C.; Shoichet, B. K. Structure-based molecular design. *Acc. Chem. Res.* **1994**, *27*, 117–123.
- (32) SYBYL 6.7; The Tripos Associates: 1699 S. Hanley Rd., St. Louis, MO.
- (33) Miller, M.; Schneider, J.; Sathyanarayana, B. K.; Toth, M. V.; Marshall, G. R.; Clawson, L.; Selk, S.; Kent, S. B. H.; Wlodawer, A. Structure of the complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 angstroms resolution. *Science* **1989**, *246*, 1149–1152.
- (34) Xu, J.; Stevenson, J. Drug-like index: a new approach to measure drug-like compounds and their diversity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1177–1187.
- (35) Xie, X. L.; Beni, G. A validity measure for fuzzy clustering. *IEEE Trans. Pattern Anal. Machine Intell.* **1991**, *PAMI-13*, 841–847.
- (36) Godden, J. W.; Bajorath, J. Differential Shannon Entropy as a sensitive measure of differences in database variability of molecular descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1060–1066.
- (37) Fisher, R. A. The use of multiple measurements in taxonomic problems. *Ann. Eugenics* **1936**, *7*, 179–188.
- (38) Maxwell, A. E. *Multivariate Analysis in Behavioural Research*; Chapman and Hall: New York, 1977.
- (39) Gustafson, D. E.; Kessel, W. C. Fuzzy clustering with a fuzzy covariance matrix. *Proc. IEEE CDC* **1979**, 761–766.
- (40) Krishnapuram, R.; Keller, J. M. A possibilistic approach to clustering. *IEEE Trans. Fuzzy Systems* **1993**, *1*, 98–110.

CI0203747