Journal of Computational Science & Engineering 9(2013) 326-355

Journal of Computational Science & Engineering

Available at http://www.asocse.org ISSN 1710-4068

The influence of molecular lowest-energy conformation on the quality of the subsequent quantitative structure-activity relationship models

Jiazhong Li[a]*, Juanjuan He[a], Beilei Lei[b], Huanxiang Liu[a] and Paola Gramatica[c]

[a] School of Pharmacy, Lanzhou University, Donggang West Road 199, Lanzhou 730000, China
[b] Center of Bioinformatics, College of Life Sciences, Northwest A & F University, Xinong Road 22, Yangling 712100, China
[c] QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Theoretical and Applied Sciences, University of Insubria, via Dunant 3, Varese 21100, Italy

Article Information

Article history:
Received 13 September 2013
Revised 20 September 2013
Accepted 25 September 2013
Available online 10 October 2013

Keywords:
Quantitative structure-activity relationship (QSAR)
Global lowest-energy molecular conformation
Local lowest-energy molecular conformation
Multiple linear regression (MLR)

Abstract

A basic problem in QSAR modeling is how to obtain proper molecular conformations to generate descriptors. In pharmaceutical chemistry, it is ideal if the active conformation can be obtained, otherwise lowest-energy conformation is commonly used. There are different ways to minimize a molecule, but if it can be reasonably minimized to its global lowest-energy conformation instead of a local one is suspectable. Unfortunately this problem is generally neglected by most of the researchers. Considering that different conformations may influence the quality of the subsequent QSAR models, here the conformation searching process is used to ensure that a molecule is minimized to its global lowest-energy conformation, and accordingly QSAR models are evloved. The molecular conformations without searching process, maybe just local lowest-energy, are used as comparison. Three datasets with various structural complexities and flexibilities are investigated. For each dataset, six different lowest-energy conformations are generated and six model populations are established accordingly. The results indicate that different optimization processes can influence the quality of the QSAR models, and global and local lowest-energy conformations have different performance in QSAR modeling. Furthermore, to get the proper global lowest-energy molecular conformation is vital for the subsequent QSAR model development and new query prediction.

* Corresponding author: lijiazhong@lzu.edu.cn

1. Introduction

Quantitative structure-activity relationship (QSAR) is among the most practical tools in computational chemistry and chemoinformatics, which have been extensively utilized in a wide range of scientific disciplines, including medicinal chemistry, biology, environmental science and toxicology etc. [1]. This method is based on the assumption that the variance in the activities/properties of chemical compounds can be described by the variance in their molecular structures. The basic strategy of QSAR is to find the optimum quantitative relationship between the biological activity or physicochemical property and molecular structures. The result is a mathematical model that can be used to predict the activity/property values of unmeasured or even unknown compounds, and to discover and understand how molecular structure influences corresponding activity/property, and very importantly for chemical synthesis, to design and optimize the needed structures.

In a QSAR study, the molecular structures are usually quantified in terms of the molecular descriptors and then trained against their activities/properties. So molecular descriptor is an important intermediate correlating the structures with their target values in QSAR. It has been many years since QSAR paradigm first found its way into the practice. Originally, the used molecular descriptors were limited, such as Hammett parameter, oil/water partition coefficient etc. In the past few decades a wide variety of methods have been developed to derive structure descriptors for a molecule [2-4]. Gasteiger summarized the different ways to represent molecules [5]. From the simple fingerprint and topological distances to three-dimensional (3D) structure representation and molecular chirality, researchers can accurately express chemicals closer to the reality than before. Molecular interactions are 3D in nature and molecular models should treat chemicals as 3D entities [6]. Up to now, the known organic and organometallic compounds uploaded in CAS are more than 70 million [7], but only about 577, 833 molecules have been determined 3D structures by X-ray diffraction or NMR studies according to the 2012 release of the Cambridge Structural Database (CSD v5.33) [8]. So the 3D conformations of vast majority of the molecules are needed to be constructed by the researchers in a QSAR study. Furthermore, most molecules are quite flexible having single bonds that allow rotation yielding different torsional angles and thus providing different conformations. How to obtain the appropriate 3D expression of a molecule is a problem, which may have vital influences on the subsequent QSAR models.

Generally, researchers try to get the biologically active conformation of a molecule when modeling a biological activity. Modeling algorithms, such as Comparative Molecular Field Analysis (CoMFA) [9] and Comparative Molecular Similarity Indices Analysis (CoMSIA) [10], explore a large number of alignments to reach an optimum outcome which may be susceptible to violation of the Topliss and Edwards criteria for causality in QSAR models [11]. COmmon REactivity PAttern (COREPA) [12-13] is another way to deal with this problem. Instead of assuming the lowest-energy conformer as the active form, all the energetically-reasonable conformers are used to establish conformer distributions across the global and local stereoelectronic descriptors associated with the activity. Becker et al. [14] tried to build a conformation space and created a QSAR type descriptor to quantify the effect of conformation constraints on bioactivity which was shown to be in excellent correlation with the observed activity of the molecules. Mekenyan et al. developed a new approach called 3DGEN [15] based on a combinatorial procedure for a systematic search of conformational space. However the systematic approach was found to only provide good performance for relative small and rigid structures. Then they developed another approach named GAS to handle highly flexible chemicals [6-16], where a genetic algorithm (GA) was employed to minimize 3D similarity among the generated (known) conformers.

But up to now, the usage of lowest-energy conformation is still very common in the QSAR analysis. There are various ways to generate lowest-energy conformation: including molecular mechanics, semi-empirical method and quantum mechanics etc. In most published QSAR works, molecules were generally pre-optimized with molecular mechanics method and then a more precise optimization is done with semi-empirical or ab initio method. Recently many researchers have highlighted the usefulness of quantum mechanics method [17-21]. Puzyn et al. [22] has proved that with the newly developed semi-empirical method, it was unnecessary to optimize a molecule at the time- and resource-consuming quantum mechanical density functional theory (DFT) level. The debate about molecular conformation in QSAR modeling is still going on.

In this study, our intention is not to investigate which method to get the final molecular conformation is the best one universally. We focus on the influence of different lowest-energy conformations on the quality of subsequent QSAR models. Considering that different molecular complexities may have different influences on the final QSAR models, here we choose three datasets with various complexities. The theoretical descriptors are calculated in DRAGON program [23]. Multiple linear regression (MLR) method is used to build QSAR models based on the important descriptors selected by genetic algorithm.

2. Materials and methods

2.1 Data sets

The three endpoints used in the present study include both chemical property (the flux behavior through an artificial membrane [24-27]) and bioactivity (inhibition activities of a series of pan-Src Lck inhibitors [28-30] and HCV NS5B polymerase inhibitors [31]). The summarized information for each data set are listed in Table 1. The detailed molecular structures and corresponding target values are summarized in Table 2-4.

Table 1 Summary of data sets used in this study.

No.	no. compounds	Data set	Endpoint	Reference
1(SMF)	256	membrane flux	silastic membrane flux (logJ)	[24-27]
2(LckI)	105	Lck inhibitors	Inhibition activity (IC_{50})	[28-30]
3(NS5BI)	67	NS5B inhibitors	Inhibition activity (IC_{50})	[31]

Dataset 1 (SMF) contains 256 data for flux through an artificial silastic membrane (logJ). Chen et al. reported this series of steady-state flux of compounds through a polydimethylsiloxane membrane at 30 °C [25-26]. The structures contained in this dataset were generally very simple like aromatics and hexatomic ring etc. Such flux measurements are important as they can be related to the flux of compounds through the skin [27]. These data were compiled by Cronin et al. [24] from the original references

Table 2 The molecular structures and corresponding logJ of dataset 1 (SMF).

No	Name	Log J
1	(2-chloroethyl)benzene	-1.292
2	1,2,4-trimethylbenzene	-0.74
3	1,3,5-triethylbenzene	-1.083
4	1,3-diethylbenzene	-0.774
5	1,5-dimethyl-2-pyrrole carbonitrile	-1.791
6	1,6-dihydroxynaphthalene	-1.883
7	1-bromonaphthalene	-1.726
8	1-ethoxynaphthalene	-2.79
9	1-fluoro-4-nitrobenzene	-1.6
10	1-methyl-2-phenoxyethylamine	-1.63
11	1-methylimidazole	-1.813
12	1-methylnaphthalene	-1.592
13	1-methylpyrrole	-0.657
14	1-naphthoic acid	-2.985
15	1-nitronaphthalene	-2.447
16	1-phenyl-2-propanol	-2.015
17	2-(3-hydroxyphenoxy)ethanol	-3.54
18	2,4-dihydroxypyridine	-4.289
19	2,4-dimethyl-6-hydroxypyrimidine	-3.3
20	2,4-quinolinediol	-5.469
21	2,5-dimethylfuran	-0.28
22	2,5-dimethylpyrrole	-1.4
23	2,5-dimethylthiophene	-0.468
24	2,5-pyridinedicarboxylic acid	-5.205
25	2,6-dimethoxypyridine	-1.129
26	2-amino-4,6-dimethylpyridine	-2.253
27	2-amino-4-methyl pyridine	-2.228
28	2-amino-5-chloropyridine	-2.625
29	2-amino-5-nitropyridine	-3.77
30	2-aminoacetophenone	-2.16
31	2-aminobenzylalcohol	-2.63
32	2-aminopyridine	-2.682
33	2-anisaldehyde	-2.03
34	2-anisidine	-2.023
35	2-butoxypyridine	-1.155
36	2-chloro-4-fluoroacetophenone	-1.937
37	2-chloroacetophenone	-1.83
38	2-chloroanisole	-1.761
39	2-chlorobenzaldehyde	-1.58

40	2-chlorolepidine	-2.3
41	2-chloronitrobenzene	-1.54
42	2-chlorophenoxyacetic acid	-2.93
43	2-chloropyridine	-1.081
44	2-ethylimidazole	-2.975
45	2-ethylpyridine	-0.718
46	2-fluoroaniline	-1.31
47	2-fluorobenzaldehyde	-1.3
48	2-fluorobenzoic acid	-2.29
49	2-fluoronitrobenzene	-1.84
50	2-fluoropropiophenone	-1.44
51	2-fluoropyridine	-0.878
52	2-fluorotoluene	-0.349
53	2-furaldehyde	-1.53
54	2-furoic acid	-2.476
55	2-hydroxy-4-methyl quinoline	-3.876
56	2-hydroxy-5-nitropyridine	-3.747
57	2-hydroxyacetophenone	-1.78
58	2-hydroxybenzimidazole	-3.922
59	2-hydroxypyridine	-2.499
60	2-hydroxyquinoline	-3.813
61	2-methoxy-5-aminopyridine	-2.23
62	2-methoxy-5-nitropyridine	-2.653
63	2-methoxynaphthalene	-1.918
64	2-methyl-5-butylpyridine	-1.113
65	2-methyl-5-nitrobenzimidazole	-3.698
66	2-methyl-5-nitroimidazole	-4.024
67	2-methylbenzimidazole	-2.979
68	2-methylindole	-1.983
69	2-naphthol	-2.477
70	2-nitrobenzoic acid	-2.86
71	2-nitrotoluene	-1.72
72	2-pyrazine carboxylic acid	-4.067
73	2-quinolinecarboxylic acid	-3.552
74	2-quinoxalinol	-4.164
75	2-thiophenecarboxaldehyde	-1.685
76	2-thiophenemethanol	-2.179
77	2-thiophenemethylamine	-1.41
78	2-xylene	-0.644
79	3,5-dichloropyridine	-1.824
80	3,5-lutidine	-0.948

81	3-acetylpyridine	-1.992
82	3-amino-1,2,4-triazole	-3.27
83	3-amino-5,6-dimethyl-1,2,4-triazine	-3.865
84	3-aminoquinoline	-2.934
85	3-anisic acid	-2.579
86	3-chloro-4-methylaniline	-1.96
87	3-chloroaniline	-2.015
88	3-chlorotoluene	-0.837
89	3-fluorobenzyl chloride	-1.12
90	3-fluoronitrobenzene	-1.62
91	3-hydroxy-4-methoxybenzoic acid	-4.37
92	3-hydroxybenzoic acid	-3.309
93	3-hydroxypyridine	-2.685
94	3-iodoanisole	-1.805
95	3-methylthiophene	-0.407
96	3-nitrobenzoic acid	-2.735
97	3-phenoxytoluene	-2.01
98	3-phenyl-1-propanol	-2.324
99	3-phenyl-1-propylamine	-1.457
100	3-phenylbutyraldehyde	-1.959
101	3-pyridinecarboxaldehyde	-1.823
102	3-quinolinecarboxylic acid	-4.41
103	3-t-butylphenol	-1.9
104	3-thiopheneacetic acid	-2.411
105	3-thiophenecarboxaldehyde	-1.612
106	4,7-dichloroquinoline	-2.59
107	4-acetoxybenzoic acid	-3.107
108	4-aminobenzoic acid	-3.488
109	4-anisaldehyde	-2.07
110	4-anisic acid	-3.226
111	4-bromotoluene	-1.421
112	4-bromoveratrole	-2.34
113	4-carboxybenzaldehyde	-3.44
114	4-chloro-3-nitroacetophenone	-3.33
115	4-chloro-4-fluorobutyrophenone	-2.21
116	4-chlorobenzoic acid	-3.088
117	4-chlorobenzyl alcohol	-2.504
118	4-chlorotoluene	-0.694
119	4-fluoro-3-methylbenzylamine	-1.42
120	4-hydroxybenzamide	-3.83
121	4-hydroxybenzoic acid	-3.53

122	4-hydroxyquinoline	-3.688
123	4-isopropylbenzaldehyde	-1.64
124	4-methoxy-2-quinolinic acid	-4.617
125	4-methoxybenzyl acetate	-2.13
126	4-methylpyrimidine	-1.022
127	4-nitrobenzoic acid	-3.358
128	4-picoline	-0.845
129	4-quinoline carboxylic acid	-4.518
130	4-t-butylbenzoic acid	-2.759
131	4-t-butylpyridine	-1.227
132	4-t-butyltoluene	-0.915
133	4-xylene	-0.457
134	5-aminoquinoline	-3.113
135	5-chloro-8-hydroxyquinoline	-3.166
136	5-methoxypyridine	-0.809
137	5-methylbenzimidazole	-3.076
138	5-nitro-8-hydroxyquinoline	-4.22
139	5-nitroquinoline	-2.862
140	6-aminoquinoline	-3.061
141	6-hydroxynicotinic acid	-5.1
142	6-isopropylquinoline	-1.897
143	6-methoxyquinoline	-2.097
144	6-methylquinoline	-1.747
145	6-nitroquinoline	-3.615
146	7-amino-2,4-dimethyl-1,8-naphthyridine	-3.663
147	7-nitroindole	-2.659
148	8-aminoquinoline	-2.278
149	8-hydroxyquinaldine	-2.375
150	8-hydroxyquinoline	-2.358
151	8-nitroquinoline	-3.395
152	8-quinoline carboxylic acid	-4.213
153	acetophenone	-1.64
154	acridine	-2.683
155	aminopyrazine	-2.587
156	aniline	-1.75
157	anisole	-1.03
158	anthracene	-3.839
159	benoic acid	-2.316
160	benzamide	-3.07
161	benzofuran	-0.948
162	benzyl alcohol	-2.222

163	benzylamine	-1.387
164	biphenyl	-2.05
165	butyl phenyl ether	-1.25
166	butylbenzene	-0.895
167	butyrophenone	-1.719
168	chlorobenzene	-0.54
169	dibenzyl	-1.98
170	diphenyl ether	-1.81
171	dl-2-phenylpropionaldehyde	-1.686
172	ethyl nicotinate	-1.53
173	ethyl paraben	-2.69
174	ethyl salicylate	-1.61
175	ethyl-2-methylbenzoate	-1.48
176	ethylbenzene	-0.555
177	fluorobenzene	-0.256
178	furfuryl alcohol	-1.86
179	imidazole	-3.019
180	isophthalic acid	-3.987
181	isoquinoline	-1.677
182	lepidine	-1.853
183	methoxymethylphenyl sulphide	-1.684
184	methyl 4-t-butylbenzoate	-1.71
185	methyl benzoate	-1.46
186	methyl paraben	-2.74
187	methylbenzylamine	-1.18
188	nicotinic acid	-3.76
189	nitrobenzene	-1.556
190	phenethylamine	-1.257
191	phenetole	-1.11
192	phenol	-1.57
193	phenoxyacetic acid	-2.458
194	phenyl acetate	-1.65
195	picolinic acid	-3.282
196	pyridazine	-1.865
197	pyridine	-0.695
198	pyrrole	-0.891
199	quinoline	-1.49
200	salicylic acid	-2.57
201	styrene	-0.711
202	t-butylbenzene	-0.753
203	terephthalic acid	-5.145

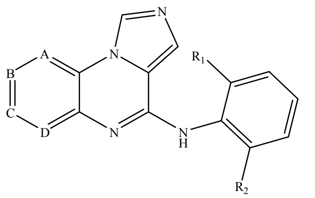
204	thioanisole	-1.39
205	toluene	-0.388
206*	1,2,5-trimethylpyrrole	-0.918
207*	1,3-disopropylbenzene	-1.06
208*	1-isoquinoline carboxylic acid	-4.132
209*	2-chloro-6-methoxypyridine	-1.211
210*	2-chlorotoluene	-0.771
211*	2-isopropylaniline	-1.69
212*	2-methoxyacetophenone	-2.02
213*	2-methyl-1-phenyl-2-propanol	-1.82
214*	2-methyl-5-ethylpyridine	-0.868
215*	2-methyl-8-nitroquinoline	-3.827
216*	2-methylimidazole	-2.797
217*	2-methylthiophene	-0.426
218*	2-naphthylacetic acid	-3.57
219*	2-thiopheneacetic acid	-2.475
220*	3,5-dimethylpyrazole	-1.791
221*	3-aminobenzoic acid	-3.727
222*	3-aminopyridine	-1.895
223*	3-anisaldehyde	-2.09
224*	3-chlorobenzoic acid	-2.371
225*	3-methoxyacetophenone	-1.99
226*	3-nitrobenzaldehyde	-2.52
227*	3-toluic acid	-2.309
228*	3-xylene	-0.58
229*	4-aminoacetophenone	-3.04
230*	4-aminophenol	-3.91
231*	4-aminoquinaldine	-3.481
232*	5-chloro-3-pyridinol	-2.621
233*	6-chloronicotinic acid	-3.098
234*	6-methoxy-8-nitroquinoline	-4.332
235*	6-methoxyquinaldine	-2.247
236*	6-quinolinecarboxylic acid	-4.672
237*	benzaldehyde	-1.48
238*	benzene	-0.256
239*	benzimidazole	-2.944
240*	benzohydroxamic acid	-3.27
241*	benzonitrile	-1.55
242*	ethyl cinnamate	-1.95
243*	furfuryl amine	-1.116
244*	indole	-1.846

245*	iodobenzene	-1.3
246*	mesitylene	-0.701
247*	methyl 2-methoxybenzoate	-2.19
248*	methyl 2-nitrobenzoate	-2.68
249*	methyl 3-methylbenzoate	-1.43
250*	methyl salicylate	-1.67
251*	naphthalene	-1.746
252*	phenylbutylamine	-1.397
253*	phenylurea	-3.31
254*	propyl paraben	-2.72
255*	pyrazole	-1.597
256*	quinaldine	-1.622

*Compounds contained in the prediction set.

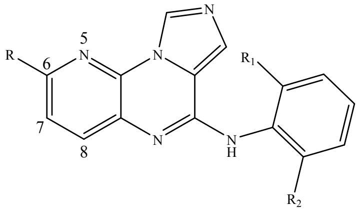
Dataset 2 (LckI) contains 105 pan-Src Lck inhibitors [28-30], which are 2-aminothiazole based chemicals. Lymphocyte-specific kinase (Lck), as one of the nine known members of Src family, expressed primarily in T-cells and natural killer cells, is required for T-cell development [32] and activation [33], which plays a critical role in signal transduction pathways. Many works have proved that dysregulation of Lck expression or its kinase activity has also been implicated in human T-cell leukemia [34-35], lymphocytic B cell leukemia [36-37], human colon carcinoma [38] and small cell lung cancer [39]. So developing selective Lck inhibitors maybe help the treatment of acute and chronic T-cell mediated autoimmune and inflammatory disorders, even leukemia and cancer etc. The molecular structures used in this dataset are more complex than dataset 1 with more flexible single bonds.

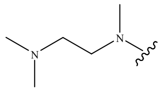
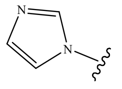
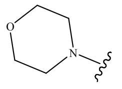
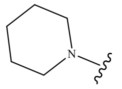
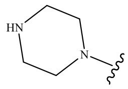
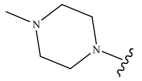
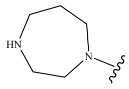
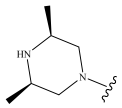
Table 3 The molecular structures and corresponding pIC50 of dataset 2 (LckI).

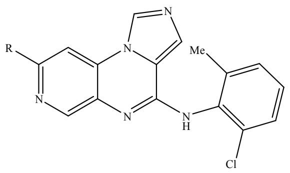
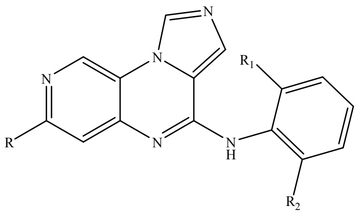
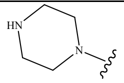
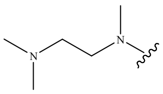
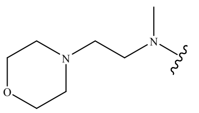
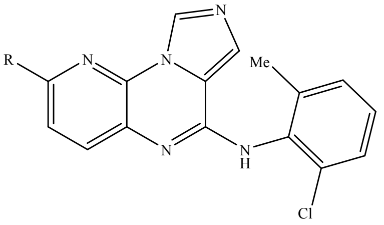
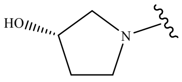
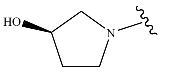
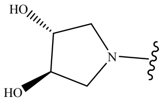
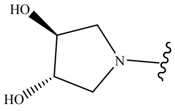
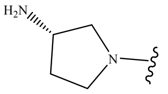


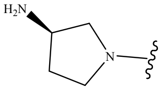
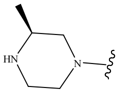
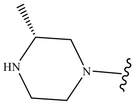
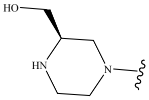
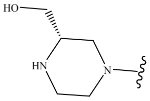
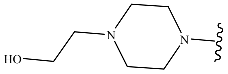
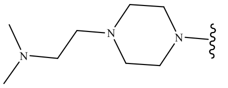
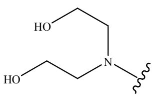
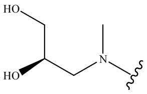
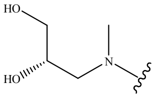
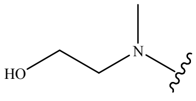
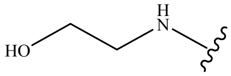
Comp.	A,B,C,D	R_1, R_2	pIC_{50}
1	A=B=C=D=CH	2-Cl,6-Me	8.05
2	A=N, B=C=D=CH	2-Cl,6-Me	7.52
3	B=N, A= C=D=CH	2-Cl,6-Me	5.85
4*	C=N, A=B= D=CH	2-Cl,6-Me	7.11
5	D=N, A=B=C= CH	2-Cl,6-Me	7.05
6*	D=N, A=B=C= CH	2,6-di-Me	6.82

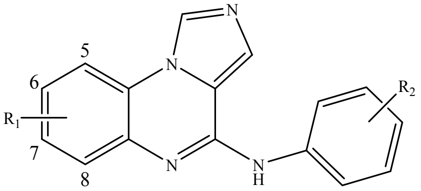
7	A=B=D=CH, C=C-OMe	2-Cl,6-Me	8.05



Comp.	R	R_1, R_2	*pIC_{50}*
8	6-Cl	2-Cl,6-Me	7.60
9	6-MeO	2-Cl,6-Me	7.72
10	6-Me_2N	2-Cl,6-Me	9.00
11	6-Et_2N	2-Cl,6-Me	8.70
12		2-Cl,6-Me	7.15
13		2-Cl,6-Me	8.30
14		2-Cl,6-Me	9.00
15		2-Cl,6-Me	8.15
16		2,6-di-Me	9.00
17		2-Cl,6-Me	9.00
18		2-Cl,6-Me	8.22
19		2-Cl,6-Me	8.70

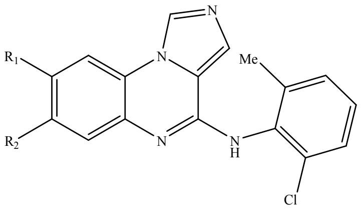
20			7.54
			
21		2-Cl,6-Me	6.79
22		2-Cl,6-Me	6.62
23		2-Cl,6-Me	6.99
			
24		---	8.40
25*		---	8.10
26		---	8.30
27		---	7.89
28		---	8.22

29		---	8.15
30		---	8.40
31*		---	8.70
32		---	8.30
33		---	8.52
34		---	8.05
35		---	8.70
36		---	8.40
37		---	8.15
38		---	7.89
39		---	8.52
40*		---	8.52

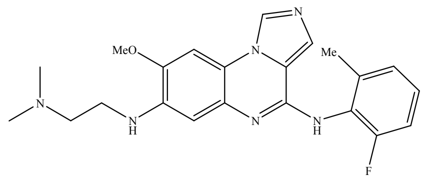
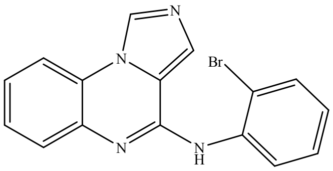


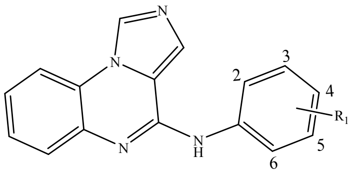
Comp.	R$_1$	R$_2$	*pIC$_{50}$*
41	6,7-di-OMe	2-Cl, 6-Me	8.70

42	6,7-di-OMe	2, 6-di-Me	8.62
43*	6,7-di-OH	2-Cl, 6-Me	8.40
44	6,7-OCH$_2$O	2-Cl, 6-Me	8.40
45	6,7-O(CH$_2$)$_2$O	2-Cl, 6-Me	8.00
46	6-OMe	2-Cl, 6-Me	8.52
47	7-OMe	2-Cl, 6-Me	8.06
48	8-OMe	2-Cl, 6-Me	6.55
49*	5-OMe	2-Cl, 6-Me	8.03
50*	5-BnO	2-Cl, 6-Me	6.62
51	5-NO$_2$	2,6-di-Me	7.00
52	5-NH$_2$	2,6-di-Me	7.15
53	6-F	2-Cl, 6-Me	7.59
54*	6-Br	2-Cl, 6-Me	7.82
55	6-CO$_2$Me	2-Cl, 6-Me	7.59
56*	6-NO$_2$	2,6-di-Cl	7.62
57	6-CN	2-Cl, 6-Me	7.00
58*	6-NH$_2$	2-Cl, 6-Me	8.15
59	6-NHAc	2-Cl, 6-Me	8.52
60	7-Br	2-Cl, 6-Me	7.85
61	7-NH$_2$	2-Cl, 6-Me	7.68
62*	7-NHAc	2-Cl, 6-Me	7.96
63	7-CONH$_2$	2-Cl, 6-Me	7.52

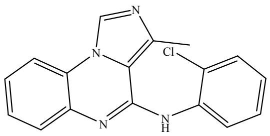
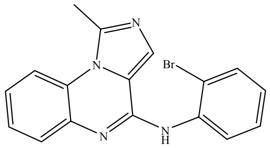


64*	NMe$_2$	H	8.30
65	NEt$_2$	H	8.70
66	NHEt	H	8.00
67	NHCH$_2$CH$_2$NMe$_2$	H	8.22
68	NHCH$_2$CH$_2$–morpholine	H	8.15
69	NHCH$_2$CH$_2$CH$_2$–morpholine	H	8.52
70	Morpholine	H	8.40
71	Piperazine	H	8.52
72*	N-Me-piperazine	H	9.00
73	N-Et piperazine	H	8.70
74	N-Formyl piperazine	H	8.05
75*	3,5-di-Me-piperazine	H	8.52
76*	N-Me-homopiperazine	H	8.05

77	H	NEt_2	8.05
78	H	$NHCH_2CH_2NMe_2$	8.00
79	H	$NHCH_2CH_2CH_2NMe_2$	8.05
80*	H	Morpholine	8.22
81	H	Piperazine	7.96
82	H	N-Me-piperazine	8.10
83	H	3,5-di-Me-piperazine	7.89
84	OCH_2CH_2–morpholine	OMe	7.44
85	3,5-di-Me-piperazine	OMe	7.74
86	OMe	OCH_2CH_2–morpholine	8.62
87	OMe	$NHCH_2CH_2$–morpholine	8.27
88*	OMe	$NHCH_2CH_2NMe_2$	8.77
89			8.30
90			6.77



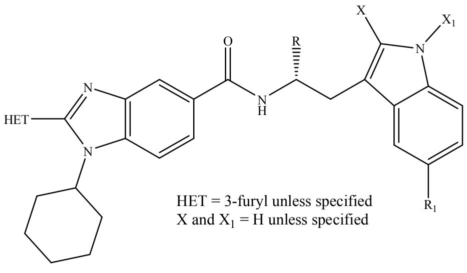
91*	H	---	6.05
92	2-F	---	6.41
93	3-F	---	5.88
94	2-Cl	---	7.22
95	2-OMe	---	5.26
96	2-Cl, 4-Me	---	6.62
97	2-Cl, 4,6-di-Me	---	7.52
98*	2,4,6-tri-Me	---	7.40
99	2,6-di-Me	---	7.80
100	2,6-di-Br	---	7.30
101	2,6-di-Cl	---	8.05
102	2,6-di-F	---	6.44

103	2,6-di-Et	---	5.77
104*			7.52
105			6.74

*Compounds contained in the prediction set.

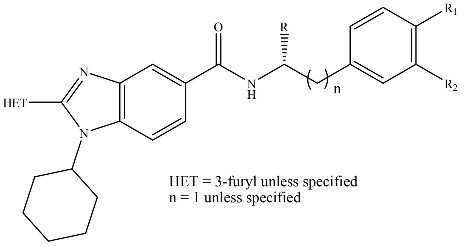
Dataset 3 contains 67 HCV NS5B polymerase inhibitors collected by Patel etc [31]. Nonstructural protein 5B (NS5B), a 66 kDa RNA-dependent RNA polymerase (RdRp), plays a pivotal role in HCV replication and the host lacks a functional counterpart of NS5B [40]. Recently NS5B has attracted the attention of medicinal chemists as a target for drug development. The urgent need for novel HCV antiviral agents has provided an impetus for understanding the structural requisites of NS5B polymerase inhibitors at the molecular level. The structures of these inhibitors containing a series of benzimidazole, tetracyclic indole, quinoxaline and indole N-acetamide derivatives, were also very complex.

Table 4 The molecular structures and corresponding pIC50 of dataset 3 (NS5BI).

Comp.	R	R_1	R2	*pIC_{50}*
				
1	COOH	OCH_2COOH	---	7.57
2	$COOCH_3$	OH	---	6.42
3	$CONH_2$	OH	---	6.66
4*	Thiazol-4-yl	OH	---	6.52
5	Thiazol-2-amino-4-yl	OH	---	6.72
6	N-methylthiazol-2-amino-4-yl	OH	---	6.30
7*	N,N-dimethyl thiazol-2-amino-4-yl	OH	---	6.05
8	thiazol-2-amino-4-yl(X1 = CH3)	OH	---	6.05
9*	COOH (HET= 2-pyridyl)	H	---	5.72
10	COOH (HET =2-pyridyl)	OH	---	6.85
11	COOH	H	---	6.40
12	COOH (X1=CH3)	OH	---	6.80

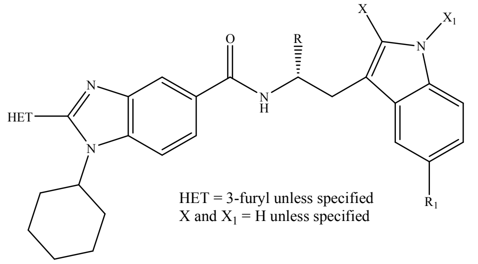
HET = 3-furyl unless specified
X and X_1 = H unless specified

13*	COOH (X=CH3)	OH	---	6.34
14	H	OH	---	5.82
15	COOH	NO_2	---	5.70
16	COOH	CH_3	---	5.66
17	COOH	F	---	5.54
18	COOH	OCH_3	---	6.40
19	COOH	NH_2	---	7.05
20	COOH	$NHSO_2CH_3$	---	6.92
21*	COOH	$NHSO_2CF_3$	---	6.82
22	COOH	NHCOCOOH	---	8.10
23*	COOH	COOH	---	7.70
24	COOH	C5-tetrazole	---	7.82
25	COOH	OC(CH3)2COOH	---	7.00



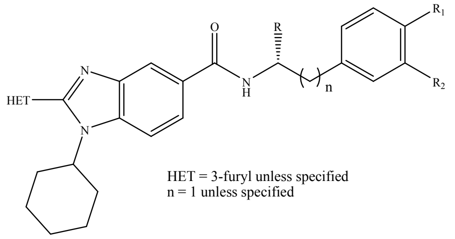
26	COOH	OH	H	6.30
27*	$CH(CH_3)_2$	OH	H	5.37
28	$CON(CH_3)_2$	OH	H	5.31
29	morpholine-4-carbonyl	OH	H	5.02
30	4-methylpiperazine carbonyl	OH	H	4.96
31	N-[2-(dimethylamino)ethyl]carboxamide	OH	H	5.68
32*	N-[(4-morpholinyl)ethyl]carboxamide	OH	H	5.49
33	N-[2-(dimethylamino)propyl]carboxamide	OH	H	5.72
34*	N-(3-pyridinylmethyl)carboxamide	OH	H	5.59
35	N-(4-pyridinylmethyl)carboxamide	OH	H	5.43
36	N-(3-pyridinyl)carboxamide	OH	H	5.52
37	2-methylthiazol-4-yl	OH	H	5.48
38	thiazol-2-amino-4-yl	OH	H	5.89
39	N,N-dimethylthiazol-2-amino-4-yl	OH	H	5.29
40	N-acetyl thiazol-2-amino-4-yl	OH	H	5.32
41	H (HET=2-pyridyl) (*n*=0)	OCH_3	OCH_3	5.15
42	COOH (HET=2-pyridyl) (*n* =0)	OCH_3	OCH_3	5.70
43*	COOH (HET=2-pyridyl)	OH	H	5.82

44	COOH (*n*=0)	OCH_3	OCH_3	6.40
45	COOH	H	H	5.40



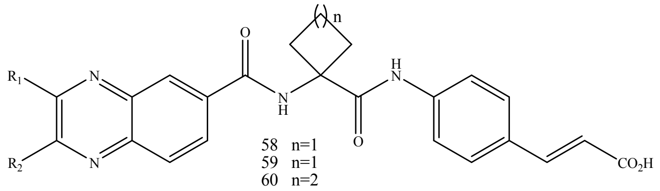
HET = 3-furyl unless specified
X and X_1 = H unless specified

46	COOH	OH	---	7.30
47	2-methylthiazol-4-yl	OH	---	6.52
48	N-acetyl thiazol-2-amino-4-yl	OH	---	6.26
49*	COOH	NHAc	---	7.22
50	COOH	$CONH_2$	---	7.43



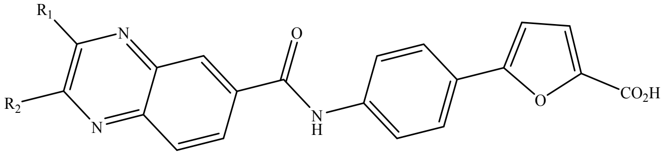
HET = 3-furyl unless specified
n = 1 unless specified

51	$CONH_2$	OH	H	6.10
52	N-(2-pyridinylmethyl) carboxamide	OH	H	5.57
53	thiazol-4-yl	OH	H	5.18
54	N-methylthiazol-2-amino-4-yl	OH	H	5.55
55	2-N-(acetamido)-1H-imidazole-4-yl)	OH	H	5.38
56	CH3 (HET =2-pyridyl) (*n*= 0)	OCH_3	OCH_3	4.92
57*	COOH (HET=2-pyridyl)	H	H	5.15

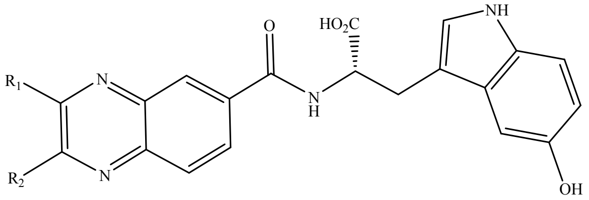
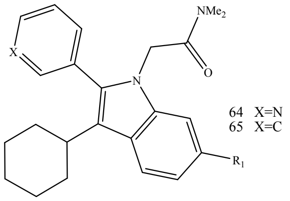
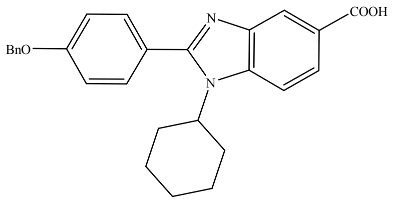
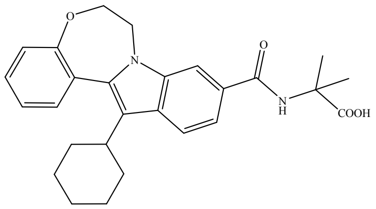


58 n=1
59 n=1
60 n=2

58	---	4-F-Ph-	4-F-Ph-	6.16
59	---	cyclohexyl-	4-F-Ph-	5.89
60	---	4-F-Ph-	4-F-Ph-	5.92



61	---	PH-	cyclohexyl-	5.74

62	---	cyclohexyl-	PH-	5.80
		 O HO$_2$C NH R$_1$ N N H R$_2$ N OH		
63	---	cyclohexyl-	PH-	6.22
				
64	---	1,2,4-oxadiazol-3-yl-5(4H)-one	---	6.11
65*	---	N-(benzylsulfonyl)-N-methylcarboxamide	---	6.19
66				5.85
67				7.34

*test set samples

2.2 Lowest-energy conformation generation

The 2D structures of the studied compounds are sketched in SYBYL 6.9 program [41]. Then we make two copies of all structures, and one copy is optimized as follows:

1) Conformation multisearch process for each molecule is executed in SYBYL. Conformation searching is used here because it can facilitate getting the global optimum conformation in the following optimization process and highly reduce the possibility to plunge into a local minimum conformation. The Multisearch method in SYBYL can locate the various energy minima available to a set of molecules. The energy minima are identified by randomly adjusting bonds and minimizing the energy of the resulting geometry. After minimization the conformation is checked against those already found and saved, if it is unique. This process is carried out for each molecule and a separate database is created. From each separate database, the conformation with lowest energy is filtered out for the following optimization process;

2) The selected conformations from step 1 are then pre-optimized with molecular mechanics MM+ force field [42], MMFF94 force field [43-45] with MMFF94 charges and Tripos force field [46] with Gasteiger-Hückel charge respectively to obtain three kinds of molecular conformations. MM+, an all atom force field, is the most general method for

molecular mechanics calculations, developed principally for organic molecules as an extension of MM2 [47]. HyperChem [48] assigns atom types and parameters not normally available to MM2 users, extending the range of chemical compounds that this force field can accommodate. MM+ also provides cutoffs for nonbonded interactions, solvation, constraints, and molecular dynamics not normally associated with MM2 calculations. Merck Molecular Force Field MMFF94, as a physically superior force field, is developed through ab initio techniques of quantum mechanics at its core and verified by experimental data sets. The effort in developing MMFF94 is facilitated by its intended use in pharmaceutical applications and the calling for its derivation and validation through computational approaches. The Tripos force field was developed for handling a broad range of organic and bioorganic molecules while not being particularly concerned about being able to reproduce some of the subtleties of molecular structure. It should be recalled that in the original design of the SYBYL program conformational energies were not regarded as important as the ability to generate all conformers of a molecule and manipulate steric volumes [49]. MM+ force field is executed in HyperChem program, and MMFF94 and Tripos force fields are performed in SYBYL;

3) The pre-optimized conformations in step 2 are separately submitted to a more precise semi-empirical method, PM3 [50], to obtain three kinds of final lowest-energy conformations. PM3 Semi-Empirical Molecular Orbital Theory is chosen because it is a robust and accurate theory, which always parallels experiment and is consequently predictive [51].

Another copy of the structures is optimized directly from the second step without conformation searching process executed in step 1. In this way another three kinds of lowest-energy conformations, maybe just local lowest-energy, are obtained. The flowchart of the optimization process is shown in Figure 1.

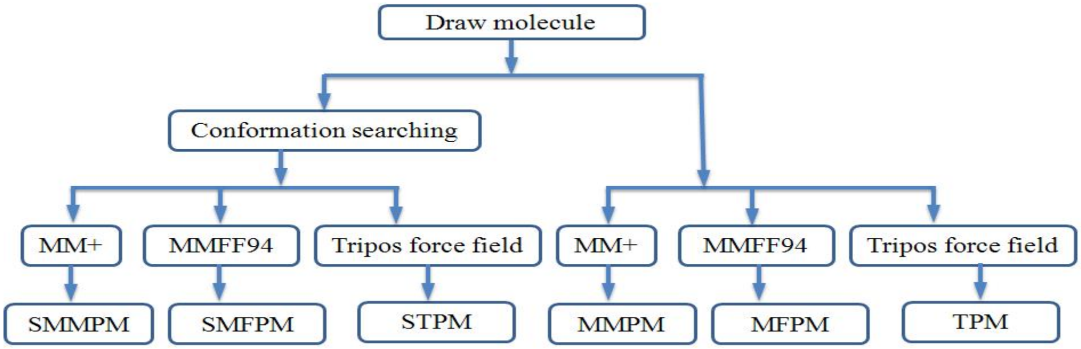


Fig. 1 The flowchart of the optimization process.

After the energy minimization, there are six different lowest-energy conformations to a certain molecule. The first three are multisearch-based conformations: multisearch-MM+-PM3 (SMMPM), multisearch-MMFF94-PM3 (SMFPM) and multisearch-Tripos-PM3 (STPM). Another three kinds of conformations are obtained without conformation searching: MM+-PM3 (MMPM), MMFF94-PM3 (MFPM), Tripos-PM3 (TPM).

2.3 Descriptor Generation

The six kinds of molecular conformations were submitted to DRAGON 5.4 program [23] respectively, in which 1354 descriptors were calculated including zero-, one-, two-, three-dimensional, charge descriptors and molecular properties. The list and meaning of the molecular descriptors is provided by the DRAGON package, and the calculation procedure is explained in detail, with related literature references, in the Handbook of Molecular Descriptors [4]. As a pre-reduction step, constant or near-constant variables are deleted, and if the pairwise correlation of two descriptors is very high (correlation coefficient greater than 0.95 here), the one showing the highest pair correlation with all the other descriptors is automatically excluded.

2.4 Variables selection and model construction by GA-MLR

After descriptor calculation, genetic algorithm (GA), which has been proved to be a very effective tool in the feature selection [52], was employed to select descriptors highly correlated with the dependent variables. The first step of GA is to randomly generate a set of solutions, which is called the initial population. Each solution, a QSAR model based on the contained descriptors by using multiple linear regressions method, is called a chromosome. Subsequently the fitness function, Friedman lack-of-fit (LOF) function, defined as follows, is used to evaluate these solutions:

$$LOF = \{SSE/(1-(c+dp/n))\}^2$$

where SSE is the sum of squares of errors, c is the number of basis function (other than the constant term), d is the smoothness factor (default 0.5), p is the number of features in the model, and n is the number of samples for model construction. After that, a new population is formed consisting of the fittest chromosomes as well as offspring of these chromosomes based on the notion of survival of the fittest. Then crossover and mutation operations are performed to generate new individuals. In the subsequent selection stage, the fittest individuals evolve to the next generation. These steps of evolution continue until the stopping criteria are satisfied. The important parameters that controlled the GA performance are listed as follows: population size (300), maximum generations (5000), mutation probability (0.1). When adding new descriptors cannot take significant improvement to the model, the optimum number of variable (Vn) used to build model is obtained.

2.5 Model validation

It is reported that the best way to evaluate the predictivity of a QSAR model, in the absence of new data, is its validation on a prediction set of compounds not included in the training process, ignoring their known activities [53-54]. So here, besides several commonly used statistic terms such as correlation coefficient (R2), leave-one-out (LOO) cross-validated, root mean squared error (RMSE) etc., all the built QSAR models are validated by an external prediction set.

In this work, the activity ranking method, which split the dataset only according to the dependent variables (property/bioactivity) without any relationship with the molecular structures, is used to divide the studied dataset into a training set and a prediction set. The first step of this method is to sort the compounds by the ascending activity order. Then from the fifth compound every fifth compound is selected into the prediction set, and the remaining compounds are included into the training set. The prediction set samples are marked by "*" in Table 2-4. After the endpoints of the validation set samples are predicted, the agreement between the experimental and predicted values is calculated as a measure of the predictive ability of a QSAR model, by using the formula recommended by the famous OECD principles [55]:

$$Q^2_{pred} = 1 - \frac{\sum_{i=1}^{m} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{m} (y_i - \overline{y}_{tr})^2}$$

where y_i and $\hat{y}_i$ are the measured and calculated values of the dependent variable for the external validation set, and $\overline{y}_{tr}$ is the mean value of the dependent variable for the training set, m is the number of the external validation set.

3. Results and discussions

As stated above, there are six kinds of conformations for each dataset, accordingly six model populations were generated. In each population, GA-MLR method provided one hundred QSAR models. So there were totally 600 individual models to a certain dataset. A statistical analysis of these models was investigated to each dataset. Furthermore, we selected one best individual model from each population to compare the results. The selection criterions are that the model should have higher cross-validated Q^2_{LOO}, higher external predictive ability, least difference between internal and external predictive ability, the fewer chemicals outside the chemical domain and the fewer chemicals with large relative errors [56-57].

3.1 Results of dataset 1 (SMF)

To analysis the models pool, all the 600 models were put together. Considering the important role of the external predictivity when evaluating a QSAR model, we ranked all these developed models according to the decreasing correlation coefficient Q^2_{pred} values for the external prediction set. Then the top 100 models based on 3 descriptors were selected for further analysis. The distribution of the top models among the six different conformations were summarized in Table 5 and shown in Figure 2 and 3. From Figure 2 it can be seen that 51 models were generated from structures after conformation searching and 49 models from structures without searching. So there was no obvious difference between these two approaches. Additionally, two kinds of conformations were optimized by molecular mechanics MM+ method first then semi-empirical PM3 method (MMPM and SMMPM), two by MMFF94 force field method first then PM3 method (MFPM and SMFPM), and two by Tripos force field first then PM3 method (TPM and STPM). From Table 5 and Figure 3 it can be seen that the three optimization processes contribute almost the same number of models among the top 100 models, which indicates that the descriptor pools from different optimization processes should be similar and the conformational differences are very small.

Table 5 The distribution of the top 100 models, according to the prediction set Q^2_{pred} , Of the studied datasets.

Models based on different conformation[a]	Data set		
	SMF	LckI	NS5BI
MMPM	17	0	3
SMMPM	16	5	3
MFPM	17	0	2
SMFPM	17	20	3
TPM	17	20	5
STPM	16	55	84

[a] multisearch-based conformations: SMMPM (multisearch-MM+-PM3), SMFPM (multisearch-MMFF94-PM3), STPM (multisearch-Tripos-PM3); and conformations without searching: MMPM (MM+-PM3), MFPM (MMFF94-PM3), TPM (Tripos-PM3).

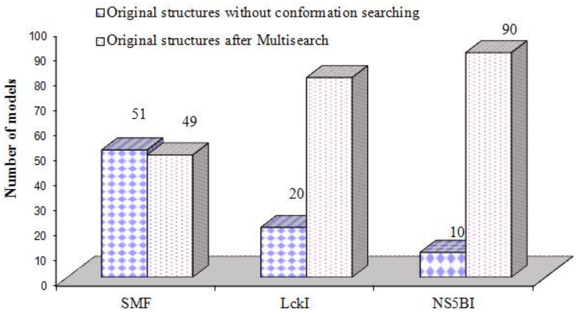


Fig. 2 The comparison of the top 100 models generated from conformations with or without conformation searching.

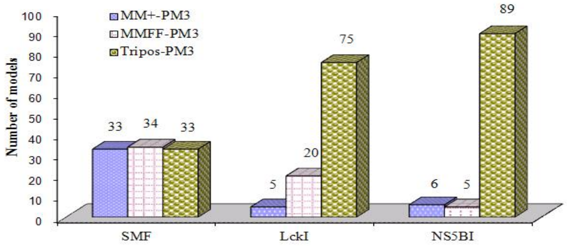


Fig. 3 The comparison of top 100 models generated from conformations by using different optimization process.

Table 6 is the statistical items of the best individual models from different conformations. The obtained best individual QSAR models for this dataset, selected from each model population, performed almost the same with only three descriptors. It is important to note that the selected descriptors for each model, listed in Table 7, are almost the same. There is one common descriptor TPSA(NO), which means the topological polar surface area using N, O polar contributions. STN is another common one except the model based on SMFPM conformation, where nAB appears instead of STN. Verifying the pair correlation, we found that these two descriptors are highly correlated with correlation coefficient R2 of 0.99 and they can be substituted by each other in the models. So we can say that there are two common descriptors present in these six best models. The third one is nARNO2, nN+ or O-061. Functional group counts descriptor nARNO2 means the number of nitro groups (aromatic). nN+ is also a functional group counts descriptor denoting the number of positive charged N (here nitro groups). O-061 belongs to atom-centred fragments descriptor, which counts the number of O atom, as in nitro and N-oxides (in this case in nitro). So we can say that the selected descriptors of each model have similar meanings and there is no big difference among the six models. Furthermore these models have no outlier for the prediction set compounds.

Table 6 The statistical items of the best individual models for the studied three datasets.

		Multisearch[a]			Nonsearch[b]		
		SMMPM	SMFPM	STPM	MMPM	MFPM	TPM
SMF	R^2	0.791	0.790	0.791	0.791	0.791	0.791
	Q^2_{LOO}	0.781	0.780	0.781	0.781	0.781	0.781
	Q^2_{EXT}	0.825	0.825	0.825	0.825	0.825	0.825
	RMSEtr	0.503	0.504	0.503	0.503	0.503	0.503
	$RMSE_{EXT}$	0.459	0.456	0.459	0.444	0.444	0.459
	Number of outliers[c]	0	0	0	0	0	0
	Number of descriptors	3	3	3	3	3	3
LckI	R^2	0.772	0.768	0.778	0.807	0.796	0.773
	Q^2_{LOO}	0.751	0.714	0.716	0.786	0.731	0.713
	Q^2_{EXT}	0.683	0.739	0.745	0.633	0.626	0.741
	RMSEtr	0.382	0.384	0.377	0.352	0.361	0.381
	$RMSE_{EXT}$	0.416	0.378	0.357	0.428	0.451	0.360
	Number of the outliers[c]	0	0	1	1	0	1
	Number of descriptors	7	7	7	8	9	7
NS5BI	R^2	0.864	0.903	0.904	0.874	0.901	0.872
	Q^2_{LOO}	0.811	0.866	0.869	0.833	0.860	0.823
	Q^2_{EXT}	0.869	0.871	0.901	0.860	0.867	0.867
	RMSEtr	0.278	0.234	0.233	0.267	0.236	0.269

$RMSE_{EXT}$	0.273	0.272	0.237	0.283	0.275	0.276
Number of the outliers[c]	1	1	1	1	1	1
Number of descriptors	6	7	7	6	7	6

[a] The meanings of different conformations are the same to the notes of Table 5.

[b] number of outliers for the external prediction set.

Table 7 The descriptors selected in six best individual models for SMF data (Dataset 1).

Models based on different conformation[a]	Descriptors		
MMPM	STN	$nArNO_2$	TPSA(NO)
SMMPM	STN	nN+	TPSA(NO)
MFPM	STN	$nArNO_2$	TPSA(NO)
SMFPM	nAB	nN+	TPSA(NO)
TPM	STN	nN+	TPSA(NO)
STPM	STN	O-061	TPSA(NO)

[a] The meanings of different conformations are the same to the notes of Table 5.

Totally six descriptors are included in these six models, but all of them belong to 2D descriptors. From Table 2 we can see that all the compounds in this dataset are very simple, so the relationship between the structure and property is not complicated and no 3D descriptor appears in the final model. The obtained results indicate that it is very easy to optimize compounds with limited complexity to the global lowest-energy conformations, and various optimization methods can obtain almost the same results as 3D descriptors have no obvious influence in the modeling.

3.2 Results of dataset 2 (LckI)

Dataset 2 is 105 2-aminothiazole based pan-Src Lck inhibitors, and their molecular structures were more complex than those in dataset 1. In the top 100 models 80% of them were generated from conformations after Multisearch, as shown in Figure 2. This fact proves that for molecules with more flexibility conformation searching process is really helpful to obtain the global lowest-energy conformations. From Table 5 and Figure 3, it was obvious that the Tripos-PM3 optimization process was the best one, which contributed 75% of the top 100 models. Then MMFF-PM3 process supported 20% models, and MM+-PM3 provided only 5% models. This fact indicated that Tripos force field might be the best method to optimize the compound to its lowest-energy conformation after Multisearch among the three methods, at least to the current dataset.

Comparing the corresponding statistical items from Table 6, we can see that the performance of these QSAR models is different and the differences among the six best individual models are comparatively large. The correlation coefficients R2 for the training set varying from 0.768 to 0.803 and the LOO cross-validated Q^2_{LOO} from 0.713 to 0.754. Especially the difference among the Q^2_{pred} for the prediction set was as large as 0.119, varying from 0.626 to 0.745, which indicated the different model predictive abilities. Overall most of the models were acceptable except two models generated from MMPM and MFPM conformations, in which the difference between the fitting on the training set and the predictions on external prediction set are larger than others. These two models were built using 8 and 9 descriptors respectively. Therefore maybe they were overfitted on the training data. Among other four individual models based on seven descriptors, the two models from Tripos-PM3 conformations had comparable performance, but the conformation after multisearch (STPM) performed slightly better than TPM with higher statistical items.

The descriptors included in these best models were listed in Table 8. From this table we can see that even the number of selected descriptors in the best models was not always the same, and there was no common descriptor among these six models. Among the selected descriptors, the Geary autocorrelation GATSkw descriptor (w is the atomic property used to weight the molecular graph and k is the lag) is the most important one, which appears in five of the six best models (except SMMPM model). Though there is no GATSkw in SMMPM model, there is a similar Moran autocorrelation descriptor named MATS4e. Both GATSkw and MATS4e are 2D autocorrelation descriptors, which describe how a considered property is distributed along a topological molecular structure. The suffix v, p and e indicate carbon-scaled atomic van der Waals volume, atomic polarizability and atomic Sanderson electronegativity respectively. The number of C atom connected with electronegative atom is another important descriptor. Descriptors C-026, C-027 and C-034, which indicate the number of R--CX—R, R--CH—X and R--CR..X respectively, appear separately in five of the six best models. In these groups, R represents any group linked through carbon and X represents any electronegative atom (O, N, S, P and halogens). From the above discussion we can say that some 2D molecular properties like van der Waals volume, polarizability and electronegativity are very important features between the binding of inhibitors and pan-Lck protein.

Table 8 The descriptors selected in six best individual models for LckI data (Dataset 2).

Model[a]	Descriptors								
MMPM	GATS7v	GATS6p	HATS7u	HATS2e	R7v$^{\#}$*	R8v$^{\#}$*	R7p+$^{\#}$*	C-027	
SMMPM	nCL	MATS4e	BELv8	H8u$^{\#}$*	HATS5u$^{\#}$*	R6v+$^{\#}$*	C-027		
MFPM	GATS1v	GATS4p	BELv8	Dm*	H1u$^{\#}$*	H8u$^{\#}$*	RCON$^{\#}$*	R5u+$^{\#}$*	H-049
SMFPM	GATS1v	GATS8v	GATS4p	H8u$^{\#}$*	R4u+$^{\#}$*	nCb-	C-034		
TPM	MATS6v	GATS7v	BELv4	E3u*	HGM$^{\#}$*	H8u$^{\#}$*	C-026		
STPM	GATS8v	E1u*	H8u$^{\#}$*	HTm$^{\#}$*	R7v+$^{\#}$*	nCb-	C-034		

a Models besed on different conformations

*3D descriptors

#GETWAY descriptor

In Table 8, the 3D descriptors are highlighted with "*". It's evident that in each model there are at least two 3D descriptors. Among all these 3D descriptors, H8u (H autocorrelation of lag 8 / unweighted) is the most important one, which appears in five of the six models. Analyzing all these 3D descriptors we find that most of them belong to GETAWAY descriptor. GETAWAY descriptors encode both the geometrical information given by the influence molecular matrix and the topological information given by the molecular graph with chemical information by using different atomic weightings (atomic mass, polarizability, van der Waals volume, and electronegativity, together with unit weights). This kind of descriptors are calculated based on spatial autocorrelation, encoding information on structural fragments and therefore seems to be particularly suitable for describing differences in congeneric series of molecules. Moreover, they are independent of molecule alignment and, to some extent, account also for information on molecular size and shape as well as for specific atomic properties. The selection of these descriptors indicates the importance of the 3D distribution of atomic polarizability (p) and van der Waals atomic volume (v) together with the unit weight (u) in a molecular.

3.3 Results of dataset 3

Dataset 3 (NS5BI) was 67 HCV NS5B polymerase inhibitors, the molecular structures of which were also very complex. From the analysis of the QSAR results, we obtained conclusion similar to dataset 2. As shown in Table 6, the differences among the six best models, which contained various descriptor numbers, were also very obvious. But fortunately there was one common descriptor (Table 9), Belv6, which appeared in all the six models. Belv6 belongs to a Burden eigenvalue descriptor, a kind of 2D descriptor, which indicate the lowest eigenvalue n. 6 of Burden matrix/weighted by atomic van der Waals volumes, defined as the following: the diagonal elements are atomic properties; the off-diagonal elements corresponding to pairs of bonded atoms are the square roots of conventional bond order; all other matrix elements are set at 0.001. The calculation of this kind of descriptors just needs the 2D expression of a molecule, so the values of the Belv6 are the same to the six kinds of conformations. Still the individual model from STPM conformation was the best one, as shown in Table 6, whose statistical items were all better than others. Furthermore, it must be highlighted that the difference of LOO cross-validated Q^2 for the six best models was very large, varied from 0.811 to 0.869.

Table 9 The descriptors selected in six best individual models for NS5BI data (Dataset 3).

Model[a]	Descriptors								
MMPM	GATS7v	GATS6p	HATS7u	HATS2e	R7v$^{\#}$*	R8v$^{\#}$*	R7p+$^{\#}$*	C-027	
SMMPM	nCL	MATS4e	BELv8	H8u$^{\#}$*	HATS5u$^{\#}$*	R6v+$^{\#}$*	C-027		
MFPM	GATS1v	GATS4p	BELv8	Dm*	H1u$^{\#}$*	H8u$^{\#}$*	RCON$^{\#}$*	R5u+$^{\#}$*	H-049
SMFPM	GATS1v	GATS8v	GATS4p	H8u$^{\#}$*	R4u+$^{\#}$*	nCb-	C-034		
TPM	MATS6v	GATS7v	BELv4	E3u*	HGM$^{\#}$*	H8u$^{\#}$*	C-026		
STPM	GATS8v	E1u*	H8u$^{\#}$*	HTm$^{\#}$*	R7v+$^{\#}$*	nCb-	C-034		

a Models besed on different conformations

*3D descriptors

#GETWAY descriptor

Among the top 100 models, 90% models were generated from conformations after conformation searching (Figure 2), even a little more than dataset 2, which indicated that the molecular structures might be more complex than other two datasets and also proved the importance of conformation searching. The contained information of Table 5 and Figure 3 shows that the Tripos-PM3 optimization process contributed 89% of the top 100 models, which further indicated the advantage and ability of Tripos force field in the optimization of complex molecular structure.

From the obtained results we can conclude that any kind of optimization method, at least the used method here, is suitable for simple compounds with limited rotatable bonds. But to compounds with higher flexibility, how to obtain the accurate global low-energy conformation is still a problem. The conformational distinctions from different optimization methods may have obvious influences on the quality of the subsequent QSAR models.

4. Conclusions

In this work, two datasets were used to investigate the conformation influence on the quality of subsequent QSAR models. The obtained results indicate that the molecular conformations from different minimization processes may result in different QSAR models, mainly based on 3D descriptors, with distinct performances, especially for molecules with much flexibility. Conformation searching, aiming to find a better original conformation near to the global lowest-energy conformation, can help to find the proper conformation in the optimization process, avoiding falling into a local minimum. Based on the above conclusions we can deduce that the new chemicals should share the same optimization process as the training samples if we want to predict the corresponding property or bioactivity accurately.

Acknowledgement

This research is supported by the Fundamental Research Funds for the Central Universities (lzujbky-2012-88) and Gansu Provincial Natural Science Foundation (Grant No. 1208RJYA038).

Reference

[1] C. Hansch, A. Leo, John Wiley & Sons (New York, 1979).

[2] J. Gasteiger, Handbook of Chemoinformatics: From Data to Knowledge, John Wiley & Sons (New York, 2003).

[3] J. Gasteiger, T. Engel, Chemoinformatics: A Textbook, Wiley-VCH: Weinheim (Germany, 2006).

[4] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH: Weinheim (Germany, 2000).

[5] J. Gasteiger, Med. Chem., 49 (2006) 6429-6434.

[6] O. Mekenyan, T. Pavlov, V. Grancharov, M. Todorov, P. Schmieder, G.J. Veith, Chem. Inf. Model., 45 (2005) 283-292.

[7] http://www.cas.org/ (Accessed 30 Nov 2012).

[8]http://www.ccdc.cam.ac.uk/products/csd/ (Accessed 30 Nov 2012).

[9] R.D. Cramer 3rd, D.E. Patterson, J.D. Bunce, Am. Chem. Soc., 110 (1998) 5959-5967.

[10] G. Klebe, U. Abraham, T. Mietzner, Med. Chem., 37 (1994) 4130-4146.

[11] J.G. Topliss, R.P. Edwards, Med. Chem., 22 (1979) 1238-1244.

[12] R. Serafimova, J. Walker, O. Mekenyan, SAR QSAR Environ Res., 13 (2002) 127-134.

[13] O. Mekenyan, J. Ivanov, S. Karabunarliev, S.P. Bradbury, G.T. Ankley, W. Karcher, Environ Sci. Technol., 31 (1997) 3702-3711.

[14] O.M. Becker, Y. Levy, O. Ravitz, Phys. Chem., 104 (2000) 2123-2135.

[15] J.M.Ivanov, S.H. Karabunarliev, O.G. Mekenyan, Chem. Inf. Comput Sci., 34 (1994) 234-243.

[16] O. Mekenyan, D. Dimitrov, N. Nikolova, S.J. Karabunarliev, Chem. Inf. Comput Sci., 39 (1999) 997-1016.

[17] G.Y. Yang, J. Yu, Z.Y. Wang, X.L. Zeng, X.H. Ju, QSAR Comb. Sci., 26 (2007) 352-357.

[18] M. Staikova, P. Messih, Y.D. Lei, F. Wania, D.J. Donaldson, Chem. Eng. Data, 50 (2005) 438-443.

[19] T. Puzyn, J. Falandysz, Phys. Chem. Ref. Data, 36 (2007) 203-214.

[20] W. Zhou, Z. Zhai, Z. Wang, L. Wang, Mol. Struc – THEOCHEM, 755 (2005) 137-145.

[21] M. Staikova, F. Wania, D.J. Donaldson, Atmos Environ, 38 (2004) 213-225.

[22] T. Puzyn, N. Suzuki, M. Haranczyk, J. Rak, Chem. Inf. Model, 48 (2008) 1174-1180.

[23] Talete srl, DRAGON for Windows (Software for molecular Descripto Calculation). Version 5.4 – 2006 – http://www.talete.mi.it.

[24] M.T.D. Cronin, J.C. Dearden, R. Gupta, G.P. Moss, Pharm. Pharmacol., 50 (1998) 143-152.

[25] Y. Chen, W.L. Yang, L.E. Matheson, Int. Pharm., 94 (1993) 81-88.

[26] Y. Chen, P. Vayumhasuwan, L.E. Matheson, Int. Pharm., 137 (1996) 149-158.

[27] M. Hewitt, M.T.D. Cronin, J.C. Madden, P.H. Rowe, C. Johnson, A. Obi, S.J. Enoch, Chem. Inf. Model, 47 (2007) 1460-1468.

[28] P. Chen, A.M. Doweyko, D. Norris, H.H. Gu, S.H. Spergel, J. Das, R.V. Moquin, J. Lin, J. Wityak, E.J. Iwanowicz, K.W. McIntyre, D.J. Shuster, K .Behnia, S. Chong, F.H. De, S. Pang, S. Pitt, D.R. Shen, S. Thrall, P. Stanley, O.R. Kocy, M.R. Witmer, S.B. Kanner, G.L. Schieven, J.C. Barrish, Med. Chem., 47 (2004) 4517-4529.

[29] P. Chen, E.J. Iwanowicz, D. Norris, H.H. Gu, J. Lin, R.V. Moquin, J. Das, J. Wityak, S.H. Spergel, H. de Fex, S. Pang, S. Pitt, D.R. Shen, G.L. Schieven, J.C. Barrisha, Bioorg. Med. Chem. Lett., 12 (2002) 3153-3156.

[30] P. Chen, D. Norris, E.J. Iwanowicz, S.H. Spergel, J. Lin, H.H. Gu, Z. Shen, J. Wityak, T. Lin, S. Pang, H.F. De Fex, S. Pitt, D.R. Shen, A.M. Doweyko, D.A. Bassolino, J.Y. Roberge, M.A. Poss, B. Chen, G.L. Schievend, J.C. Barrisha, Bioorg. Med. Chem. Lett., 12 (2002) 1361-1364.

[31] P.D. Patel, M.R. Patel, N. Kaushik-Basu, T.T. Talele, Chem. Inf. Model, 48 (2008) 42-55.

[32] T.J. Molina, K. Kishihara, D.P. Siderovski, W. van Ewijk, A. Narendran, E. Timms, A. Wakeham, C.J. Paige, K.U. Hartman, A. Veilette, D. Davison, T.W. Mak, Nature, 357 (1992) 161-164.

[33] D.B. Straus, A. Weiss, Cell, 70 (1992) 585-593.

[34] C.L. Yu, R. Jove, S.J. Burakoff, Immunol., 159 (1997) 5206-5210.

[35] M.B. Majolini, M.M. D'Elios, P. Galieni, M. Boncristiano, F. Lauria, G. Del Prete, J.L. Telford, C.T. Baldari, Blood, 91 (1998) 3390-3396.

[36] S. McCracken, C.S. Kim, Y. Xu, M. Minden, N.G. Miyamoto, Oncogene, 15 (1997) 2929-2937.

[37] A. Von Knethen, H. Abts, D. Kube, V. Diehl, H. Tesch, Leuk Lymphoma, 26 (1997) 551-562.

[38] G.W. Krystal, C.S. DeBerry, D. Linnekin, J. Litz, Cancer Res., 58 (1998) 4660-4666.

[39] A. Weiss, D.R. Littman, Cell, 76 (1994) 263-274.

[40] D.C. Myles, Curr Opin Drug Discov Devel, 4 (2001) 411-416.

[41] Sybyl version 6.9, Tripos Associates, St. Louis (MO) (1999).

[42] A. Hocquet, M. Langgård, Mol Model Annual, 4 (1998) 94-112.

[43] T.A. Halgren, Comput Chem., 17 (1996) 490-519.

[44] T.A. Halgren, Comput Chem., 17 (1996) 520-552.

[45] T.A. Halgren, Comput Chem., 17 (1996) 553-586.

[46] M. Clark, R.D. Cramer III, N.V. Opdenbosch, Comput Chem., 10 (1989) 982-1012.

[47] N.L. Allinger, Am. Chem. Soc., 99 (1977) 8127-8134.

[48] HyperChem 7.0, Hypercube. Inc.. 2002. Gainesville, FL 32601, USA.

[49] G.R. Marshall, C.D. Barry, H.E. Bosshard, R.A. Dammkoehler, D.A. Dunn, in Computer-Assisted Drug Design, ACS Symposium Series, (Olson EC and Christoffersen RE (Eds)), American Chemical Society, Washington, DC, 112 (1979) 205-266.

[50] G.A. Segal, Semiempirical Methods of Electronic Structure Calculation, Part A: Techniques, University of Southern California (Los Angeles, 1997).

[51] A.J. Dickie, A.K. Kakkar, M.A. Whitehead, Langmuir, 18 (2002) 5657-5660.

[52] R. Leardi, R. Boggia, M. Terrile, Chemometr., 6 (1992) 267-281.

[53] P. Gramatica, QSAR Comb Sci., 26 (2007) 694-701.

[54] A. Tropsha, P. Gramatica, V.K. Gombar, QSAR Comb Sci., 22 (2003) 69-77.

[55] http://www.oecd.org/dataoecd/33/37/37849783.pdf (Accessed 22 Jan 2007).

[56] P. Gramatica, E. Giani, E. Papa, Mol. Graph. Model, 25 (2007) 755-766.

[57] J.Z. Li, B.L. Lei, H.X. Liu, S.Y. Li, X.J. Yao, M.C. Liu, P. Gramatica, Comput Chem., 29 (2008) 2636-2647.