Supplementary Information for ***Acta Phys. -Chim. Sin.*** **2013**, *29* (1): 217-223
doi: 10.3866/PKU.WHXB201211122

基于机器学习方法的 H1N1 神经氨酸苷酶抑制剂的分类预测

吕 巍[1,2] 薛 英[3,4] 孟庆伟[1,2,*]

([1]山东农业大学生命科学学院，作物生物学国家重点实验室，山东 泰安 271018; [2]山东农业大学生物学博士后科研流动站，山东 泰安 271018; [3]四川大学化学学院，教育部绿色化学与技术重点实验室，成都 610064; [4]四川大学生物治疗国家重点实验室，成都 610041)

Classification Prediction of Inhibitors of H1N1 Neuraminidase by Machine Learning Methods

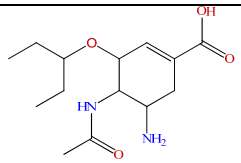
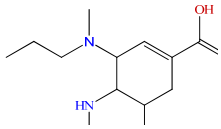
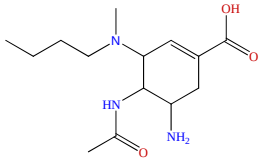
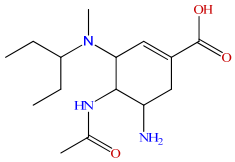
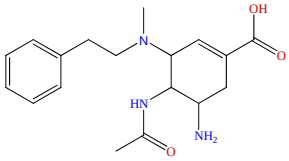
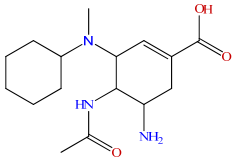
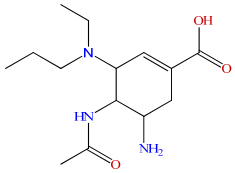
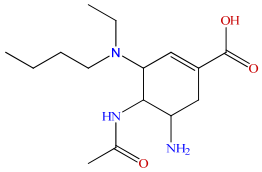
LÜ Wei[1,2] XUE Ying[3,4] MENG Qing-Wei[1,2,*]

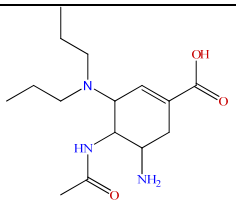
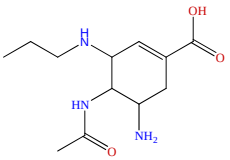
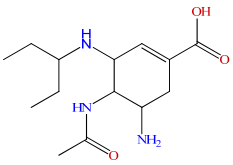
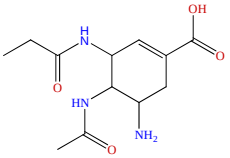
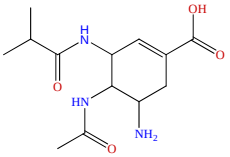
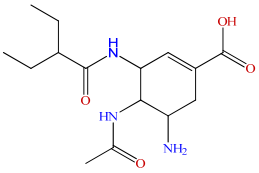
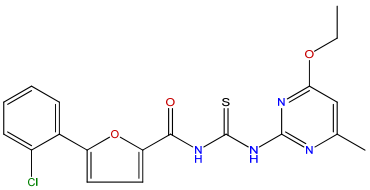
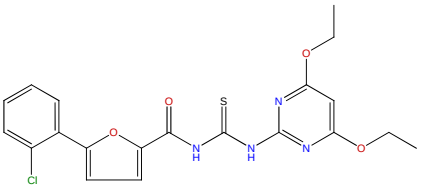
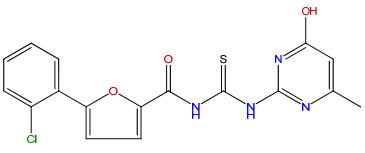
([1]*College of Life Sciences, State Key Laboratory of Crop Biology, Shandong Agricultural University, Tai'an 271018, Shandong Province, P. R. China*; [2]*Postdoctoral Research Bachelor of Biology, Shandong Agricultural University, Tai'an 271018, Shandong Province, P. R. China*; [3]*College of Chemistry, Key Laboratory of Green Chemistry and Technology, Ministry of Education, Sichuan University, Chengdu 610064, P. R. China*; [4]*State Key Laboratory of Biotherapy, Sichuan University, Chengdu 610041, P. R. China*)

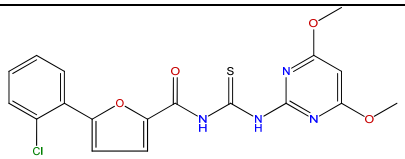
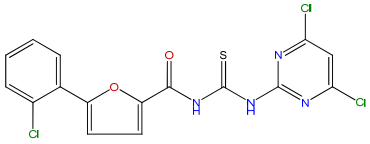
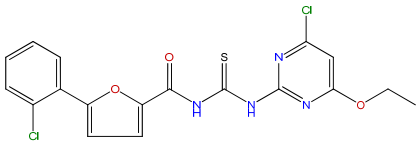
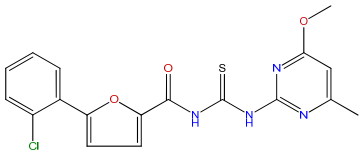
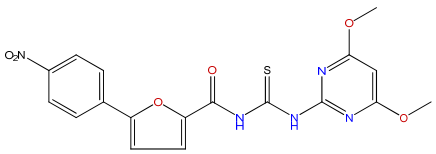
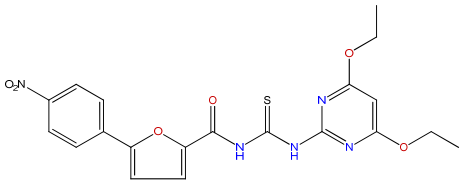
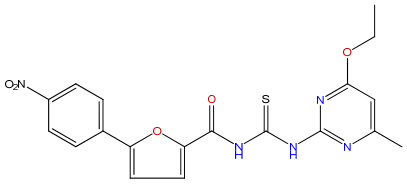
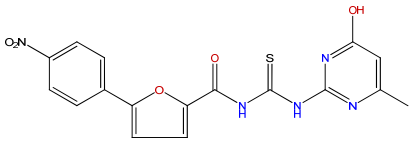
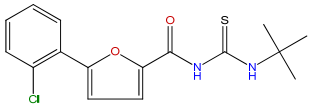
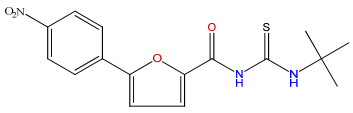
*Corresponding author. Email: qwmeng@sdau.edu.cn; Tel: +86-538-8249606.

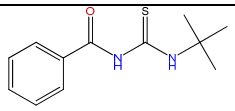
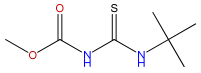
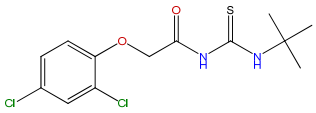
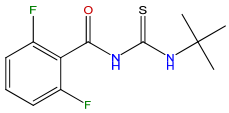
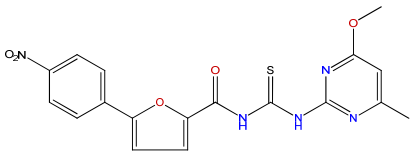
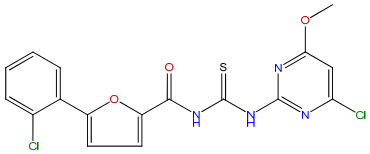
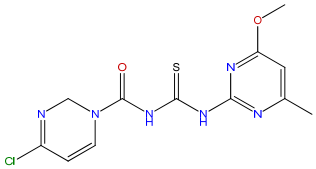
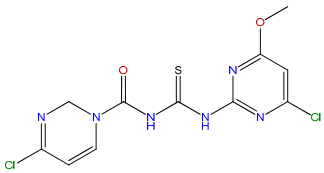
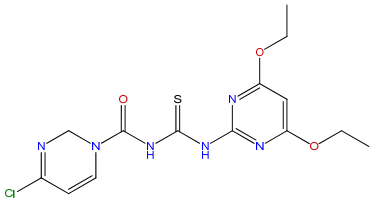
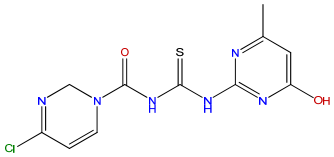
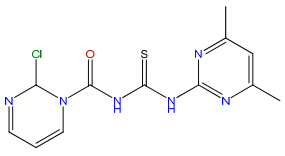
附表 1　文章中用到的所有化合物

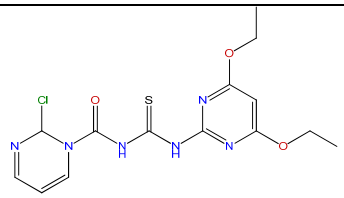
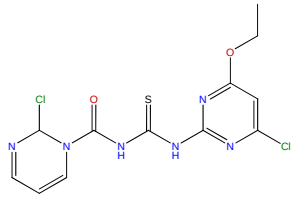
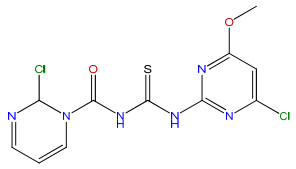
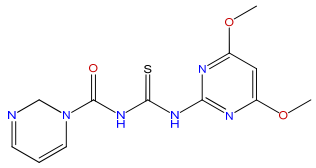
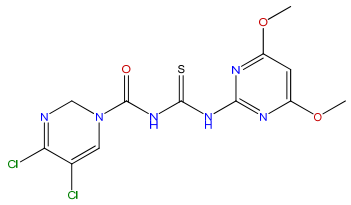
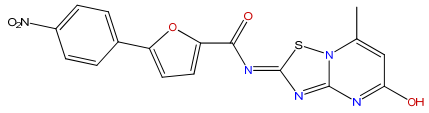
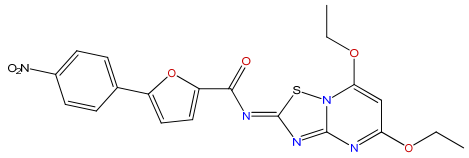
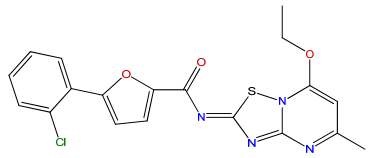
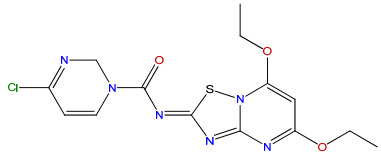
Table S1　All Molecules Used in this Article

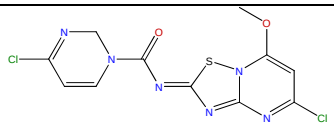
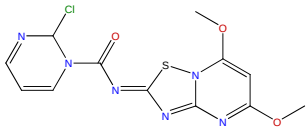
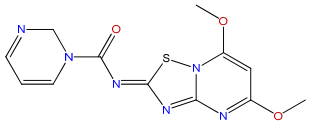
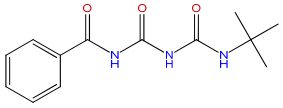
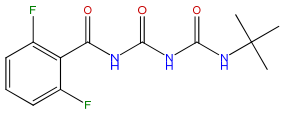
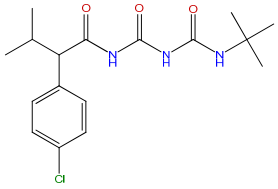
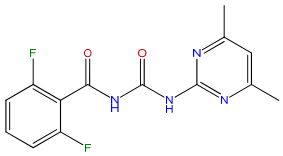
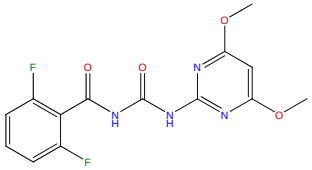
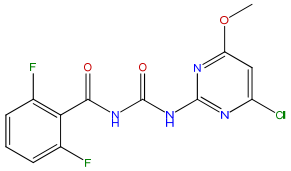
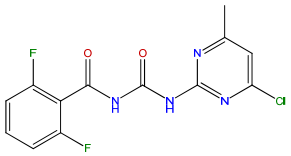
ID	Molecule Structure	ID in Ref.	IC50 (nM)	Reference
1		1	1	1
2		3a	65	1
3		3b	180	1
4		3c	6	1
5		3d	100	1
6		3e	200	1
7		3f	90	1
8		3g	85	1

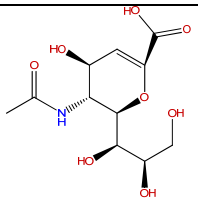
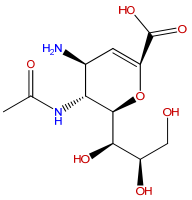
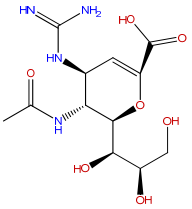
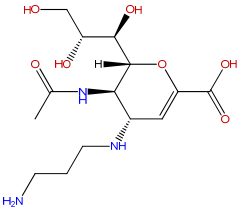
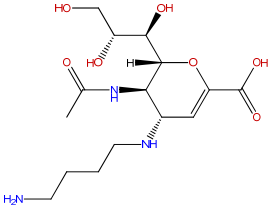
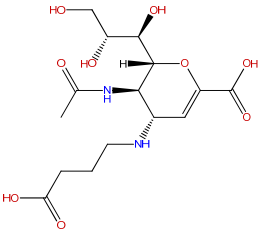
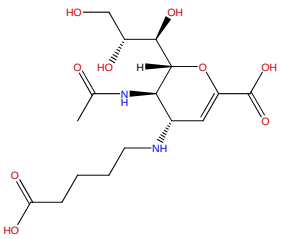
9		3h	12	1
10		3i	200	1
11		3j	11	1
12		3k	2700	1
13		3l	6400	1
14		3m	4000	1
15		15	1650	2
16		16	80	2
17		17	320	2

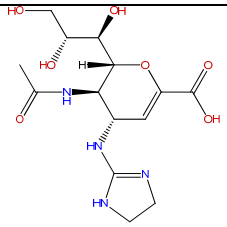
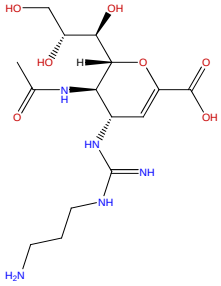
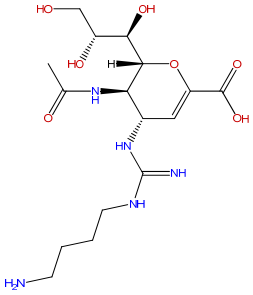
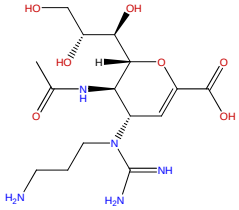
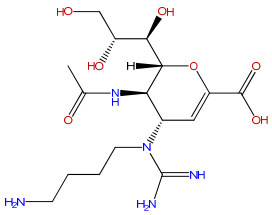
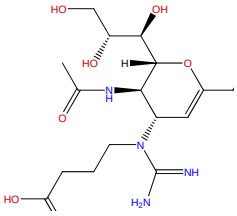
18		18	17700	2
19		19	1450	2
20		20	20000	2
21		21	20000	2
22		22	20000	2
23		23	20000	2
24		24	11300	2
25		25	360	2
26		26	1420	2
27		27	1300	2

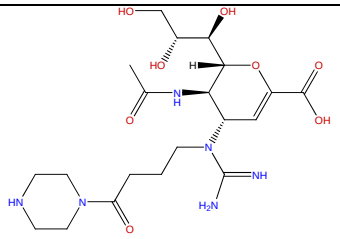
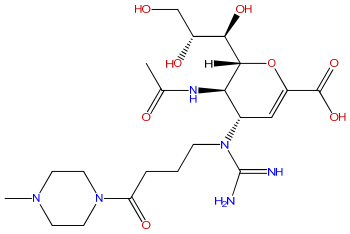
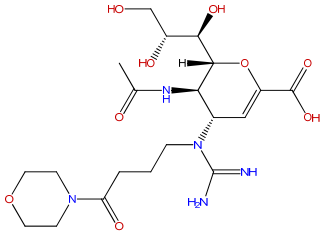
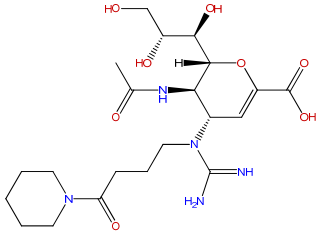
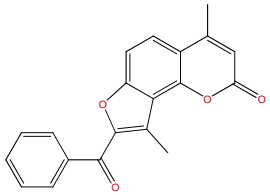
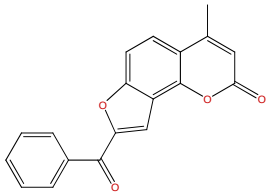
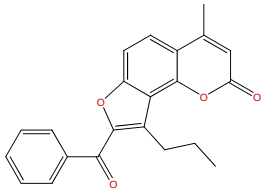
28		28	1790	2
29		29	1830	2
30		30	1670	2
31		31	1430	2
32		35	1220	2
33		36	1290	2
34		37	20000	2
35		38	20000	2
36		39	20000	2
37		40	8580	2
38		41	7190	2

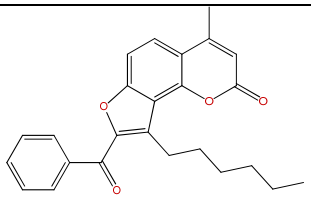
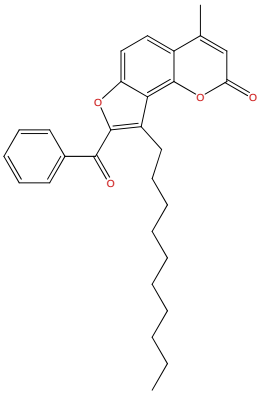
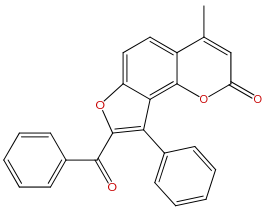
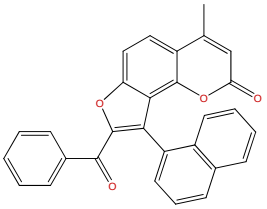
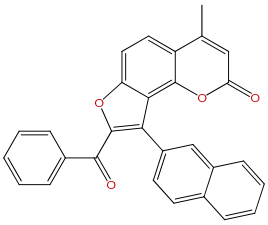
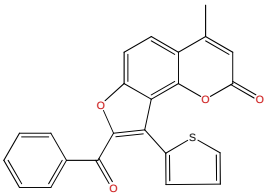
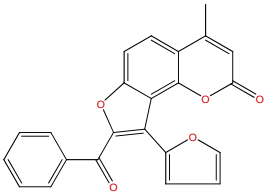
39		42	20000	2
40		43	20000	2
41		44	2590	2
42		45	20000	2
43		46	1850	2
44		47	670	2
45		48	5250	2
46		49	350	2
47		50	90	2

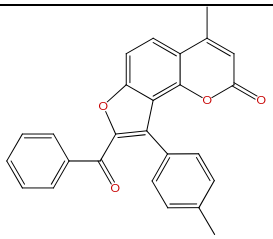
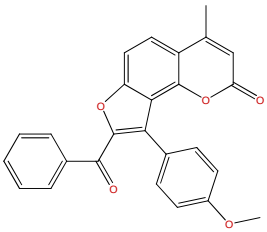
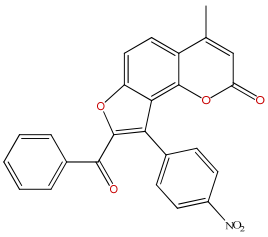
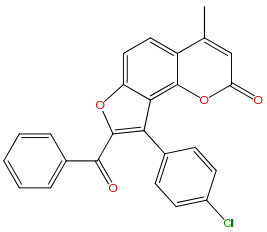
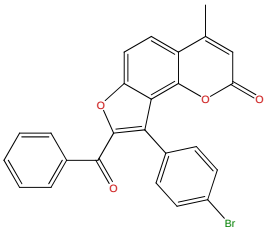
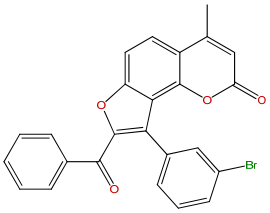
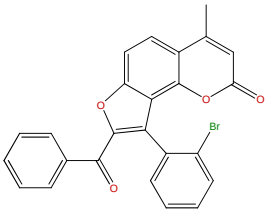
48		51	9320	2
49		52	2410	2
50		53	20000	2
51		54	20000	2
52		55	14100	2
53		56	15500	2
54		57	20000	2
55		58	20000	2
56		59	15900	2
57		60	20000	2

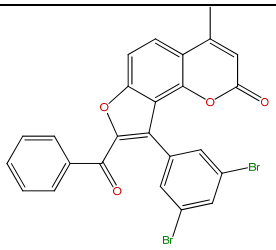
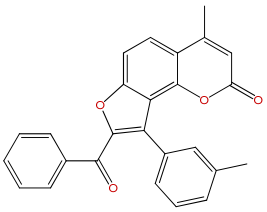
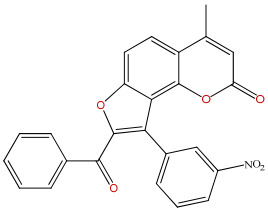
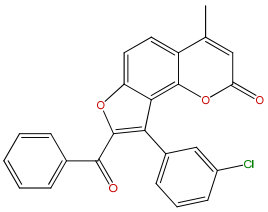
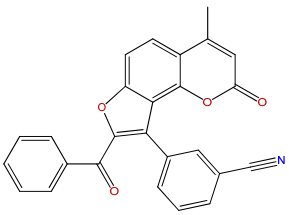
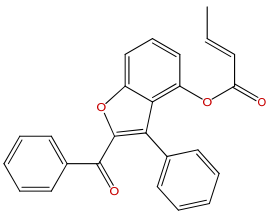
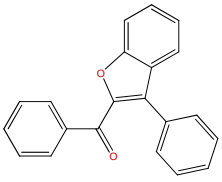
58		1	190000	3
59		2	100000	3
60		3	4	3
61		4a	200000	3
62		4b	200000	3
63		4c	70000	3
64		4d	80000	3

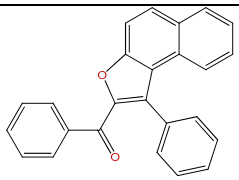
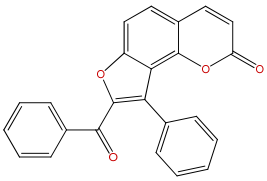
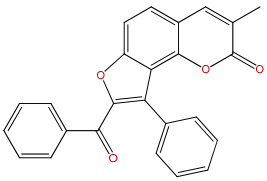
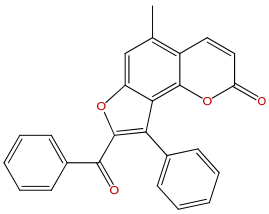
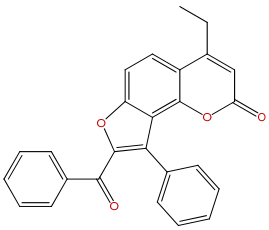
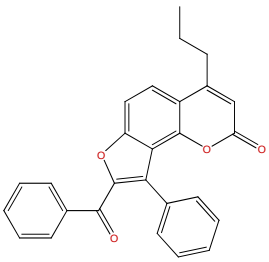
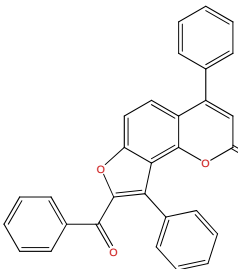
65		5	200000	3
66		6a	200000	3
67		6b	200000	3
68		7a	200000	3
69		7b	200000	3
70		7c	7120	3

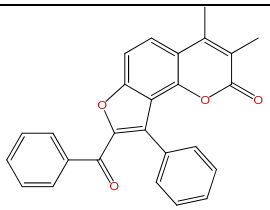
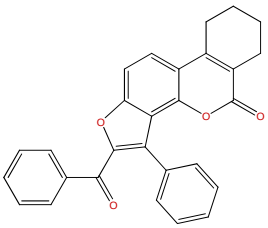
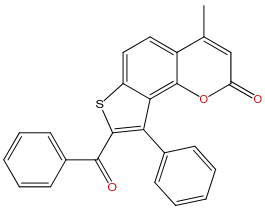
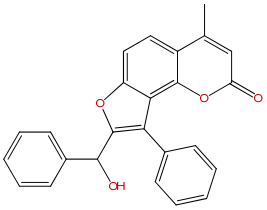
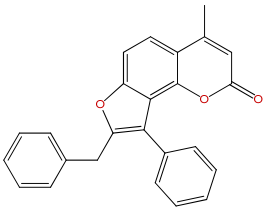
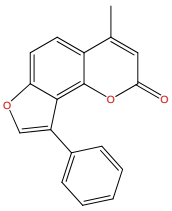
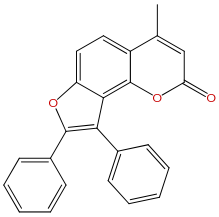
71		8a	2150	3
72		8d	6500	3
73		8c	33300	3
74		8d	23100	3
75		6a	4510	4
76		6b	2430	4
77		6c	290	4

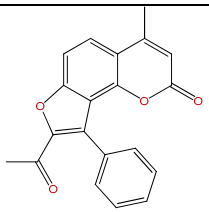
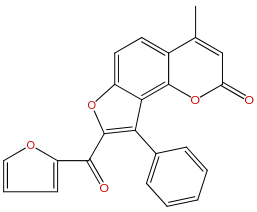
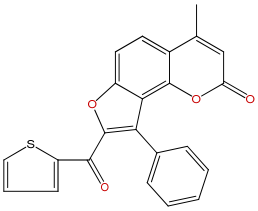
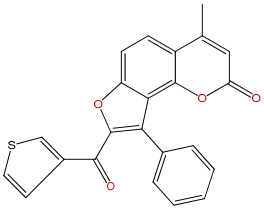
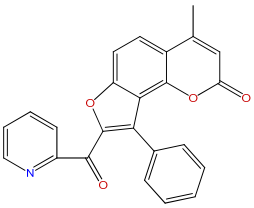
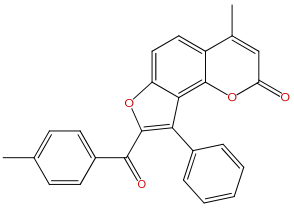
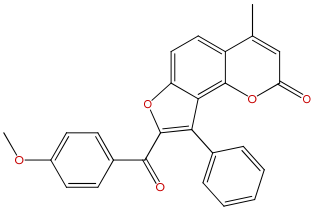
78		6d	650	4
79		6e	25000	4
80		6f	140	4
81		6g	880	4
82		6h	1910	4
83		6i	150	4
84		6j	650	4

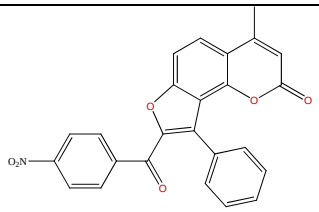
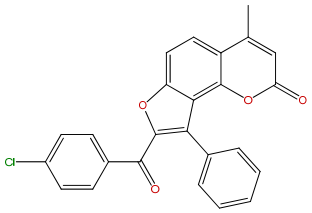
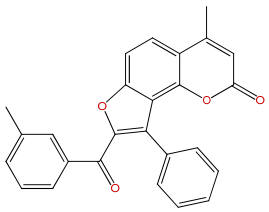
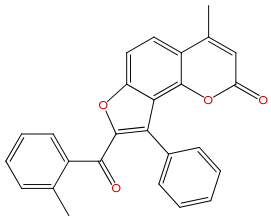
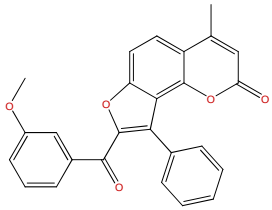
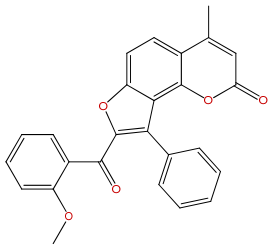
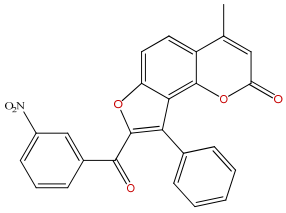
85		6k	820	4
86		6l	10080	4
87		6m	760	4
88		6n	3920	4
89		6o	460	4
90		6p	80	4
91		6q	730	4

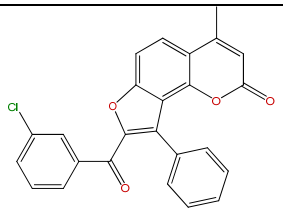
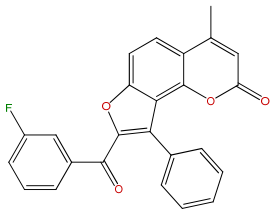
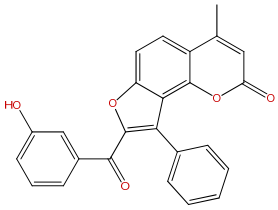
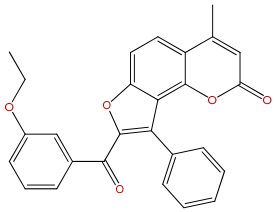
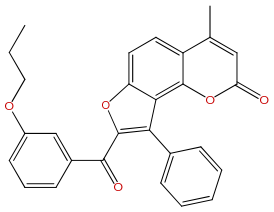
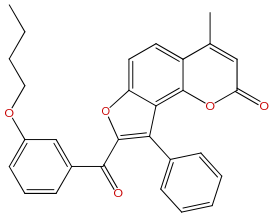
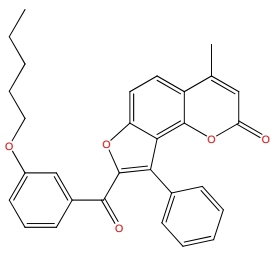
92		6r	100	4
93		6s	100	4
94		6t	250	4
95		6u	1410	4
96		6v	530	4
97		7a	20820	4
98		7b	25000	4

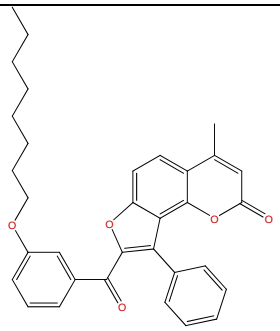
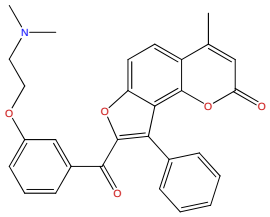
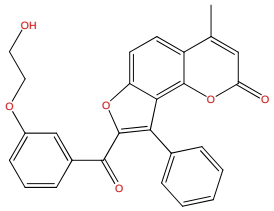
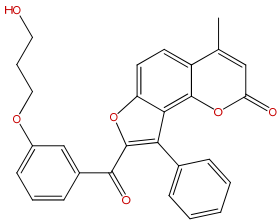
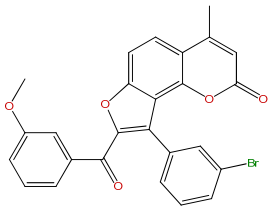
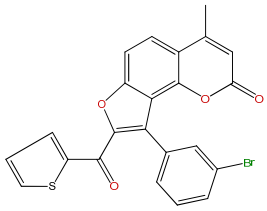
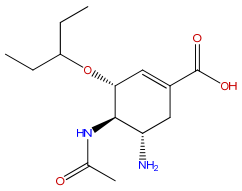
99		7c	25000	4
100		7d	17600	4
101		7e	2790	4
102		7f	7760	4
103		7g	240	4
104		7h	470	4
105		7i	25000	4

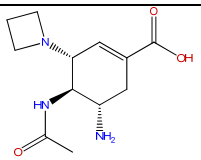
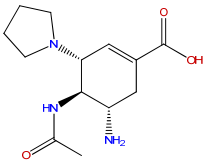
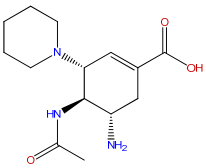
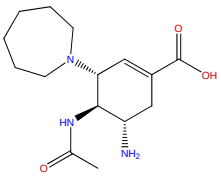
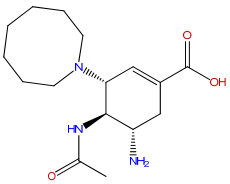
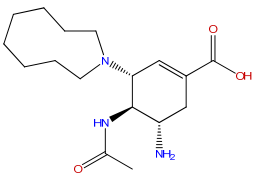
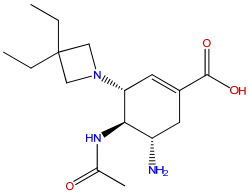
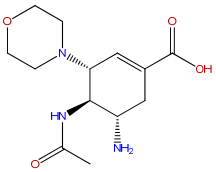
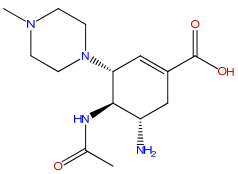
106		7j	290	4
107		7k	25000	4
108		7l	720	4
109		8a	320	4
110		8b	630	4
111		8c	25000	4
112		8d	25000	4

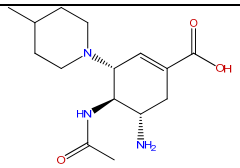
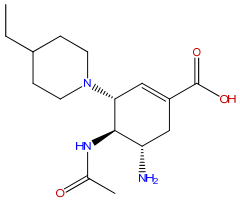
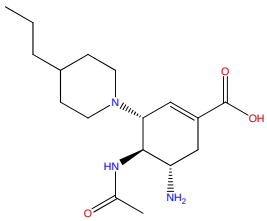
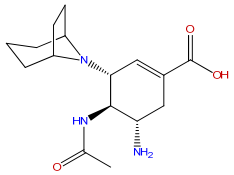
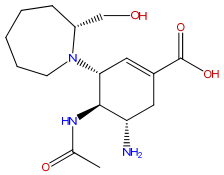
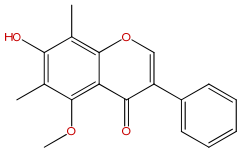
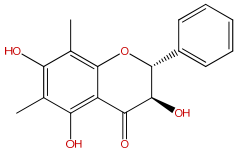
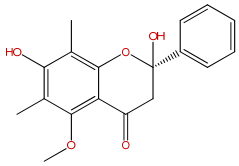
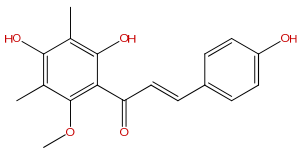
113		8e	25000	4
114		8f	410	4
115		8g	70	4
116		8h	150	4
117		8i	610	4
118		8j	870	4
119		8k	1670	4

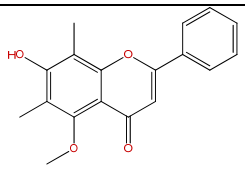
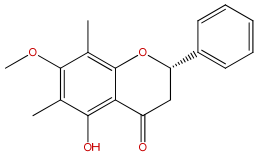
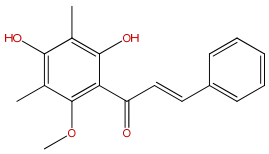
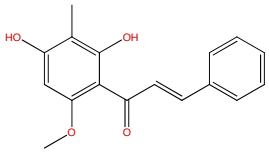
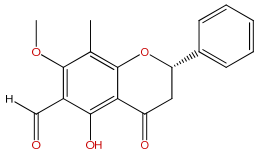
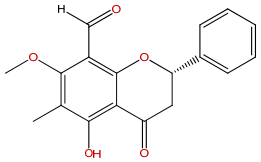
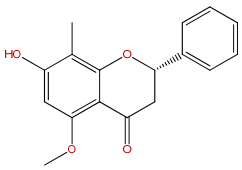
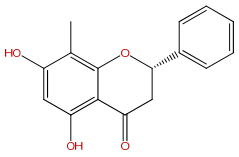
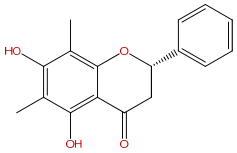
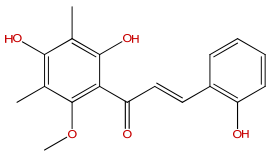
120		8l	6250	4
121		8m	22340	4
122		8n	260	4
123		8o	6620	4
124		8p	110	4
125		8q	640	4
126		8r	1710	4

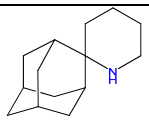
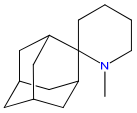
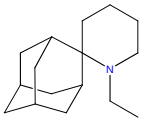
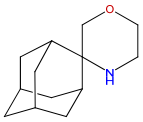
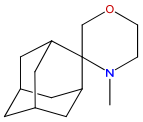
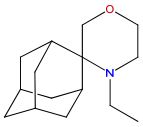
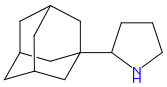
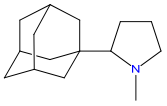
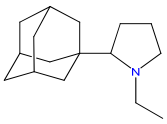
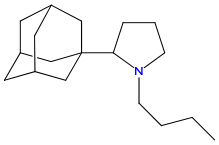
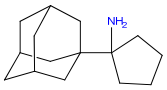
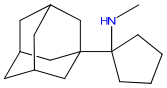
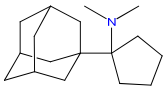
127		8s	260	4
128		8t	290	4
129		8u	720	4
130		9a	120	4
131		9b	150	4
132		9c	80	4
133		9d	140	4

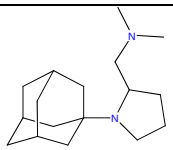
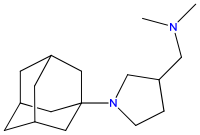
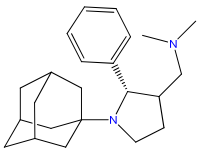
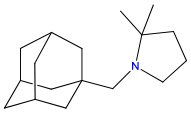
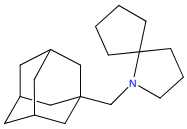
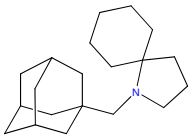
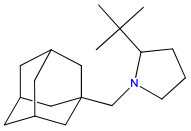
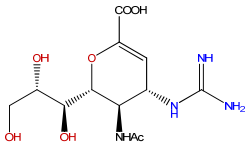
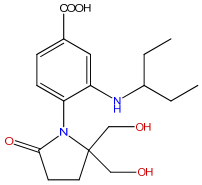
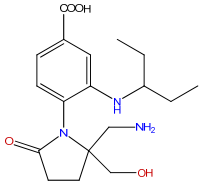
134		9e	7970	4
135		9f	5280	4
136		9g	6090	4
137		9h	2650	4
138		10a	90	4
139		10b	60	4
140		1	1	5

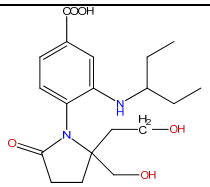
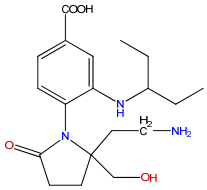
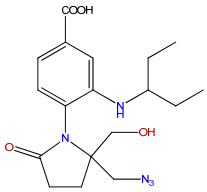
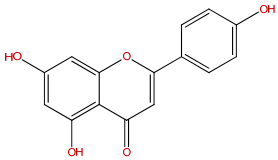
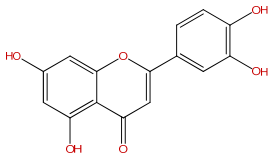
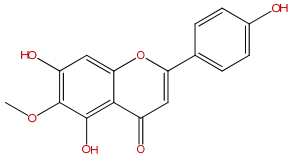
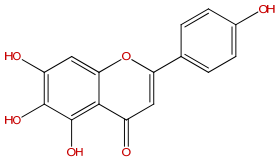
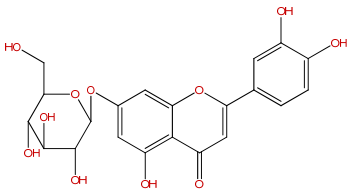
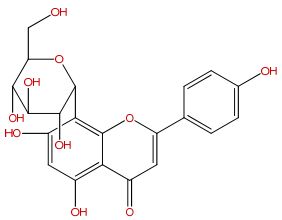
141		4a	1000	5
142		4b	310	5
143		4c	25	5
144		4d	35	5
145		4e	26	5
146		4f	75	5
147		4g	265	5
148		4h	260	5
149		4i	1900	5

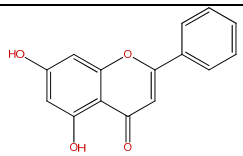
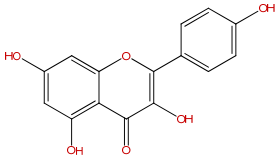
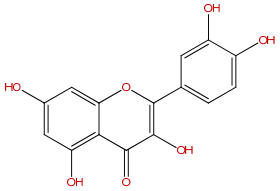
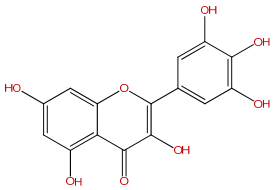
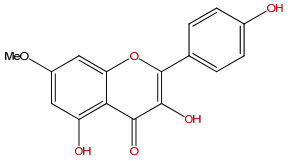
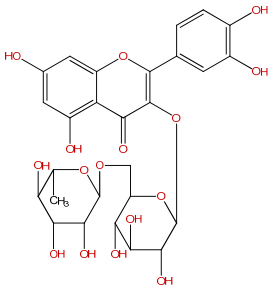
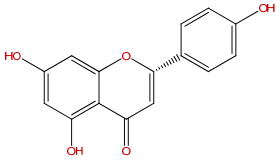
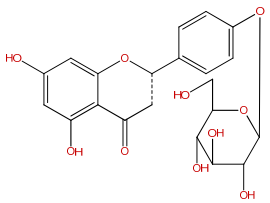
150		4j	51	5
151		4k	52	5
152		4l	40	5
153		4m	32	5
154		4n	8	5
155		1	430	6
156		2	347	6
157		3	282	6
158		4	20	6

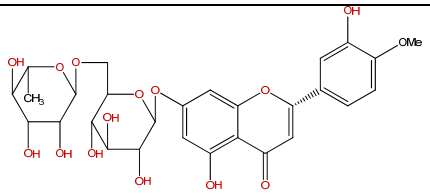
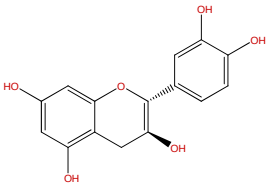
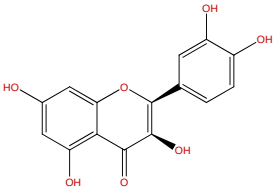
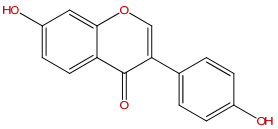
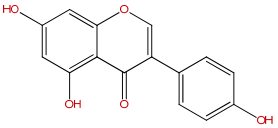
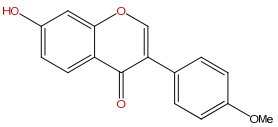
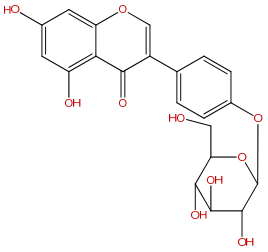
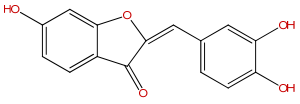
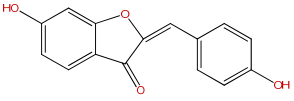
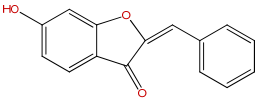
159		5	414	6
160		6	500	6
161		7	32	6
162		8	85	6
163		9	268	6
164		10	289	6
165		11	500	6
166		12	500	6
167		13	333	6
168		14	28	6

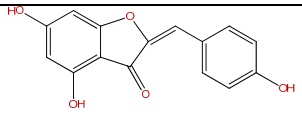
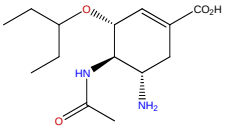
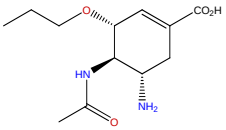
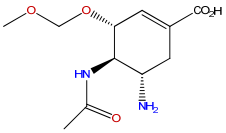
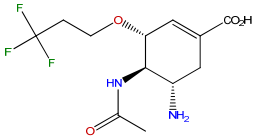
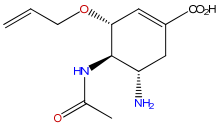
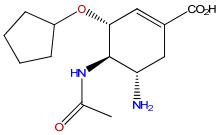
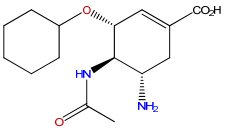
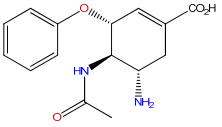
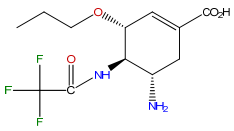
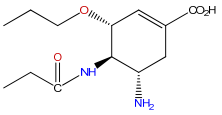
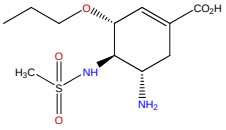
169		4a	35000	7
170		4b	6200	7
171		4c	4400	7
172		5a	29000	7
173		5b	22000	7
174		5c	10000	7
175		6a	15000	7
176		6b	34000	7
177		6c	34000	7
178		6d	4000	7
179		7a	64000	7
180		7b	59000	7
181		7c	80000	7

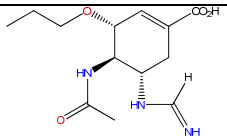
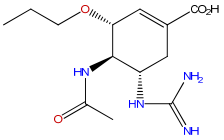
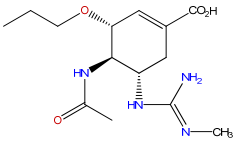
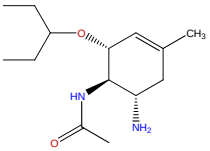
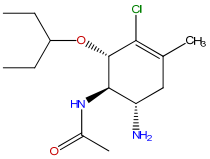
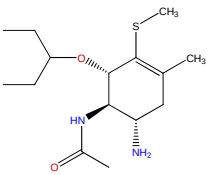
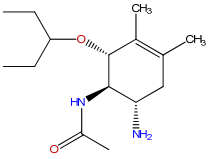
182		8a	70000	7
183		8b	36000	7
184		8c	36000	7
185		9a	150000	7
186		9b	150000	7
187		9c	150000	7
188		9d	104000	7
189		3	6	8
190		8	79000	8
191		9	880	8

192		10	1000000	8
193		11	180000	8
194		19	268000	8
195		1	31000	9
196		2	33000	9
197		3	46000	9
198		4	50000	9
199		5	47000	9
200		6	46000	9

201		7	45000	9
202		8	58000	9
203		9	58000	9
204		10	82000	9
205		11	51000	9
206		12	52000	9
207		13	100000	9
208		14	100000	9

209		15	100000	9
210		16	100000	9
211		17	100000	9
212		18	37000	9
213		19	77000	9
214		20	100000	9
215		21	100000	9
216		22	29000	9
217		23	22000	9
218		24	72000	9

219		25	25000	9
220		1a	1	10
221		3	130	10
222		4	2000	10
223		5	225	10
224		6	2200	10
225		7	22	10
226		8	60	10
227		9	530	10
228		12	100	10
229		13	1500	10
230		14	25000	10

231		16	140	10
232		18	2	10
233		19	5	10
234		1	1	11
235		2	3100	11
236		3	3400	11
237		4	2300	11

[1] Lew, W.; Wu, H. W.; Mendel, D. B.; Escarpe, P. A.; Chen X. W.; Laver, W. G.; Graves, B. J.; Kim, C. U. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 3321.

[2] Sun, C. W.; Huang, H.; Feng, M. Q.; Shi, X. L.; Zhang, X. D.; Zhou, P. *Bioorg. Med. Chem. Lett.* **2006,** *16*, 162.

[3] Wen, W. H.; Wang, S. Y.; Tsai, K. C.; Cheng, Y. S. E.; Yang, A. S.; Fang, J. M.; Wong, C. H. *Bioorg. Med. Chem.* **2010,** *18*, 4074.

[4] Yeh, J. Y.; Coumar, M. S.; Horng, J. T.; Shiao, H. Y.; Kuo, F. M.; Lee, H. L.; Chen, I. C.; Chang, C. W.; Tang, W. F.; Tseng, S. N.; Chen, C. J.; Shih, S. R.; Hsu, J. T. A.; Liao, C. C.; Chao, Y. S.; Hsieh, H. P. *J. Med. Chem.* **2010,** *53*, 1519.

[5] Lew, W.; Wu, H. W.; Chen, X. W.; Graves, B. J.; Escarpe, P. A.; MacArthur, H. L.; Mendel,

D. B.; Kim, C. U. *Bioorg. Med. Chem. Lett.* **2000,** *10*, 1257.

[6] Dao, T. T.; Tung, B. T.; Nguyen, P. H.; Thuong, P. T.; Yoo, S. S.; Kim, E. H.; Kim, S. K.; Oh, W. K. *J. Nat. Prod.* **2010,** *73*, 1636.

[7] Kolocouris, N.; Kolocouris, A.; Foscolos, G. B.; Fytas, G.; Neyts, J.; Padalko, E.; Balzarini, J.; Snoeck, R.; Andrei, G.; Clercq, E. D. *J. Med. Chem.* **1996,** *39*, 3307.

[8] Brouillette, W. J.; Bajpai, S. N.; Ali, S. M.; Velu, S. E.; Atigadda, V. R.; Lommer, B. S.; Finley, J. B.; Luo, M,; Aird, G. M. *Bioorg. Med. Chem.* **2003,** *11*, 2739.

[9] Liu, A. L.; Wang, H. D.; Lee, S. M. Y.; Wang, Y. T.; Du, G. H. *Bioorg. Med. Chem.* **2008,** *16*, 7141.

[10] Williams, M. A.; Lew, Willard.; Mendel, Dirk. B.; Tai, C. Y.; Escarpe, P. A.; Laver, W. G.; Stevens, R. C.; Kim, C. U. *Bioorg. Med. Chem. Lett.* **1997,** *14*, 1837.

[11] Zhang, L. J.; Williams, M. A.; Mendel, D. B.; Escarpe, P. A.; Kim, C. U. *Bioorg. Med. Chem. Lett.* **1997,** *14*, 1847.

附表2　训练集、测试集、独立验证集的化合物

Table S2　The Compounds of Training Set, Testing Set, and Independent Validation Set

	ID of the compounds
Training set	1, 2, 3, 7, 9, 14, 15, 16, 17, 19, 20, 24, 25, 27, 28, 32, 33, 34, 36, 37, 38, 39, 40, 42, 44, 45, 46, 47, 53, 56, 59, 60, 64, 65, 70, 71, 72, 75, 77, 78, 80, 81, 85, 87, 88, 90, 91, 93, 94, 95, 100, 101, 102, 106, 108, 109, 115, 117, 119, 121, 123, 124, 125, 127, 128, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 145, 149, 150, 152, 153, 154, 161, 167, 168, 170, 171, 173, 174, 178, 182, 184, 190, 191, 193, 195, 196, 198, 200, 203, 206, 208, 212, 217, 221, 222, 223, 224, 228, 229, 232, 233, 234, 237
Testing set	10, 11, 13, 22, 23, 26, 29, 35, 43, 49, 50, 54, 57, 58, 63, 66, 67, 74, 83, 86, 92, 97, 98, 103, 112, 118, 126, 129, 144, 146, 147, 151, 156, 157, 163, 169, 172, 176, 180, 183, 185, 187, 189, 192, 194, 201, 202, 207, 209, 210, 211, 214, 215, 218, 220, 230, 231, 236
Independent validation set	4, 5, 6, 8, 12, 18, 21, 30, 31, 41, 48, 51, 52, 55, 61, 62, 68, 69, 73, 76, 79, 82, 84, 99, 105, 107, 110, 111, 113, 116, 120, 122, 142, 143, 148, 158, 162, 164, 175, 177, 179, 181, 186, 188, 197, 199, 204, 205, 213, 216, 219, 225, 226, 235

附表 3 研究中所有分子描述符

Table S3 Molecular descriptors use in this work

descriptor Class	number of descriptor	descriptors
Simple molecular properties	18	molecular weight, numbers of rings, rotatable bonds, H-bond donors, and H-bond acceptors, element counts
Molecular connectivity and shape	27	molecular connectivity indices, valence molecular connectivity indices, molecular shape Kappa indices, Kappa alpha indices, flexibility index
Electro-topological state	97	Electrotopological state indices, and atom type electrotopological state indices, Weiner index, centric index, Altenburg index, Balaban index, Harary number, Schultz index, PetitJohn R2 index, PetitJohn D2 index, mean distance index, PetitJohn I2 index, information Weiner, Balaban rmsd index, graph distance index
Quantum chemical properties	22	polarizability index, hydrogen bond acceptor basicity (covalent HBAB), hydrogen bond donor acidity (covalent HBDA), molecular dipole moment, absolute hardness, softness, ionization potential, electron affinity, chemical potential, electronegativity index, electrophilicity index, most positive charge on H, C, N, O atoms, most negative charge on H, C, N, O atoms, most positive and negative charge in a molecule, LSum of squares of charges on H, C, N, O and all atoms, mean of positive charges, mean of negative charges, mean absolute charge, relative positive charge, relative negative charge
Geometrical properties	25	length vectors (longest distance, longest third atom, 4th atom), molecular van der Waals volume, solvent accessible surface area, molecular surface area, van der Waals surface area, polar molecular surface area, sum of solvent accessible surface areas of positively charged atoms, sum of solvent accessible surface areas of negatively charged atoms, sum of charge weighted solvent accessible surface areas of positively charged atoms, sum of charge weighted solvent accessible surface areas of negatively charged atoms, sum of van der Waals surface areas of positively charged atoms, sum of van der Waals surface areas of negatively charged atoms, sum of charge weighted van der Waals surface areas of positively charged atoms, sum of charge weighted van der Waals surface areas of negatively

	charged atoms, molecular rugosity, molecular globularity, hydrophilic region, hydrophobic region, capacity factor, hydrophilic-hydrophobic balance, hydrophilic intery moment, hydrophobic intery moment, amphiphilic moment