Journal of Molecular Graphics and Modelling 28 (2010) 714–727

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

QSAR models for predicting cathepsin B inhibition by small molecules—Continuous and binary QSAR models to classify cathepsin B inhibition activities of small molecules

Zhigang Zhou, Yanli Wang, Stephen H. Bryant *

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

ARTICLE INFO

Article history:
Received 3 June 2009
Received in revised form 22 January 2010
Accepted 24 January 2010
Available online 1 February 2010

Keywords:
Binary QSAR
Regression
Partial least squares
Docking
Cathepsin B protein
Screening
PubChem bioassay

ABSTRACT

Cathepsin B is a potential target for the development of drugs to treat several important human diseases. A number of inhibitors targeting this protein have been developed in the past several years. Recently, a group of small molecules were identified to have inhibitory activity against cathepsin B through high throughput screening (HTS) tests. In this study, traditional continuous and binary QSAR models were built to classify the biological activities of previously identified compounds and to distinguish active compounds from inactive compounds for drug development based on the calculated molecular and physicochemical properties. Strong correlations were obtained for the continuous QSAR models with regression correlation coefficients (r^2) and cross-validated correlation coefficients (q^2) of 0.77 and 0.61 for all compounds, and 0.82 and 0.68 for the compound set excluding 3 outliers, respectively. The models were further validated through the leave-one-out (LOO) method and the training-test set method. The binary models demonstrated a strong level of predictability in distinguishing the active compounds from inactive compounds with accuracies of 0.89 and 0.94 for active and inactive compounds, respectively, in non-cross-validated models. Similar results were obtained for the cross-validated models. Collectively, these results demonstrate the models' ability to discriminate between active and inactive compounds, suggesting that the models may be used to pre-screen compounds to facilitate compound optimization and to design novel inhibitors for drug development.

Published by Elsevier Inc.

1. Introduction

As members of the papain superfamily [1], cathepsins are involved in many biological processes related to human diseases and disorders [1–4]. Previously identified cathepsins include cathepsin B, D, H, K, L, and S. Several of these proteins have been selected as biological targets to develop therapeutic treatments, and a number of inhibitors have been identified and developed for many of these enzymes. Cathepsin B inhibitors are highly sought after chemical agents since many diseases, such as neurodegenerative disorder, cardiovascular disease, cancer, inflammation, rheumatoid arthritis, and Alzheimer's disease [5–12], have been connected with unusual levels or abnormal function of cathepsin B. As an ubiquitous lysosomal cysteine proteases, cathepsin B has been found to be responsible for intracellular as well as extracellular proteolysis in mammalian cells and can facilitate cell migration by dissolving the extracellular barriers, which result in tumor metastasis and angiogenesis [13–15]. The biological activity and function of cathepsin B is also important during viral infection and replication for several viruses, such as Ebola, SARS (Severe Acute Respiratory Syndrome) in human cells [16,17]. Due to its important biological functions, which are directly related to several important human diseases, cathepsin B has been chosen as a drug development target in many efforts.

A number of compounds have been found to inhibit cathepsin B activity, and some of these compounds have been tested and are effective in animal experiments [5–12,18–20]. Most of these cathepsin inhibitors disable the biological activity of cathepsin B through forming irreversible covalent chemical bond in the catalytic site of the enzyme. These irreversible inhibitors include dipeptidyl nitriles [21], vinyl sulfones, expoxysuccinates, acyloxymethyl ketones, fluoromethyl ketones, hydrazides, and *bis*-α-amidoketones [22]. Structural studies have provided detailed insights into the biological mechanisms of these inhibitors. Inhibitors bind to the catalytic active site of cathepsin B, and form an irreversible covalent bond with the protein in the active site [21,23]. Meanwhile, computational studies, such as docking and virtual screening, were also used to explore the binding and

* Corresponding author at: NCBI/NIH, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA. Tel.: +1 301 435 7792; fax: +1 301 480 9241.
E-mail addresses: zhougeor@ncbi.nlm.nih.gov (Z. Zhou), bryant@ncbi.nlm.nih.gov (S.H. Bryant).

1093-3263/$ – see front matter. Published by Elsevier Inc.
doi:10.1016/j.jmgm.2010.01.009

Table 1
The compound structure and biological activity of the active and inactive compounds from PubChem bioassay database.[a].

Compd	CID[b]	Mol. formula	SMILES	MW	IC_{50} (μM)	log IC_{50}	Active
1	286532	$C_{18}H_{14}N_2O_6$	o1[n+]([O−])c(C(=O)c2ccc(OC)cc2)c(n1)C(=O)c1ccc(OC)cc1	354.09	11.46	−4.941	1
2	573353	$C_{16}H_{12}FN_5O$	Fc1ccc(cc1)Cn1c2c(nc1–c1nonc1N)cccc2	309.10	33.86	−4.470	1
3	646525	$C_{15}H_{13}N_3O_4S_2$	s1cccc1C(Oc1nn(S(=O)(=O)c2ccc(cc2)C)c(N)c1)=O	363.04	1.99	−5.701	1
4	646749	$C_{18}H_{17}N_3O_5S$	S(=O)(=O)(n1nc(OC(=O)c2ccc(cc2)C)cc1N)c1ccc(OC)cc1	387.09	12.27	−4.911	1
5	647599	$C_{14}H_{10}FN_3O_5S$	S(=O)(=O)(n1nc(OC(=O)c2occc2)cc1N)c1ccc(F)cc1	351.03	1.26	−5.899	1
6	648315	$C_{17}H_{14}N_2O_3S_2$	s1cccc1CNC(=O)C(OC(=O)c1sccc1)c1ncccc1	358.05	6.36	−5.197	1
7	651936	$C_{15}H_{13}N_3O_5S$	S(=O)(=O)(n1nc(OC(=O)c2occc2)cc1N)c1ccc(cc1)C	347.06	1.75	−5.757	1
8	653316	$C_{16}H_{18}N_6O_2$	o1nc(–c2nc3c(n2CC(=O)N2CCCCC2)cccc3)c(n1)N	326.15	44.58	−4.351	1
9	653862	$C_{15}H_{13}N_3O_6S$	S(=O)(=O)(n1nc(OC(=O)c2occc2)cc1N)c1ccc(OC)cc1	363.05	0.92	−6.035	1
10	654815	$C_7H_6Cl_2N_2O_4$	ClC1=C(Cl)C(OC1NC(=O)NC(=O)C)=O	251.97	2.12	−5.674	1
11	655490	$C_{17}H_{15}N_3O_5S$	S(=O)(=O)(n1nc(OC(=O)c2ccccc2)cc1N)c1ccc(OC)cc1	373.07	9.56	−5.019	1
12	658111	$C_{17}H_{21}N_3O_4S$	S(C(OC)=O)c1nc(N2CCOCC2)c2c(CC(OC2)(C)C)c1C#N	363.13	6.72	−5.173	1
13	658152	$C_{14}H_{20}N_6O_4S$	S(C)c1nc(nc(n1)N(C)C)N(C(C(OCC)=O)C(OCC)=O)C#N	368.13	19.69	−4.706	1
14	658724	$C_{19}H_{16}N_2O_4$	o1c(nc(C=Nc2ccccc2OC)c1OC(=O)C)–c1ccccc1	336.11	8.93	−5.049	1
15	658964	$C_{20}H_{18}N_2O_4$	o1c(nc(C=Nc2ccccc2OC)c1OC(=O)CC)–c1ccccc1	350.13	39.99	−4.398	1
16	660829	$C_{19}H_{12}N_2O_5$	o1c(nc(C=Nc2ccccc2)c1OC(=O)c1occc1)–c1occc1	348.08	38.47	−4.415	1
17	665480	$C_{20}H_{29}N_3O_5S$	S(=O)(=O)(Cc1cc(ccc1)C)c1oc(nn1)[C@H](NC(OC(C)(C)C)=O)C(CC)C	423.18	2.09	−5.680	1
18	714967	$C_{11}H_{15}N_7O_2$	O(C)c1nc(nc(n1)N(CC(=O)N)C#N)N1CCCC1	277.13	14.20	−4.848	1
19	794694	$C_{12}H_{14}ClNO_3S_2$	Clc1ccc(S(=O)(=O)N2S(=O)CC(C)=C(C2)C)cc1	319.01	4.17	−5.380	1
20	971438	$C_{18}H_{17}N_5O_2$	o1nc(–c2nc3c(n2Cc2cc(OC)ccc2C)cccc3)c(n1)N	335.14	37.19	−4.430	1
21	1506381	$C_{17}H_{15}N_5O_2$	o1nc(-c2nc3c(n2Cc2cc(OC)ccc2)cccc3)c(n1)N	321.12	45.97	−4.338	1
22	2212050	$C_{15}H_{13}N_3OS$	S(CC)c1ccccc1C(=O)n1nnc2c1cccc2	283.08	7.11	−5.148	1
23	2998380	$C_{16}H_{16}N_2O_4S_3$	S(=O)(=O)(N1S(=NS(=O)(=O)c2ccccc2)CC=CC1)c1ccccc1	396.03	9.39	−5.027	1
24	3236798	$C_{16}H_{14}N_6O_2$	O=C1N(C)C(=O)N(c2c1c(n(–n1cnnc1)c2)–c1ccccc1)C	322.12	1.19	−5.926	1
25	3240114	$C_{15}H_{13}N_3O_5S_2$	s1cccc1C(Oc1nn(S(=O)(=O)c2ccc(OC)cc2)c(N)c1)=O	379.03	0.69	−6.160	1
26	3241895	$C_{14}H_{10}FN_3O_4S_2$	s1cccc1C(Oc1nn(S(=O)(=O)c2ccc(F)cc2)c(N)c1)=O	367.01	0.44	−6.362	1
27	3243025	$C_{11}H_8F_3NO_4$	FC(F)(F)C1(Oc2c(NC1=O)cccc2)OC(=O)C	275.04	0.85	−6.073	1
28	3243128	$C_{14}H_{11}N_3O_4S_2$	s1cccc1C(Oc1nn(S(=O)(=O)c2ccccc2)c(N)c1)=O	349.02	0.26	−6.608	1
29	3243168	$C_{16}H_{12}N_2O_6$	o1cccc1C(=O)NCC(OCN1C(=O)c2c(cccc2)C1=O)=O	328.07	8.56	−5.067	1
30	3250046	$C_{22}H_{19}NO_6$	o1cccc1C(=O)NCC(OC(C(=O)c1ccc(OC)cc1)c1ccccc1)=O	393.12	18.35	−4.736	1
31	5293426	$C_{19}H_{14}N_4O_3S$	S(CC(=O)Nc1ccccc1)c1nc(c(nn1)–c1occc1)–c1occc1	378.08	2.25	−5.648	1
32	11834381	$C_{15}H_{13}N_3O_4S_2$	s1cc(cc1C(Oc1nn(S(=O)(=O)c2ccccc2)c(N)c1)=O)C	363.41	2.82	−5.550	1
33	11834392	$C_{20}H_{15}N_3O_6S_3$	s1cccc1C(=O)Nc1n(S(=O)(=O)c2ccc(OC)cc2)nc(OC(=O)c2sccc2)c1	489.54	3.23	−5.490	1
34	3685806	$C_9H_9N_3O_4S_2$	s1cccc1C(Oc1nn(S(=O)(=O)C)c(N)c1)=O	287.32	22.28	−4.652	1
35	11834389	$C_{15}H_{12}N_2O_4S_2$	s1cccc1C(Oc1nn(S(=O)(=O)c2ccc(cc2)C)cc1)=O	348.40	33.10	−4.480	1
36	10145	$C_{21}H_{25}NO_4$	CN1CCC2=CC(=C(C3=C2C1CC4=CC(=C(C=C43)OC)OC)OC)OC	355.18			0
37	1205147	$C_{23}H_{26}N_6O_4$	CCOC1=C(C=C(C=C1)CCNC(=O)CN2C3=CC=CC=C3N=C2C4=NON=C4N)OCC	450.20			0
38	1248970	$C_{20}H_{22}N_2O_5S$	CC1=C(C=C(C=C1)C(=O)NC2=CC3=C(C=C2)OCO3)S(=O)(=O)N4CCCCC4	402.12			0
39	1306035	$C_{26}H_{20}N_2O_4$	COC1=CC=CC=C1OCC(=O)NC2=C(C(=C(O2)C3=CC=CC=C3)C4=CC=CC=C4)C#N	424.14			0
40	1505224	$C_{19}H_{16}N_6O_4$	COC(=O)C1=CC=CC=C1NC(=O)CN2C3=CC=CC=C3N=C2C4=NON=C4N	392.12			0
41	2612950	$C_{10}H_6N_4OS$	C1=CSC(=C1C#N)NC(=O)C2=NC=CN=C2	230.03			0
42	3236055	$C_{21}H_{21}ClN_2O_3$	CCN1C2=CC=CC=C2C(=CC1=O)OCC(=O)N(CC)C3=CC(=CC=C3)Cl	384.12			0
43	3236935	$C_{19}H_{25}N_3O_6S$	CC1=C(C(=NO1)C)S(=O)(=O)N(CC(=O)NCC2CCCO2)C3=CC=C(C=C3)OC	423.15			0
44	3238028	$C_{21}H_{27}N_3O_6S_2$	CCOC(=O)C1=C(N(C(=NC(=O)C2=CC=C(C=C2)S(=O)(=O)N3CC(OC(C3)C)C)S1)C)C	481.13			0
45	3239534	$C_{21}H_{29}N_5O_3S$	CC1=C(C(=CC=C1)N2CCN(CC2)C(=O)C3CCN(CC3)S(=O)(=O)C4=CN=CN4)C	431.20			0
46	3239997	$C_{20}H_{19}ClN_2O_3$	CCN(C1=CC(=CC=C1)Cl)C(=O)COC2=CC(=O)N(C3=CC=CC=C32)C	370.11			0
47	3240677	$C_{15}H_{12}FNO_3$	C1=CC=C(C(=C1)C(=O)NCC2=CC=C(C=C2)F)C(=O)O	273.08			0
48	3240711	$C_{22}H_{26}N_4O_4S$	CC1=C(C(=CC=C1)NC(=O)CN(C2=CC=C(C=C2)OC)S(=O)(=O)C3=C(NN=C3C)C)C	442.17			0
49	3241211	$C_{20}H_{18}N_6O_4$	CCOC(=O)C1=CC=CC=C1NC(=O)CN2C3=CC=CC=C3N=C2C4=NON=C4N	406.14			0
50	3244032	$C_{14}H_{12}N_2S_2$	CSC1=C(C(=CC(=N1)C2CC2)C3=CC=CS3)C#N	272.04			0
51	3333	$C_{18}H_{19}Cl_2NO_4$	CCOC(=O)C1=C(NC(=C(C1C2=C(C(=CC=C2)Cl)Cl)C(=O)OC)C)C	383.07			0
52	3422818	$C_{20}H_{21}FN_2O_2$	C1CC(=O)N(C1C(=O)NCC2=CC=C(C=C2)F)CCC3=CC=CC=C3	340.16			0
53	380199	$C_{11}H_{13}N_3OS$	CN(C)C=NC1=NC2=C(S1)C=C(C=C2)OC	235.08			0
54	4226014	$C_{21}H_{27}N_3O_2$	CC1=C(C=C(C=C1)NC(=O)NCC2CCN(C2)C3=CC=C(C=C3)OC)C	353.21			0
55	5308432	$C_{15}H_{17}N_7$	CC1=NN=C2N1N=C(C=C2)N3CCN(CC3)C4=CC=CC=N4	295.15			0
56	5308489	$C_{18}H_{21}BrN_6O_2$	C1CN(CCN1C2=NN=C(C=C2)N3CCOCC3)C(=O)C4=CC(=CN=C4)Br	432.09			0
57	571349	$C_{17}H_{15}N_5O$	CC1=CC=C(C=C1)CN2C3=CC=CC=C3N=C2C4=NON=C4N	305.13			0
58	599637	$C_8H_4BrN_5O_3$	C1=C(OC(=C1)Br)C2=NC(=NO2)C3=NON=C3N	296.95			0
59	6023689	$C_8H_9ClN_2O_2S$	C/C(=N\S(=O)(=O)C1=CC=C(C=C1)Cl)/N	232.01			0
60	651703	$C_9H_{12}N_4O_5$	C1COCCN1C(=O)COC(=O)C2=NON=C2N	256.08			0
61	655872	$C_{25}H_{26}N_6O_4$	COCCNC(=O)C(C1=CC(=CC=C1)OC)N(C2=CN=CC=C2)C(=O)CN3C4=CC=CC=C4N=N3	474.20			0
62	658581	$C_9H_{12}N_6O_2$	C1CCN(CC1)C2=NC(=NO2)C3=NON=C3N	236.10			0
63	6603449	$C_{19}H_{21}ClFN_5O_3$	C1=CC=C(C(=C1)CNCCNC(=O)C2=NON=C2N)OCC3=CC=CC=C3FCl	421.13			0
64	660518	$C_{19}H_{19}N_3O_5S$	CCOC(=O)CNS(=O)(=O)C1=C(C=CC(=C1)C2=NNC(=O)C3=CC=CC=C32)C	401.10			0
65	664136	$C_{20}H_{21}N_5O_2$	CCCCC1=C2C(=NN1)OC(=C(C23C4=CC=CC=C4N(C3=O)CC)C#N)N	363.17			0
66	664291	$C_{21}H_{19}N_3O_2$	C1CN(CC2=CC=CC=C21)C3=NC=C4C(=N3)CC(CC4=O)C5=CC=CO5	345.15			0
67	665203	$C_{15}H_{18}N_2O_4$	CC(C)(CC1=C[N+](=C2C=C(C=CC2=N1)OC)[O−])CC(=O)O	290.13			0
68	693069	$C_{11}H_{13}NO_4$	COC1=CC=CC=C1NC(=O)CCC(=O)O	223.08			0
69	715594	$C_{15}H_{17}N_5O_3$	CC1=CC(=C(C=C1)C(C)C)OCC2=NC(=NO2)C3=NON=C3N	315.13			0
70	730135	$C_{16}H_{16}N_2O$	C1CC1C(=O)NC2=CC=C(C=C2)NC3=CC=CC=C3	252.13			0
71	1093235	$C_{11}H_{13}N_3O_3S$	S(=O)(=O)(n1nc(OC)cc1N)c1ccc(cc1)C	267.30			0
72	11834379	$C_{10}H_{11}N_3O_3S$	S(=O)(=O)(N1NC(=O)C=C1N)c1ccc(cc1)C	253.28			0
73	11834380	$C_{16}H_{14}N_4O_4S$	s1cccc1C(Oc1nn(C(=O)Nc2ccc(OC)cc2)c(N)c1)=O	358.37			0
74	11834382	$C_{18}H_{13}N_3O_4S_2$	s1c2c(cc1C(Oc1nn(S(=O)(=O)c3ccccc3)c(N)c1)=O)cccc2	399.44			0
75	11834384	$C_{16}H_{17}N_3O_7S$	S(=O)(=O)(n1nc(OC(=O)C)cc1N(C(=O)C)C(=O)C)c1ccc(OC)cc1	395.39			0

Table 1 (*Continued*)

Compd	CID[b]	Mol. formula	SMILES	MW	IC_{50} (μM)	log IC_{50}	Active
76	11834385	$C_{21}H_{17}N_3O_5S_2$	s1cccc1C(Oc1nn(S(=O)(=O)c2ccc(cc2)–c2ccc(OC)cc2)c(N)c1)=O	455.51			0
77	11834386	$C_{16}H_{14}N_4O_5S$	S(=O)(=O)(n1nc(OC(=O)c2ccncc2)cc1N)c1ccc(OC)cc1	374.37			0
78	11834387	$C_{15}H_{15}N_3O_4S_2$	s1cccc1COc1nn(S(=O)(=O)c2ccc(OC)cc2)c(N)c1	365.43			0
79	11834388	$C_{10}H_{10}N_2O_3S$	S(=O)(=O)(N1NC(=O)C=C1)c1ccc(cc1)C	238.26			0
80	11834390	$C_{15}H_{14}N_4O_5S_2$	s1cccc1NC(=O)NC=1N(S(=O)(=O)c2ccc(OC)cc2)NC(=O)C=1	394.43			0
81	11834391	$C_{15}H_{13}N_3O_4S_2$	s1cccc1C(=O)NC=1N(S(=O)(=O)c2ccc(cc2)C)NC(=O)C=1	363.41			0
82	11834393	$C_{14}H_{13}N_3O_6S_3$	s1cccc1S(Oc1nn(S(=O)(=O)c2ccc(OC)cc2)c(N)c1)(=O)=O	415.47			0
83	11834394	$C_{11}H_{10}F_3N_3O_6S_2$	S(=O)(=O)(n1nc(OS(=O)(=O)C(F)(F)F)cc1N)c1ccc(OC)cc1	401.34			0
84	11834395	$C_9H_{12}O_3S_3$	s1cccc1C(SCC(O)C(O)CS)=O	264.38			0
85	5035758	$C_{10}H_{11}N_3O_3S$	S(=O)(=O)(NNC(=O)CC#N)c1ccc(cc1)C	253.28			0
86	573755	$C_{11}H_8O_2S$	s1cccc1C(Oc1ccccc1)=O	204.25			0

[a] The information was obtained from PubChem database (BioAssay #: 820 and 523).
[b] CID: Pubchem compound accession. Mol. formula: molecular formula, MW: molecular weight.

inhibition mechanism of these inhibitors [24–26]. These computational approaches have proved efficient and essential for optimizing and designing chemical agents with improved biological activities [27–29].

In an effort to identify chemical probes through high throughput screening (HTS) technology, a number of chemicals have been found to be active against cathepsin B in the screening campaign for inhibitor identification, with the NIH Molecular Libraries Program (MLP). The screening results were deposited into the PubChem (http://pubchem.ncbi.nlm.nih.gov), a database for molecular structures and associated biological activities which is available to general public. The database contains the 2D structures and the biological activity (IC_{50}) derived from dose–response HTS experiments for the active compounds tested.

In a previous study [30,31], docking simulation was used to model the binding structures of the active compounds identified by the high throughput screening (HTS) tests to the binding site of the protein. A relative binding affinity was calculated for each compound studied based on the respective modeled bound structure using the linear response molecular mechanics Poisson Boltzmann-surface area method (LR-MM-PBSA). Strong correlations between the calculated binding affinities and experimental biological activities were obtained [30,31]. Three-dimensional (3D) Comparative Molecular Field Analysis (CoMFA) quantitative structure–activity relationships (QSAR) models were also established based on the multi-conformation method with high correlation coefficients for these active compounds (to be published). Given the importance of the enzyme in disease treatment and the broad research interest to screen and identify potent inhibitors, efforts have been taken further to build the QSAR models to correlate the molecular and physicochemical properties with the biological activities of the compounds tested in the Cathepsin B inhibition screening experiments. Efforts were also undertaken to build statistical models to classify inhibition activities against Cathepsin B for specific compounds. This work reports the results of the continuous (continuous in bioactivity space) and binary (two discrete bioactivity status, e.g. 'active' *vs.* 'inactive') QSAR models in order to obtain insight into the relationship of chemical structure and physicochemical properties/biological activity. Additionally, the results of this work may provide means to pre-screen compounds to facilitate *in silico* design of de novo cathepsin B inhibitors and the development of drugs for the treatment of various human diseases using some other molecular targets.

2. Materials and methods

2.1. Data set and molecule preparation

The 3D chemical structures of all small molecule compounds were first converted from the 2D molecular structures obtained from the PubChem database (PubChem bioAssay accession: 820 and 523) [32], which were subsequently subjected to energy minimization calculation until the root mean square deviation (RMSD) of potential energy was smaller than 0.001, using the Optimized Potentials for Liquid Simulations (OPLS) force field [33–36] by the Molecular Operating Environment (MOE) program (Version 2007.09, developed by Chemical Computing Group, Montreal, Canada). In Bioassay AID (Pubchem bioassay accession) 820, 75 compounds were tested by a dose–response confirmatory screening experiment, whereas 37 compounds were confirmed to have inhibitory activity, 35 compounds were confirmed to have no

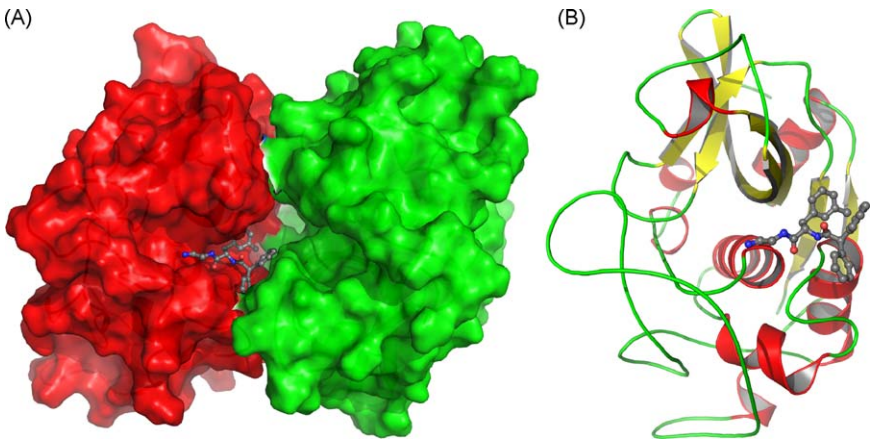

Fig. 1. The active site located in the dimer interface of cathepsin protein complexed with ligand DNP (stick-ball mode) in crystal complex (PDB code: 1GMY). (A) The dimer proteins are shown with surface mode with the primary protein in red and the second protein in green. (B) The primary protein is depicted with a ribbon diagram (red for helixes, green for loops, and yellow for β-sheets) to show the binding location and conformation of the ligand DNP (stick-ball mode in grey).

Table 2
The calculated molecular and physicochemical properties (descriptors) of 86 compounds used to build QSAR models.

Compd.	E_{sol}	Apol	MR	Dipole	SA_{pol}	SA	Vol	$\log P_{(o/w)}$	N_{acc}	N_{don}	$N_{acc+don}$
1	−2.42	48.03	9.25	1.39	43.03	355.29	316.25	2.51	5	0	5
2	−2.60	43.02	8.38	0.91	45.20	302.53	277.88	3.52	3	1	4
3	−2.73	47.38	9.28	1.69	72.75	351.85	309.00	2.69	4	1	5
4	−3.26	53.23	10.09	1.98	75.25	380.19	337.88	3.12	5	1	6
5	−6.53	42.07	8.33	1.12	72.75	316.24	281.63	1.73	4	1	5
6	4.86	49.66	9.86	0.38	38.50	364.58	320.75	2.46	3	1	4
7	−7.46	45.28	8.73	1.71	72.75	334.83	295.00	1.88	4	1	5
8	−1.70	48.37	8.82	0.28	58.76	333.24	301.25	1.88	4	1	5
9	−8.40	46.08	8.91	1.37	75.25	348.53	301.63	1.54	5	1	6
10	−0.10	26.09	5.23	0.70	52.07	228.61	186.25	0.76	3	2	5
11	−6.83	50.13	9.64	1.73	75.25	369.86	325.25	2.82	5	1	6
12	1.95	53.33	9.56	0.75	42.00	366.33	334.88	2.34	5	0	5
13	4.09	50.68	9.36	0.45	61.92	388.44	335.00	0.03	6	0	6
14	2.25	49.52	9.46	0.86	27.44	357.59	316.88	3.39	4	0	4
15	1.63	52.61	9.94	0.80	27.44	381.60	338.00	3.86	4	0	4
16	−0.35	47.65	9.44	1.53	24.93	352.29	316.25	2.53	3	0	3
17	−24.17	64.75	11.27	2.46	70.11	417.28	394.63	4.32	5	1	6
18	−2.42	38.67	7.00	0.82	66.10	295.12	252.38	−1.65	5	1	6
19	−18.91	41.94	7.89	2.23	48.02	301.79	265.00	0.72	3	0	3
20	−3.85	50.12	9.44	1.11	47.70	334.35	312.63	3.65	4	1	5
21	−2.94	47.03	8.98	0.86	47.70	330.19	299.88	3.36	4	1	5
22	0.68	42.07	8.21	1.23	36.15	282.17	263.75	3.92	3	0	3
23	−20.71	52.94	10.09	2.47	74.50	371.50	332.00	2.03	2	0	2
24	−0.23	45.70	8.66	0.22	45.98	345.12	310.75	0.93	4	0	4
25	−2.98	48.18	9.46	1.65	75.25	360.35	313.88	2.35	5	1	6
26	−0.96	44.17	8.88	0.92	72.75	322.59	287.13	2.55	4	1	5
27	7.73	30.67	5.77	0.87	35.32	237.88	209.25	2.13	3	1	4
28	−6.75	44.28	8.82	2.56	72.75	322.89	288.50	2.40	4	1	5
29	−3.59	43.17	8.28	1.22	59.95	326.87	286.25	0.58	4	1	5
30	3.06	57.30	10.79	0.79	48.89	323.03	337.13	2.90	4	1	5
31	0.31	52.48	10.44	0.98	43.77	373.78	339.00	2.97	4	1	5
32	−4.71	47.38	9.28	1.81	72.75	347.05	304.50	2.73	4	1	5
33	−25.98	62.01	12.39	4.04	76.76	442.14	398.63	3.51	6	1	7
34	−7.17	34.15	6.71	1.17	72.75	256.56	221.38	0.74	4	1	5
35	0.45	45.61	9.01	1.48	55.00	336.87	293.50	2.92	4	0	4
36	3.02	98.35	17.60	0.86	57.94	410.70	404.25	5.31	10	0	10
37	−11.90	108.04	19.91	1.78	117.37	441.79	460.75	4.99	11	2	13
38	−38.13	99.39	18.17	1.01	104.19	462.33	426.75	4.64	10	1	11
39	0.75	104.92	19.89	1.11	89.92	465.95	474.50	6.88	9	1	10
40	7.49	94.33	17.92	0.62	125.93	416.34	421.00	4.32	10	2	12
41	−2.68	70.12	13.76	1.25	96.28	306.93	294.13	2.10	9	1	10
42	5.75	98.16	18.26	0.76	77.56	433.12	432.75	6.08	8	0	8
43	5.63	101.53	18.44	0.64	115.08	423.36	430.75	2.75	11	1	12
44	−2.94	109.29	19.92	1.24	115.26	510.59	485.88	5.18	11	0	11
45	1.91	107.52	19.22	0.86	104.87	478.28	481.75	2.39	10	2	12
46	4.65	95.07	17.78	0.74	77.56	411.88	411.00	5.74	8	0	8
47	6.67	78.88	14.95	0.99	94.30	332.24	324.13	4.85	8	3	11
48	−2.90	106.98	19.66	1.72	120.53	480.71	471.75	4.77	11	3	14
49	6.87	97.42	18.39	0.53	125.93	398.43	414.13	4.67	10	2	12
50	7.26	81.05	15.46	0.53	71.35	355.71	350.75	5.49	7	0	7
51	7.87	93.43	17.46	0.08	80.74	483.06	457.13	6.19	7	1	8
52	2.64	93.98	17.25	0.72	80.74	371.23	379.75	4.88	7	1	8
53	0.40	75.44	14.21	0.25	61.79	334.58	312.00	4.47	8	0	8
54	−0.71	100.28	17.96	0.80	75.36	446.78	459.25	5.36	7	2	9
55	2.35	85.85	15.75	0.56	81.87	345.63	349.25	3.01	9	0	9
56	3.24	97.35	18.02	0.77	88.52	431.17	427.75	3.51	10	0	10
57	−6.95	86.64	16.30	0.84	93.12	371.62	355.50	5.69	8	1	9
58	−16.67	68.12	13.47	0.71	104.00	315.16	295.50	2.93	9	1	10
59	−5.24	69.38	13.24	2.33	105.75	319.01	296.00	3.93	8	1	9
60	1.82	72.66	13.42	0.74	117.07	325.83	312.75	0.28	10	1	11
61	3.20	111.56	20.76	1.79	114.01	480.30	484.63	3.59	12	1	13
62	−0.34	72.46	13.37	1.10	104.00	319.83	304.13	2.47	9	1	10
63	3.43	95.65	17.86	0.52	114.87	377.95	388.38	4.04	10	3	13
64	−4.71	96.73	18.06	1.40	133.99	393.38	397.25	4.88	10	2	12
65	−1.42	96.72	17.66	0.93	115.81	494.46	475.63	5.22	9	3	12
66	1.76	94.95	17.51	0.76	72.85	414.17	398.75	3.66	8	0	8
67	−3.81	84.22	15.38	0.47	100.21	370.84	365.25	4.53	10	2	12
68	−1.49	72.75	13.47	1.35	96.81	324.38	301.00	2.73	9	3	12
69	−10.12	86.05	15.95	0.58	106.50	389.15	371.13	4.47	10	1	11
70	1.44	82.24	15.16	0.37	72.85	344.02	331.75	4.96	6	2	8
71	−11.49	36.63	6.85	2.49	59.18	273.47	232.63	1.52	3	1	4
72	−13.03	33.54	6.27	2.48	72.75	246.89	214.50	0.37	3	2	5
73	4.03	48.00	9.39	0.19	62.48	351.27	314.13	2.72	4	2	6
74	−27.88	52.66	10.45	1.55	72.75	362.57	330.50	3.87	4	1	5
75	−31.16	51.31	9.63	3.59	84.64	382.88	334.25	1.30	7	0	7
76	−3.82	61.41	11.98	2.26	75.25	433.90	390.13	4.31	5	1	6

Table 2 (*Continued*)

Compd.	E_{sol}	Apol	MR	Dipole	SA_{pol}	SA	Vol	$\log P_{(o/w)}$	N_{acc}	N_{don}	$N_{acc+don}$
77	−7.18	48.81	9.46	2.16	80.93	355.59	315.75	1.59	6	1	7
78	−0.04	48.71	9.40	1.46	61.68	346.64	309.63	2.50	4	1	5
79	−3.28	31.77	6.01	1.44	55.00	234.94	202.25	0.35	3	1	4
80	−51.53	49.95	9.69	5.22	82.44	363.13	320.13	1.16	5	3	8
81	−32.63	47.38	9.21	4.50	74.25	335.87	301.75	1.53	4	2	6
82	−27.45	50.12	9.81	3.67	93.70	367.80	322.75	2.08	6	1	7
83	−28.50	41.61	8.03	2.54	93.70	327.67	278.00	2.01	6	1	7
84	1.48	34.95	6.71	1.04	40.70	256.19	222.13	1.32	3	2	5
85	−6.64	33.54	6.42	1.54	84.56	262.41	219.25	0.40	4	2	6
86	−8.75	29.20	5.84	1.13	13.57	217.66	189.63	3.03	1	0	1

Apol: sum of atomic polarizabilities. MR: molar refractivity. Dipole: the dipole moment. SA_{pol}: hydrophilic surface area. SA: solvent accessible surface area. Vol: molecular volume. N_{acc}: the number of hydrogen bond acceptors. N_{don}: the number of hydrogen bond donors. $N_{acc+don}$: the summary of the number of hydrogen bond acceptors and doners. $\log P_{(o/w)}$: log octanol/water partition coefficient.

inhibitory activity, and the bioactivity outcome for the remaining 3 compounds were reported as "inconclusive". By further examining these active compounds using different experimental protocol, assay depositors suggested that some of them were likely artifacts. Further investigation using several independent bioassay tests (different biological experiments), which may be affected by similar causes of artifact, suggested that five compounds bear greater chance as being false positives, and they were removed from the current analysis. So, 32 active ($37 - 5$) and 35 inactive compounds from bioassay 820 were included in the work. Bioassay 523 employed a similar experimental protocol to the Bioassay 820, where 27 compounds were tested, 10 compounds were confirmed having inhibitory activity, 16 compounds were confirmed having no inhibitory activity, and one compound was reported as "inconclusive". Four out of the 10 active compounds are unique structures and were added to the active compound list. Total 36 active compounds (32 from bioassay 820 and 4 from bioassay 523) and 51 inactive compounds (35 from bioassay 820 and 16 from bioassay 523) were used in the modeling. The average IC_{50} reported in the bioassay depositions was used for the bioactivity of the active compounds. The PubChem compound accession (CID), molecular formula, and biological activity (IC_{50}) for the 86 compounds are listed in Table 1 (one active compound was not included in QSAR modeling as no 3D conformer was obtained in docking simulation). Molecular structure and other information can be obtained from PubChem website based on CIDs (http://pubchem.ncbi.nlm.nih.gov) and the bioassay protocol information can be obtained from PubChem website by BioAssay ID (AID) 820 and 523 (http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=820 and http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=523) [32]. The compounds tested in the confirmatory assays were cherry picked based on the screening results for over 60,000 compounds in a single dose HTS assay for potential cathepsin B inhibition activity (PubChem AID: 453) by measuring the release of the fluorophore aminomethyl coumarin (AMC) from the hydrolysis of an AMC-labeled dipeptide. The large number of inactive compounds from the primary HTS assay was not included in the model since the statistical methods typically do not perform well on unbalanced data sets. Therefore claiming all compounds as inactive will yield an accuracy of 99.94% as pointed out by Weis et al. [37].

The three-dimensional structure coordinate of cathepsin protein bound with dipeptidyl nitrile (DPN) was obtained from Protein Databank (PDB code: 1GMY [21]). The PDB file for the crystallographic structure complex of cathepsin B contains three chains A, B and C. The first two chains form a dimer and were used in docking simulation (Fig. 1). Hydrogen atoms and partial charges were added to the protein and then short steps of minimization were performed to relax the newly added hydrogen atoms and the potential steric contacts in the original PDB coordinates. The minimizations were performed using an OPLS force field following the standard protocol of the Glide program (version 8 in FirstDiscovery suite) [38,39]. The minimized protein structure was then used to generate a docking grid which was used to docked all active compounds into the DPN binding site. A docking box (exterior box) of $16 \times 18 \times 22$ Å for the placement of all ligand atoms and a restraining box (interior box) of $8 \times 10 \times 14$ Å for the placement of ligand geometric center were used to generate the docking grid. The boxes were centered at the geometric center of DPN ligand. Default values were used for all other settings within Glide (version 8).

2.2. Conformation determination and molecule alignment

All active compounds were flexibly docked into the active site of the cathepsin protein based on the generated docking grid. All docking simulations were carried out using Glide program. Details of the docking simulation can be obtained from previously reported work [30,31,40]. Docking simulation was used to determine the "active" conformation of each compound according to the structural and physicochemical properties of the binding site. The 3D conformation for each inactive compound was modeled by flexibly aligning it to the 3D conformer of DPN, the ligand in the crystal complex. The alignment was carried out based on molecular shape (volume) and pharmacophore elements using the flexible alignment module implemented in MOE, default parameters were used for the flexible alignment. The conformation for each active compound was also obtained using the same alignment for the comparison with the conformation obtained from docking simulations in QSAR modeling.

2.3. Statistical analysis

Partial least-squares (PLS) analysis method was used to conduct statistical analysis and to derive a continuous QSAR model based on the descriptors calculated for each inhibitor. The PLS method has been proved to be a priority method for such statistical analysis especially in the case where the number of descriptors is large. All PLS calculations were carried out with the module implemented in MOE. The $\log_{10} IC_{50}$ value was used as a dependent variable and the calculated molecular and physicochemical properties were used as independent variables for the analysis. Over 50 descriptors were calculated and preliminary factor analysis was conducted to obtain the optimum correlation between each individual descriptor and the biological activity ($\log IC_{50}$). The nine descriptors listed in Table 2 were selected and used to build QSAR models. An optimal number of components (9 components for continuous models and 8 for binary models) were used to build final models.

Model validation was carried out by the Leave-one-out (LOO) cross-validation procedure and training-test set approach in which the compounds were split into training sets for model building and test sets for which activities were predicted by the built model. Both continuous and binary models were validated by LOO and

three training-test experiments in which nine compounds were randomly selected as test compounds and all other compounds were used as training set for QSAR model building. The LOO technique provides a good way to quantitatively evaluate the predictive ability and robustness of a model by predicting each compound's activity using a QSAR model built based on information of the remaining compounds, which avoids the effect of a compound on its own activity prediction. This approach tries to eliminate the over-fitting problem normally existing in conventional regression method which includes all compounds in model construction. A cross-validated correlation coefficient (q^2) was used to measure the predictability of a model and the conventional correlation coefficient (r^2) was used to measure the quality of a model.

3. Results and discussion

3.1. Conformation determination and alignment construction

In order to calculate the physicochemical and molecular property-based descriptors to build QSAR models, 3D confirmation for each studied compound was needed. The 3D structures of these compounds were first converted from their 2D structures. As stated in Section 2, totally 36 active compounds and 51 inactive compounds with unique structures were obtained from two confirmatory bioassay data entries (PubChem AID 820 and 523). Docking simulations and molecular alignment methods were applied to obtain active 3D confirmers for the 36 active compounds. In the docking simulations, the active compounds were docked into the active site of cathepsin B using the same methodology and confirmation determination procedure as described previously [30]. Reliable docked poses were obtained for all but one compound. Detail results and discussion on conformer selection and comparison for most compounds can be found in previous report [30]. As a result, the 35 compounds with docking results were included in the QSAR modeling. The docked cluster of the active compounds is depicted in Fig. 2. It is seen that all these compounds were docked in the binding site with comparable binding conformers and roughly the similar locations.

In the molecular alignment, all 86 (active and inactive) compounds were superposed onto DPN using flexible molecular alignment method. The best fitted conformation of each compound was selected. The fitting was scored based on molecular shape (occupied volume) and pharmacophore matching which includes hydrogen donor, hydrogen acceptor, hydrophobic atoms, polar hydrogen atoms, and aromaticity of the two molecules. The 3D conformation for each compound was selected based on the best fit from MOE. Two separate sets of molecular structures were prepared, one based on docked structures for active compounds and molecular alignment for inactive compounds and the other based on the molecular alignment for all compounds. The two sets of 3D structures were used to build QSAR models for comparison and they yielded the very similar results. The information, including PubChem compound accession, e.g. CID, molecular formula, SMILES (Simplified Molecular Input Line Entry Specification) string, molecular weight (MW), and the activity (IC_{50}), as obtained from PubChem BioAssay database (AID: 820 and 523) of all 86 compounds used in the work are listed in Table 1. To build a binary QSAR model, the activities of the active and the inactive compounds were assigned as 1 and 0, respectively.

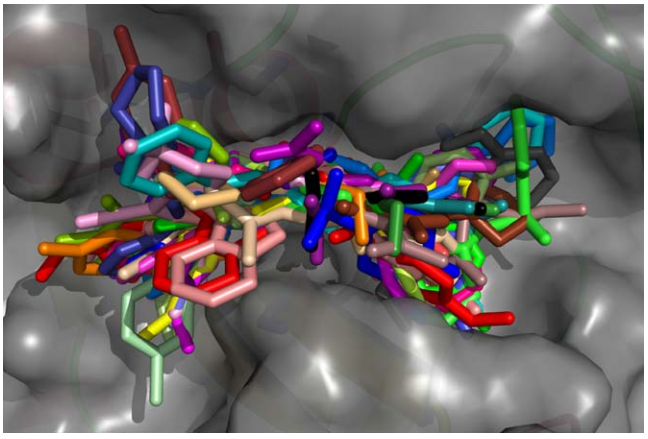

Fig. 2. The cluster of docked poses of the active compounds in the protein binding site.

3.2. Descriptor calculation

Over 250 descriptors are available with MOE QSAR module. Initial selection for the descriptors was attempted empirically based on the nature of the descriptors, the features of the studied molecular set, and the problem studied. Those which are not relevant to the modeling, such as the total energy, heat of formation, the numbers of bromine, boron, fluorine, or phosphorus atoms, were eliminated from further consideration in the work.

Table 3
Continuous all model: the predicted activities and errors (difference between predicted and observed activity ($\log IC_{50}$)) of 42 compounds predicted by two statistic methods based on the QSAR model.

Compnd #	$\log IC_{50}$	Non-validated		Cross-validated	
		Predicted	Error	Predicted	Error
1	−4.94	−5.13	0.19	−5.15	0.20
2	−4.47	−5.16	0.69	−5.26	0.79
3	−5.70	−5.57	−0.13	−5.55	−0.15
4	−4.91	−5.49	0.57	−5.62	0.70
5	−5.90	−5.60	−0.30	−5.54	−0.35
6	−5.20	−4.77	−0.43	−4.68	−0.52
7	−5.76	−5.55	−0.21	−5.53	−0.22
8	−4.35	−4.90	0.55	−5.00	0.65
9	−6.04	−5.45	−0.59	−5.38	−0.66
10	−5.67	−5.53	−0.15	−5.48	−0.20
11	−5.02	−5.50	0.48	−5.55	0.53
12	−5.17	−4.76	−0.41	−4.67	−0.51
13	−4.71	−5.06	0.35	−5.19	0.49
14	−5.05	−4.68	−0.37	−4.62	−0.43
15	−4.40	−4.61	0.21	−4.65	0.25
16	−4.42	−4.97	0.56	−5.16	0.74
17	−5.68	−5.03	−0.65	−4.03	−1.65
18	−4.85	−5.31	0.46	−5.50	0.65
19	−5.38	−5.10	−0.28	−4.95	−0.43
20	−4.43	−5.02	0.59	−5.10	0.67
21	−4.34	−5.03	0.70	−5.12	0.78
22	−5.15	−5.11	−0.04	−5.10	−0.05
23	−5.03	−5.28	0.26	−5.61	0.58
24	−5.93	−5.04	−0.88	−4.80	−1.12
25	−6.16	−5.56	−0.60	−5.49	−0.67
26	−6.36	−5.51	−0.85	−5.33	−1.03
27	−6.07	−5.37	−0.70	−5.03	−1.04
28	−6.61	−5.85	−0.75	−5.65	−0.95
29	−5.07	−5.35	0.28	−5.38	0.31
30	−4.74	−4.97	0.23	−5.23	0.50
31	−5.65	−5.00	−0.65	−4.88	−0.77
32	−5.55	−5.52	−0.03	−5.52	−0.03
33	−5.49	−5.72	0.23	−5.99	0.50
34	−4.65	−5.59	0.94	−5.80	1.15
35	−4.48	−5.27	0.79	−5.36	0.88
37	−3.00	−3.06	0.06	−3.10	0.10
39	−3.00	−3.02	0.02	−3.03	0.03
43	−3.00	−3.03	0.03	−3.04	0.04
50	−3.00	−3.22	0.22	−3.28	0.28
51	−3.00	−3.20	0.20	−3.33	0.33
52	−3.00	−2.59	−0.41	−2.41	−0.59
53	−3.00	−2.83	−0.17	−2.68	−0.32

Over 50 descriptors of physicochemical molecular properties were selected and calculated for all studied compounds based on the constructed 3D conformations using the methods implemented in MOE (version 2007.09). Factor analyses were performed to examine the correlation between each individual descriptor and the biological activity ($\log IC_{50}$) of these compounds. The nine descriptors that demonstrated an apparent correlation with the observed biological activity were chosen to build QSAR models, which are listed in Table 2. Among these descriptors, $N_{acc+don}$ is the sum of the counts of hydrogen acceptor and donor. Apol, MR, and Dipole are the sum of atomic polarizabilities [41], molar refractivity, and the dipole moment of a molecule calculated from the partial charges of the molecule, respectively. SA and SA_{pol} are the water accessible surface areas (WASA) of whole molecule and the hydrophilic part of WASA calculated using a radius of 1.4 Å for the probe, respectively. Vol, E_{sol}, and $\log P_{(o/w)}$ are the molecular volume, the empirical solvation energy calculated based on OPLS force field, and the log value of octanol/water partition coefficient of a compound, respectively. Most of these descriptors are related to the molecular solvation properties to some degree. Calculation of molecular solvation or partition of a compound between hydrophobic and hydrophilic compartments still remains challenging. Combination of various computational methods would be a better choice to estimate the chemical's partition property which is well regarded to have direct and large effect on its biological activities.

3.3. Continuous QSAR models

Conventional QSAR models were constructed for the 35 active compounds based on the measured biological activities ($\log IC_{50}$) using the PLS regression method. A model for the entire compound set was first built to examine model coherence of all compounds and to identify the potential outliers. The model was then validated using the LOO method and three training-test sets to examine the predictability and robustness of the QSAR models. To extend the prediction spectrum to include inactive compounds, seven inactive compounds were randomly chosen and included in the models. The inhibitory activity (reading) of these compounds at the max tested concentration (50 μM) was close to the reading of the control experiments. It was assumed that their IC_{50} (if exists) was much higher than 50 μM. To enable the numerical analysis, a value of −3.00 was assigned as the IC_{50} of these five inactive compounds in the QSAR analysis.

The predicted results and prediction errors for the model based on all 42 compounds (all model) are listed in Table 3. Two thresholds, one and half units, respectively were used to validate the calculated results. The former corresponds to a ±10-fold variation ranges or with a range of a single order of magnitude on each direction from the experimental value of IC_{50} and the latter corresponds to ±3.33-fold variation ranges or within a window of a single order of magnitude on IC_{50}. The non-validated model based on 9 descriptors (properties) showed strong ability to predict the biological activities with errors (difference between the predicted value and experimental value) of smaller than one unit for all 42 compounds. Twenty-six compounds (62%) have predicted errors smaller than or equal to 0.5 unit by the model. Model evaluation using the LOO validation method generated satisfactory results. The observed activities *vs.* the predicted activities by the non-validated fitting model and cross-validated model are plotted in Fig. 3A. By further checking the predicted results of the cross-validation model, it was noted that the three compounds (compounds **24**, **26** and **34**) had predicted errors close to one unit. When these three compounds were treated as outliers and excluded from the model construction, the

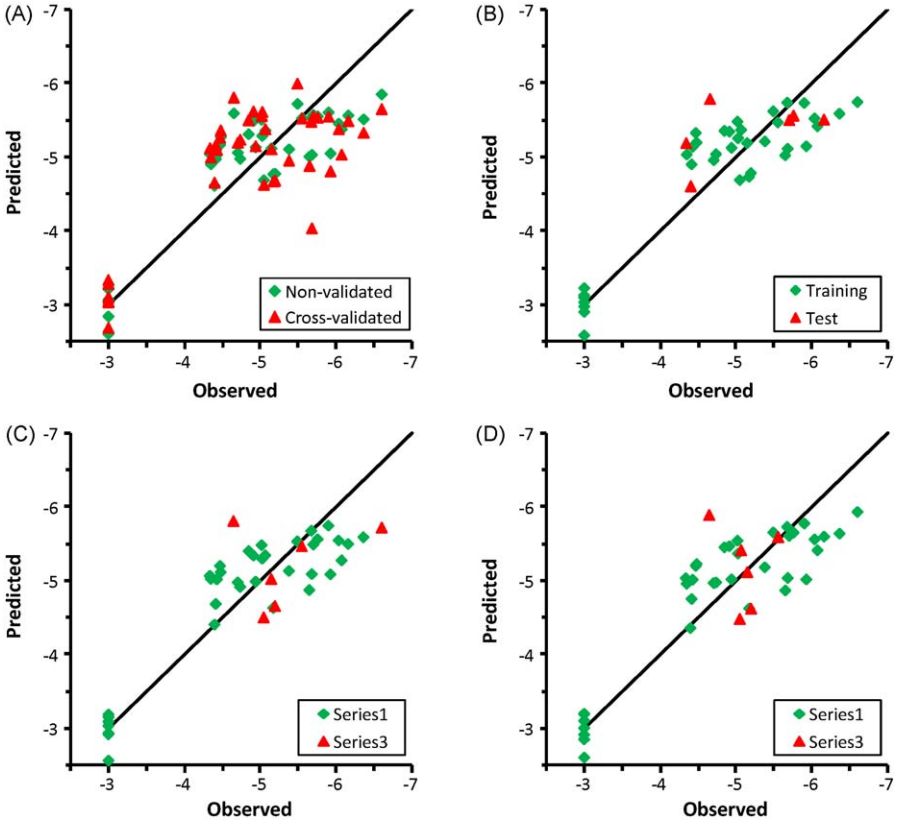

Fig. 3. The plots of the predicted *vs.* observed bioactivity ($\log IC_{50}$) for the continuous models. (A) The continuous all model. (B) The training-test model 1. (C) The training-test model 2. (D) The training-test model 3. For (B–D): the predicted activity for the 6 compounds in test set (red triangles) based on the model built on the 36 compounds in the training set (green squares). The lines are unit slops for perfect predictions (predicted = observed).

Table 4
3 Outlier model: the predicted activities and residues of 37 compounds predicted by two statistic methods based on the QSAR model excluding three outliers (compounds **24**, **26**, and **34**) from the all model.

Compd. #	log IC_{50}	Non-validated		Cross-validated	
		Predicted	Error	Predicted	Error
1	−4.94	−5.11	0.17	−5.13	0.18
2	−4.47	−5.10	0.63	−5.21	0.73
3	−5.70	−5.56	−0.14	−5.54	−0.16
4	−4.91	−5.55	0.64	−5.70	0.79
5	−5.90	−5.47	−0.43	−5.35	−0.55
6	−5.20	−4.55	−0.65	−4.39	−0.81
7	−5.76	−5.55	−0.21	−5.53	−0.23
8	−4.35	−4.77	0.41	−4.86	0.51
9	−6.04	−5.40	−0.63	−5.30	−0.73
10	−5.67	−5.63	−0.04	−5.61	−0.06
11	−5.02	−5.50	0.48	−5.56	0.54
12	−5.17	−4.72	−0.45	−4.60	−0.57
13	−4.71	−4.90	0.20	−4.99	0.28
14	−5.05	−4.64	−0.41	−4.57	−0.48
15	−4.40	−4.54	0.14	−4.57	0.18
16	−4.42	−4.90	0.49	−5.08	0.67
17	−5.68	−5.10	−0.58	−4.17	−1.51
18	−4.85	−5.32	0.48	−5.54	0.69
19	−5.38	−5.19	−0.19	−5.08	−0.30
20	−4.43	−4.99	0.56	−5.06	0.63
21	−4.34	−4.98	0.64	−5.06	0.72
22	−5.15	−5.14	−0.01	−5.14	−0.01
23	−5.03	−5.19	0.16	−5.40	0.38
24	−5.93				
25	−6.16	−5.55	−0.61	−5.47	−0.69
26	−6.36				
27	−6.07	−5.55	−0.52	−5.28	−0.79
28	−6.61	−5.96	−0.64	−5.78	−0.82
29	−5.07	−5.30	0.23	−5.32	0.26
30	−4.74	−4.75	0.02	−4.77	0.04
31	−5.65	−4.82	−0.83	−4.64	−1.01
32	−5.55	−5.54	−0.01	−5.54	−0.01
33	−5.49	−5.80	0.31	−6.19	0.70
34	−4.65				
35	−4.48	−5.28	0.80	−5.37	0.89
37	−3.00	−3.14	0.14	−3.24	0.24
39	−3.00	−2.91	−0.09	−2.88	−0.12
43	−3.00	−3.05	0.05	−3.08	0.08
50	−3.00	−3.24	0.24	−3.31	0.31
51	−3.00	−3.01	0.01	−3.01	0.01
52	−3.00	−2.69	−0.31	−2.54	−0.46
53	−3.00	−2.97	−0.03	−2.95	−0.05

Table 5
The training-test model 1: the predicted activities and residues of 36 compounds in the training set and 6 compounds in the test set.

Compd. #	log IC_{50}	Non-validated		Cross-validated	
		Predicted	Error	Predicted	Error
	Training set				
1	−4.94	−5.13	0.19	−5.15	0.21
2	−4.47	−5.33	0.86	−5.48	1.01
4	−4.91	−5.35	0.44	−5.47	0.56
5	−5.90	−5.73	−0.16	−5.69	−0.21
6	−5.20	−4.78	−0.41	−4.69	−0.51
8	−4.35	−5.04	0.69	−5.18	0.83
9	−6.04	−5.52	−0.51	−5.44	−0.59
10	−5.67	−5.74	0.07	−5.77	0.09
11	−5.02	−5.48	0.46	−5.55	0.53
12	−5.17	−4.73	−0.44	−4.61	−0.56
13	−4.71	−4.96	0.25	−5.07	0.36
14	−5.05	−4.69	−0.36	−4.61	−0.44
16	−4.42	−4.90	0.49	−5.09	0.68
17	−5.68	−5.12	−0.57	−4.15	−1.53
18	−4.85	−5.36	0.51	−5.59	0.74
19	−5.38	−5.21	−0.17	−5.12	−0.26
20	−4.43	−5.14	0.71	−5.24	0.81
22	−5.15	−5.19	0.04	−5.20	0.05
23	−5.03	−5.26	0.23	−5.58	0.56
24	−5.93	−5.15	−0.78	−4.93	−1.00
26	−6.36	−5.59	−0.77	−5.37	−0.99
27	−6.07	−5.42	−0.65	−5.09	−0.99
28	−6.61	−5.75	−0.85	−5.46	−1.14
29	−5.07	−5.37	0.30	−5.40	0.33
30	−4.74	−5.04	0.30	−5.40	0.66
31	−5.65	−5.03	−0.62	−4.90	−0.75
32	−5.55	−5.47	−0.08	−5.46	−0.09
33	−5.49	−5.62	0.13	−5.79	0.30
35	−4.48	−5.20	0.72	−5.30	0.82
37	−3.00	−3.13	0.13	−3.23	0.23
39	−3.00	−2.98	−0.02	−2.97	−0.03
43	−3.00	−2.91	−0.09	−2.84	−0.16
50	−3.00	−3.23	0.23	−3.30	0.30
51	−3.00	−3.10	0.10	−3.18	0.18
52	−3.00	−2.59	−0.41	−2.40	−0.60
53	−3.00	−3.04	0.04	−3.07	0.07
	Test set				
3	−5.70	−5.49	−0.21		
7	−5.76	−5.55	−0.21		
15	−4.40	−4.60	0.20		
21	−4.34	−5.18	0.84		
25	−6.16	−5.50	−0.66		
34	−4.65	−5.78	1.13		

new model (3 outlier model) consisting of the remaining 39 compounds demonstrated an improved quality. The predicted activities for these 39 compounds are listed in Table 4. All compounds have predicted errors smaller than one unit in the non-validated model and 37 compounds have predicted errors smaller than on unit in the cross-validated model. Twenty-seven and more than half compounds have predicted errors smaller than half unit according to the non-validated and cross-validated results, respectively. These cross-validation results demonstrated the strong predictability of the constructed models for the application of cathepsin B inhibition activity prediction.

To further evaluate the model's predictability and robustness, three additional models were built using the training-test-set method, in which the compounds were randomly split into a training set and test set. In this work, six compounds were randomly selected for the test set and this process was repeated three times. The predicted results for the three models are listed in Tables 5–7. The plots of the observed activities *vs.* the predicted activities for the training compounds and testing compounds are shown in Figs. 3B–3D. It is seen that the predicted results for the testing compounds are close to their experimental activities with reasonable deviations. All three models based on the training set of compounds demonstrated similar predictability results.

As a summary, the non-validated regression coefficient, cross-validated coefficient, and root mean square error (RMSE) for the all model are 0.771, 0.608, and 0.483, respectively (Table 8). The non-validated, cross-validated coefficients, and RMSE for the 3 outlier model are 0.821, 0.683, and 0.424, respectively (Table 8). The statistics results for the three training-test models are also listed in Table 8. All three models based on the training set of compounds have similar statistical properties as the all model described above, with non-validated correlation coefficients of 0.804–0.807 and cross-validated coefficients of 0.637–0.675. The predicted errors for all test compounds but one (compound **34**) in the three training sets are smaller than one unit. The Models 1 and 3 (Tables 5 and 7) have two test compounds predicted with errors larger than half unit and the Model 2 has the most compounds predicted with errors larger than half unit. These results show that the predictability of these models is acceptable. Considering the fact that the compound structures and bioactivities studied are reasonably diverse, it is reasonable to conclude that the models demonstrated relatively strong prediction power for estimating the activity of the "unknown" compounds based on the 'known' compounds.

Table 6
The training-test model 2: the predicted activities and residues of 36 compounds in the training set and 6 compounds in the test set.

Compd. #	log IC_{50}	Non-validated		Cross-validated	
		Predicted	Error	Predicted	Error
	Training set				
1	−4.94	−4.99	0.05	−5.00	0.06
2	−4.47	−5.21	0.74	−5.34	0.87
3	−5.70	−5.49	−0.21	−5.46	−0.24
4	−4.91	−5.35	0.44	−5.48	0.57
5	−5.90	−5.75	−0.15	−5.72	−0.18
7	−5.76	−5.57	−0.19	−5.55	−0.21
8	−4.35	−5.03	0.68	−5.17	0.82
9	−6.04	−5.55	−0.48	−5.48	−0.55
10	−5.67	−5.68	0.01	−5.69	0.01
11	−5.02	−5.49	0.47	−5.55	0.53
12	−5.17	−4.64	−0.54	−4.49	−0.68
13	−4.71	−4.98	0.28	−5.11	0.41
15	−4.40	−4.41	0.01	−4.42	0.02
16	−4.42	−4.69	0.28	−4.82	0.40
17	−5.68	−5.10	−0.58	−4.08	−1.60
18	−4.85	−5.41	0.56	−5.66	0.81
19	−5.38	−5.14	−0.24	−5.00	−0.38
20	−4.43	−5.02	0.59	−5.10	0.67
21	−4.34	−5.07	0.74	−5.17	0.83
23	−5.03	−5.30	0.28	−5.69	0.66
24	−5.93	−5.09	−0.83	−4.81	−1.12
25	−6.16	−5.51	−0.65	−5.41	−0.75
26	−6.36	−5.60	−0.77	−5.40	−0.96
27	−6.07	−5.28	−0.79	−4.74	−1.33
29	−5.07	−5.35	0.28	−5.38	0.31
30	−4.74	−4.92	0.19	−5.16	0.42
31	−5.65	−4.88	−0.77	−4.70	−0.95
33	−5.49	−5.54	0.05	−5.60	0.11
35	−4.48	−5.12	0.64	−5.23	0.75
37	−3.00	−3.19	0.19	−3.34	0.34
39	−3.00	−2.92	−0.08	−2.90	−0.10
43	−3.00	−3.04	0.04	−3.07	0.07
50	−3.00	−3.16	0.16	−3.21	0.21
51	−3.00	−3.10	0.10	−3.17	0.17
52	−3.00	−2.57	−0.43	−2.36	−0.64
53	−3.00	−2.94	−0.06	−2.88	−0.12
	Test set				
6	−5.20	−4.65	−0.55		
14	−5.05	−4.50	−0.55		
28	−6.61	−5.72	−0.89		
22	−5.15	−5.02	−0.13		
32	−5.55	−5.47	−0.08		
34	−4.65	−5.81	1.16		

Table 7
The training-test model **3**: the predicted activities and residues of 36 compounds in the training set and 6 compounds in the test set.

Compd. #	log IC_{50}	Non-validated		Cross-validated	
		Predicted	Error	Predicted	Error
	Training set				
1	−4.94	−5.02	0.08	−5.03	0.09
2	−4.47	−5.21	0.74	−5.34	0.87
3	−5.70	−5.61	−0.09	−5.60	−0.10
4	−4.91	−5.47	0.56	−5.62	0.70
5	−5.90	−5.78	−0.12	−5.75	−0.15
7	−5.76	−5.66	−0.10	−5.65	−0.11
8	−4.35	−4.96	0.61	−5.08	0.73
9	−6.04	−5.57	−0.47	−5.50	−0.54
10	−5.67	−5.74	0.06	−5.76	0.09
11	−5.02	−5.55	0.53	−5.61	0.59
12	−5.17	−4.63	−0.55	−4.48	−0.69
13	−4.71	−4.97	0.27	−5.10	0.40
15	−4.40	−4.36	−0.04	−4.35	−0.05
16	−4.42	−4.76	0.34	−4.91	0.49
17	−5.68	−5.04	−0.64	−3.92	−1.76
18	−4.85	−5.46	0.61	−5.76	0.91
19	−5.38	−5.19	−0.19	−5.08	−0.30
20	−4.43	−5.02	0.59	−5.10	0.67
21	−4.34	−5.04	0.70	−5.13	0.79
23	−5.03	−5.37	0.35	−5.84	0.82
24	−5.93	−5.02	−0.90	−4.73	−1.20
25	−6.16	−5.60	−0.56	−5.53	−0.63
26	−6.36	−5.64	−0.72	−5.46	−0.90
27	−6.07	−5.42	−0.66	−5.02	−1.05
28	−6.61	−5.94	−0.66	−5.73	−0.87
30	−4.74	−4.98	0.24	−5.29	0.55
31	−5.65	−4.87	−0.78	−4.69	−0.96
33	−5.49	−5.66	0.17	−5.87	0.38
35	−4.48	−5.23	0.75	−5.34	0.86
37	−3.00	−3.19	0.19	−3.34	0.34
39	−3.00	−2.91	−0.09	−2.88	−0.12
43	−3.00	−3.10	0.10	−3.18	0.18
50	−3.00	−3.20	0.20	−3.26	0.26
51	−3.00	−3.00	0.00	−3.00	0.00
52	−3.00	−2.60	−0.40	−2.41	−0.59
53	−3.00	−2.85	−0.15	−2.71	−0.29
	Test set				
6	−5.20	−4.62	−0.58		
14	−5.05	−4.48	−0.57		
22	−5.15	−5.11	−0.04		
29	−5.07	−5.41	0.34		
32	−5.55	−5.58	0.03		
34	−4.65	−5.88	1.23		

3.4. Binary QSAR models to classify active and inactive compounds

The availability of confirmed active and inactive compounds provides an opportunity to develop binary models for classifying active cathepsin B inhibitors from inactive compounds. Although modern HTS technology allows rapid screening of tens of thousands of compounds in a matter of a few days, such experiments are expensive, and can hardly cover all chemical space. *In silico* screening helps to eliminate potentially inactive compounds and to build compound libraries with enriched promising lead candidates. Thus *in silico* screening is a cost efficient strategy to facilitate the identification of potential drug candidates. Such pre-screening provides a way to eliminate potentially inactive molecules in novel compound design for synthesis and make the limited experimental resources available for screening only the potentially active molecules.

The same nine descriptors used in the continuous models were also used to build binary QSAR models relying on the method developed by the Chemical Computing Group which was implemented as a module in MOE [42,43]. The binary QSAR model was developed as an economic QSAR model to be used distinguish active *vs.* inactive compounds in HTS experiments to assist model drug development. All binary models were built using the binary method module of MOE [42,43]. Similarly to the process of continuous model construction, the "all binary model" was first built with all 86 compounds to examine the predictability of the model for the "inside" compounds, the compounds used to build a model, and followed by validations using the LOO method and three training-test sets to examine the predictability and robustness of the QSAR models for "outside" compounds. The predicted results for the all binary model are listed in Table 9 and the statistical results are listed in Table 13. All 86 compounds, except for four active and three inactive ones, were predicted correctly with this non-validated model (Table 9). The prediction accuracies (Table 13) for all compounds are 0.919, and for active and inactive compounds are 0.886 and 0.941 respectively. This demonstrates the strong classification power of this model for discriminating inactive compounds from active compounds. In addition, four cross-validated models were built, one from LOO and three from the training-test-set method. In the latter three models, the compounds were randomly split into training set (80 compounds) and test set (6 compounds). The four models were designated as

Table 8
Summary of statistics and relative importance of descriptors for three fitting and 3 training-test QSAR models.

Model	All model	3 Outlier model	Training-test 1	Training-test 2	Training-test 3
Training set	42	39	36	36	36
Test set			6	6	6
$r^{2,*}$	0.771	0.821	0.804	0.804	0.807
q^2	0.608	0.683	0.637	0.647	0.675
RMSE	0.483	0.424	0.460	0.459	0.473
Cross-RMSE	0.546	0.574	0.639	0.627	0.629
Relative importance of descriptors					
Apol	1.00	1.00	1.00	1.00	1.00
MR	0.118	0.295	0.196	0.136	0.131
E_{sol}	0.045	0.109	0.022	0.023	0.017
$\log P_{(o/w)}$	0.011	0.054	0.026	0.019	0.012
Dipole	0.093	0.260	0.025	0.018	0.070
SA_{pol}	0.144	0.207	0.111	0.162	0.180
SA	0.222	0.228	0.245	0.245	0.273
Vol	0.608	0.651	0.562	0.586	0.605
$N_{acc-don}$	0.0208	0.092	0.045	0.034	0.022

* r^2: the square of the regression (non-cross-validated) correlation coefficient. q^2: the square of the cross-validated correlation coefficient. Linear model: $\log IC_{50} = a_1 \cdot Apol + a_2 \cdot MR + a_3 \cdot E_{sl} + a_4 \cdot \log P_{(o/w)} + a_5 \cdot Dipole + a_6 \cdot SA_{pol} + a_7 \cdot SA + a_8 \cdot Vol + a_9 \cdot N_{acc-don}$, a_1–a_9 are the linear factors with different values in each model which were optimized in each regression.

"binary model", "bin. train-test 1", "bin. train-test 2", and "bin. train-test 3", and the predicted results from these four cross-validated models are listed in Tables 9–12. The three training-test models produced consistent results that are similar to those from the LOO-based binary model. For the 80 training set compounds, the non-validated prediction accuracies for all compounds are 0.906–9.25, and for active and inactive compounds are 0.906 and 0.938, respectively. The cross-validated prediction accuracies for all compounds are 0.900, active and inactive compounds are 0.844 and 0.938, respectively. For the six test compounds, five compounds were predicted correctly, the prediction accuracies for the randomly selected test compounds in all three training-test models are 0.833 (Table 13). These validated models clearly demonstrated their abilities to distinguish active *vs.* inactive compounds with high efficiency and accuracy, which indicates their strong potential for *in silico* screening of small molecules.

It is noted that the prediction accuracies for active compounds were slightly, but consistently, lower than those for the inactive compounds. Such phenomenon could arise from the unbalanced nature of the data set regarding to bioactivity classes (active and inactive in this work). The larger population of inactive compounds over that of the active ones shifts the splitting point used to determine the active *vs.* inactive to the active side. Such shift results in more active compounds at the border line of being

Table 9
Binary model: the predicted active *vs.* inactive for all active and inactive compounds predicted by two statistic methods based on the binary QSAR model.

Compd. #	Active	Non-validated		Cross-validated	
		Predicted	Error	Predicted	Error
1	1	1	0	1	0
2	1	1	0	1	0
3	1	1	0	1	0
4	1	1	0	1	0
5	1	1	0	1	0
6	1	1	0	1	0
7	1	1	0	1	0
8	1	1	0	1	0
9	1	1	0	1	0
10	1	1	0	0	1
11	1	1	0	1	0
12	1	1	0	1	0
13	1	1	0	1	0
14	1	1	0	1	0
15	1	1	0	1	0
16	1	1	0	1	0
17	1	1	0	1	0
18	1	1	0	1	0
19	1	1	0	1	0
20	1	1	0	1	0
21	1	1	0	1	0
22	1	0	1	0	1
23	1	1	0	1	0
24	1	1	0	1	0
25	1	1	0	1	0
26	1	1	0	1	0
27	1	0	1	0	1
28	1	1	0	1	0
29	1	1	0	1	0
30	1	1	0	1	0
31	1	1	0	1	0
32	1	1	0	1	0
33	1	0	1	0	1
34	1	0	1	0	1
35	1	1	0	1	0
36	0	0	0	0	0
37	0	0	0	0	0
38	0	0	0	0	0
39	0	0	0	0	0
40	0	0	0	0	0
41	0	0	0	0	0
42	0	0	0	0	0
43	0	0	0	0	0
44	0	0	0	0	0
45	0	0	0	0	0
46	0	0	0	0	0
47	0	0	0	0	0
48	0	0	0	0	0
49	0	0	0	0	0
50	0	0	0	0	0
51	0	0	0	0	0
52	0	0	0	0	0
53	0	0	0	0	0
54	0	0	0	0	0
55	0	0	0	0	0
56	0	0	0	0	0
57	0	0	0	0	0
58	0	0	0	0	0
59	0	0	0	0	0
60	0	0	0	0	0
61	0	0	0	0	0
62	0	0	0	0	0
63	0	0	0	0	0
64	0	0	0	0	0
65	0	0	0	0	0
66	0	0	0	0	0
67	0	0	0	0	0
68	0	0	0	0	0
69	0	0	0	0	0
70	0	0	0	0	0
71	0	0	0	0	0
72	0	0	0	0	0
73	0	1	−1	1	−1
74	0	0	0	0	0
75	0	0	0	0	0
76	0	0	0	0	0
77	0	1	−1	1	−1
78	0	1	−1	1	−1
79	0	0	0	0	0
80	0	0	0	0	0
81	0	0	0	0	0
82	0	0	0	0	0
83	0	0	0	0	0
84	0	0	0	0	0
85	0	0	0	0	0
86	0	0	0	0	0

Table 10
Binary training-test model 1: the predicted active *vs.* inactive for training set compounds and the predicted results for the tested compounds.

Compd. #	Active	Non-validated		Cross-validated	
		Predicted	Error	Predicted	Error
	Training set				
1	1	1	0	1	0
2	1	1	0	1	0
4	1	1	0	1	0
5	1	1	0	1	0
6	1	1	0	1	0
7	1	1	0	1	0
8	1	1	0	1	0
9	1	1	0	1	0
10	1	1	0	1	0
11	1	1	0	1	0
12	1	1	0	1	0
14	1	1	0	1	0
15	1	1	0	1	0
16	1	1	0	1	0
17	1	1	0	1	0
18	1	1	0	1	0
19	1	1	0	1	0
20	1	1	0	1	0
21	1	1	0	1	0
22	1	0	1	0	1
23	1	1	0	1	0
24	1	1	0	0	1
25	1	1	0	1	0
26	1	1	0	1	0
27	1	0	1	0	1
28	1	1	0	1	0
30	1	1	0	1	0
31	1	1	0	1	0
32	1	1	0	1	0
33	0	0	1	0	1
34	0	1	0	0	1
35	0	1	0	1	0
36	0	0	0	0	0
37	0	0	0	0	0
38	0	0	0	0	0
39	0	0	0	0	0
40	0	0	0	0	0
42	0	0	0	0	0
43	0	0	0	0	0
44	0	0	0	0	0
45	0	0	0	0	0
46	0	0	0	0	0
47	0	0	0	0	0
48	0	0	0	0	0
49	0	0	0	0	0
50	0	0	0	0	0
52	0	0	0	0	0
53	0	0	0	0	0
54	0	0	0	0	0
55	0	0	0	0	0
56	0	0	0	0	0
57	0	0	0	0	0
58	0	0	0	0	0
59	0	0	0	0	0
60	0	0	0	0	0
61	0	0	0	0	0
62	0	0	0	0	0
63	0	0	0	0	0
64	0	0	0	0	0
65	0	0	0	0	0
66	0	0	0	0	0
67	0	0	0	0	0
68	0	0	0	0	0
69	0	0	0	0	0
70	0	0	0	0	0
71	0	0	0	0	0
72	0	0	0	0	0
73	0	1	−1	1	−1
74	0	0	0	0	0
75	0	0	0	0	0
76	0	0	0	0	0
77	0	1	−1	1	−1
78	0	1	−1	1	−1
79	0	0	0	0	0
80	0	0	0	0	0
81	0	0	0	0	0
82	0	0	0	0	0
83	0	0	0	0	0
84	0	0	0	0	0
86	0	0	0	0	0
	Test set				
3	1	1	0		
13	1	0	1		
29	1	1	0		
41	0	0	0		
51	0	0	0		
85	0	0	0		

Table 11
Binary training-test model 2: the predicted active *vs.* inactive for training set compounds and the predicted results for the tested compounds.

Compd. #	Active	Non-validated		Cross-validated	
		Predicted	Error	Predicted	Error
	Training set				
1	1	1	0	1	0
2	1	1	0	1	0
3	1	1	0	1	0
5	1	1	0	1	0
6	1	1	0	1	0
7	1	1	0	1	0
8	1	1	0	1	0
9	1	1	0	1	0
10	1	1	0	0	1
11	1	1	0	1	0
12	1	1	0	1	0
13	1	1	0	1	0
14	1	1	0	1	0
15	1	1	0	1	0
16	1	1	0	1	0
17	1	1	0	1	0
19	1	1	0	1	0
20	1	1	0	1	0
21	1	1	0	1	0
22	1	0	1	0	1
23	1	1	0	1	0
24	1	1	0	1	0
25	1	1	0	1	0
26	1	1	0	1	0
27	1	0	1	0	1
28	1	1	0	1	0
29	1	1	0	1	0
30	1	1	0	1	0
32	1	1	0	1	0
33	0	0	1	0	1
34	0	1	0	0	1
35	0	1	0	1	0
36	0	0	0	0	0
37	0	0	0	0	0
38	0	0	0	0	0
39	0	0	0	0	0
40	0	0	0	0	0
41	0	0	0	0	0
42	0	0	0	0	0
43	0	0	0	0	0
44	0	0	0	0	0
45	0	0	0	0	0
46	0	0	0	0	0

Table 11 (*Continued*)

Compd. #	Active	Non-validated		Cross-validated	
		Predicted	Error	Predicted	Error
	Training set				
48	0	0	0	0	0
49	0	0	0	0	0
50	0	0	0	0	0
51	0	0	0	0	0
52	0	0	0	0	0
53	0	0	0	0	0
54	0	0	0	0	0
55	0	0	0	0	0
56	0	0	0	0	0
57	0	0	0	0	0
58	0	0	0	0	0
59	0	0	0	0	0
60	0	0	0	0	0
61	0	0	0	0	0
62	0	0	0	0	0
63	0	0	0	0	0
65	0	0	0	0	0
66	0	0	0	0	0
67	0	0	0	0	0
68	0	0	0	0	0
69	0	0	0	0	0
70	0	0	0	0	0
71	0	0	0	0	0
72	0	0	0	0	0
73	0	1	−1	1	−1
74	0	0	0	0	0
75	0	0	0	0	0
76	0	0	0	0	0
77	0	1	−1	1	−1
78	0	1	−1	1	−1
80	0	0	0	0	0
81	0	0	0	0	0
82	0	0	0	0	0
83	0	0	0	0	0
84	0	0	0	0	0
85	0	0	0	0	0
86	0	0	0	0	0
	Test set				
4	1	1	0		
18	1	0	1		
31	1	1	0		
47	0	0	0		
64	0	0	0		
79	0	0	0		

Table 12
Binary training-test model 3: the predicted active *vs.* inactive for training set compounds and the predicted results for the tested compounds.

Compd. #	Active	Non-validated		Cross-validated	
		Predicted	Error	Predicted	Error
	Training set				
1	1	1	0	1	0
2	1	1	0	1	0
3	1	1	0	1	0
4	1	1	0	1	0
6	1	1	0	1	0
7	1	1	0	1	0
8	1	1	0	1	0
9	1	1	0	1	0
10	1	1	0	0	1
11	1	1	0	1	0
12	1	1	0	1	0
13	1	1	0	1	0
15	1	1	0	1	0
16	1	1	0	1	0
17	1	1	0	1	0
18	1	1	0	1	0

Table 12 (*Continued*)

Compd. #	Active	Non-validated		Cross-validated	
		Predicted	Error	Predicted	Error
	Training set				
19	1	1	0	1	0
20	1	1	0	0	1
21	1	1	0	1	0
23	1	1	0	1	0
24	1	1	0	1	0
25	1	1	0	1	0
26	1	1	0	1	0
27	1	0	1	0	1
28	1	1	0	1	0
29	1	1	0	1	0
30	1	0	0	1	0
31	1	1	0	1	0
32	1	1	0	1	0
33	0	0	1	0	1
34	0	1	1	0	1
35	0	1	0	1	0
36	0	0	0	0	0
37	0	0	0	0	0
38	0	0	0	0	0
39	0	0	0	0	0
40	0	0	0	0	0
41	0	0	0	0	0
42	0	0	0	0	0
43	0	0	0	0	0
44	0	0	0	0	0
45	0	0	0	0	0
46	0	0	0	0	0
47	0	0	0	0	0
48	0	0	0	0	0
49	0	0	0	0	0
50	0	0	0	0	0
51	0	0	0	0	0
52	0	0	0	0	0
53	0	0	0	0	0
54	0	0	0	0	0
56	0	0	0	0	0
57	0	0	0	0	0
58	0	0	0	0	0
59	0	0	0	0	0
60	0	0	0	0	0
61	0	0	0	0	0
62	0	0	0	0	0
63	0	0	0	0	0
64	0	0	0	0	0
65	0	0	0	0	0
66	0	0	0	0	0
68	0	0	0	0	0
69	0	0	0	0	0
70	0	0	0	0	0
71	0	0	0	0	0
72	0	0	0	0	0
73	0	1	−1	1	−1
74	0	0	0	0	0
75	0	0	0	0	0
77	0	1	−1	1	−1
78	0	1	−1	1	−1
79	0	0	0	0	0
80	0	0	0	0	0
81	0	0	0	0	0
82	0	0	0	0	0
83	0	0	0	0	0
84	0	0	0	0	0
85	0	0	0	0	0
86	0	0	0	0	0
	Test set				
5	1	1	0		
14	1	1	0		
22	1	0	1		
55	0	0	0		
67	0	0	0		
76	0	0	0		

Table 13
Summary of statistics and importance of descriptors for the binary model and three training-test binary models.

Model	Binary model	Bin. train-test 1	Bin. train-test 2	Bin. train-test 3
Training set	86	80	80	80
Active	35	32	32	32
Inactive	51	48	48	48
Test set	0	6	6	6
Tested accuracy		1.00	0.833	0.833
Total accuracy	0.919	0.906	0.925	0.925
p-value*	1.28e−13	4.15e−13	4.14e−13	4.14e−13
Accuracy on active	0.886	0.906	0.906	0.906
Accuracy on inactive	0.941	0.938	0.938	0.938
p-value	8.02e−13	2.63e−12	2.63e−12	2.63e−12
Cross-validated results				
X-validated total accuracy	0.907	0.900	0.900	0.900
p-value	8.44e−13	1.86e−11	1.86e−11	1.86e−11
X-validated accuracy on active	0.857	0.844	0.844	0.844
X-validated accuracy on inactive	0.941	0.938	0.938	0.938
p-value	4.33e−12	7.99e−11	7.99e−11	7.99e−11
Relative importance of descriptors				
Apol	0.127	0.127	0.129	0.141
MR	0.088	0.085	0.100	0.102
E_{sol}	0.137	0.157	0.141	0.158
$\log P_{(o/w)}$	0.049	0.067	0.059	0.050
Dipole	0.111	0.131	0.100	0.105
SA_{pol}	0.198	0.186	0.194	0.208
SA	0.162	0.162	0.164	0.171
Vol	0.138	0.145	0.145	0.146
$N_{acc-don}$	0.055	0.072	0.064	0.052

*p-value: significance of statistics. X-validated: cross-validated.

classified as "inactive". To solve the problem, one could design a weighting schema to "overestimate" the type of compounds that are underrepresented in the data set. In many cases, however, such unbalance would not cause a severe effect on the application of the method when the focus is to select a molecule with a given property, such as eliminating inactive molecules to further develop active compounds in designing cathepsin B inhibitors.

4. Conclusion

The availability of molecular structures and their inhibition activity against cathepsin B identified through HTS experiments provided an opportunity to conduct a structure–activity relationship study to facilitate chemical probe development, and potential drug development for related disease treatment. The "active" conformers of active compounds were modeled using docking simulations, and the conformations of the inactive compounds were determined based on flexible alignment of each compound to the 3D conformer of the original ligand (DPN) in the crystal structure complex. Conventional continuous models were built for predicting cathepsin B inhibition activity of active small molecules. Binary models were constructed for the "active" *vs.* "inactive" classification. Nine molecular and physicochemical properties were calculated based on 3D conformational information, and were selected and used as descriptors to build the QSAR models.

The continuous models demonstrated reasonable correlations between the predicted and the observed activities with non-validated (r^2) and cross-validated (q^2) regression correlation coefficients of 0.771 and 0.608 for all compounds, and 0.821 and 0.683 for the compounds excluding 3 outliers, respectively. The model showed strong predictability with reasonable predicted error rates over the entire test set and across several models. Almost all the 42 compounds used in the model have predicted errors smaller than one unit, while the majority of the compounds have predicted errors less than a half unit from both the non-validated and cross-validated results. The predictability of the models were further validated using the three training-test methods with the results that the predicted errors for all test compounds are smaller than one unit, while a significant fraction of them are less than a half unit in all three models together.

The consistent performance of the four cross-validated models demonstrated that the models are robust in predicting the inhibition activities of small molecules for the biological system of cathepsin B. The success of the QSAR models proves that the modeling approach based on docking conformation determination and the nine physicochemical properties may be a promising method to predict biological activity of a small molecule for drug development based on the knowledge derived from the identified cathepsin B inhibitors.

Furthermore, binary models were built for classifying cathepsin B inhibition activities, e.g. active *vs.* inactive categories. The statistical results showed that the models have demonstrated excellent performance with the prediction accuracies for bioactivity classes of 0.919, 0.886, and 0.941 for all 86 tested, 35 active, and 51 inactive compounds in the non-validated binary models, respectively. All compounds, except for four active and three inactive compounds, were correctly predicted by the non-validated fitting model. Similar prediction results were obtained with the cross-validated model. With the three training-test models, the bioactivity categories for the five out of six tested compounds were predicted correctly, with a prediction accuracy of 0.833. The factor that all of these models yielded high accuracies upon the prediction of cathepsin B inhibition activity demonstrates that such models enable the classification of active *vs.* inactive inhibitory compounds with high accuracies. The robustness and feasibility of this binary method indicates its potential in several applications to facilitate molecular design and drug development. In particular, this approach may prove highly efficient for pre-screening *in silico* designed compounds to optimize the biological activities based on the information of known bioactive compounds. Such *in silico* pre-screening may be performed to filter out the ones that are unlikely to have the desired biological activities, thus to allow the limited efforts to be focused on the highly promising candidates. With such advantages, we anticipate the

utilization of the combination of the continuous and binary classification models will potentially benefit modern drug development for therapeutic purposes.

Acknowledgements

This research was supported by the Intramural Research Program of the NIH and NLM. The authors would like to thank the NIH Fellows Editorial Board (FEB) for reviewing the manuscript and the developers of Pymol software for sharing the program to prepare the molecular figures used in this paper.

References

[1] H.K. Rooprai, D. McCormick, Proteases and their inhibitors in human brain tumours: a review, Anticancer Res. 17 (1997) 4151–4162.
[2] A.J. Barrett, H. Kirschke, B. Cathepsin, H. Cathepsin, L. cathepsin, Methods Enzymol. 80 Pt C (1981) 535–561.
[3] H.A. Chapman Jr., J.S. Munger, G.P. Shi, The role of thiol proteases in tissue injury and remodeling, Am. J. Respir. Crit. Care Med. 150 (1994) S155–159.
[4] L.R. Roberts, P.N. Adjei, G.J. Gores, Cathepsins as effector proteases in hepatocyte apoptosis, Cell Biochem. Biophys. 30 (1999) 71–88.
[5] I. Giusti, S. D'Ascenzo, D. Millimaggi, G. Taraboletti, G. Carta, N. Franceschini, A. Pavan, V. Dolo, Cathepsin B mediates the pH-dependent proinvasive activity of tumor-shed microvesicles, Neoplasia 10 (2008) 481–488.
[6] E. Gounaris, C.H. Tung, C. Restaino, R. Maehr, R. Kohler, J.A. Joyce, H.L. Plough, T.A. Barrett, R. Weissleder, K. Khazaie, Live imaging of cysteine-cathepsin activity reveals dynamics of focal inflammation, angiogenesis, and polyp growth, PLoS ONE 3 (2008) e2916.
[7] S.D. Ha, A. Martins, K. Khazaie, J. Han, B.M. Chan, S.O. Kim, Cathepsin B is involved in the trafficking of TNF-alpha-containing vesicles to the plasma membrane in macrophages, J. Immunol. 181 (2008) 690–697.
[8] A. Haque, N.L. Banik, S.K. Ray, New insights into the roles of endolysosomal cathepsins in the pathogenesis of Alzheimer's disease: cathepsin inhibitors as potential therapeutics, CNS Neurol. Disord. Drug Targets 7 (2008) 270–277.
[9] E. Sandes, C. Lodillinsky, R. Cwirenbaum, C. Arguelles, A. Casabe, A.M. Eijan, Cathepsin B is involved in the apoptosis intrinsic pathway induced by Bacillus Calmette-Guerin in transitional cancer cell lines, Int. J. Mol. Med. 20 (2007) 823–828.
[10] S.P. Lutgens, K.B. Cleutjens, M.J. Daemen, S. Heeneman, Cathepsin cysteine proteases in cardiovascular disease, FASEB J. 21 (2007) 3029–3041.
[11] V. Hook, M. Kindy, G. Hook, Cysteine protease inhibitors effectively reduce in vivo levels of brain beta-amyloid related to Alzheimer's disease, Biol. Chem. 388 (2007) 247–252.
[12] L.S. Downs Jr., P.H. Lima, R.L. Bliss, C.H. Blomquist, Cathepsins B and D activity and activity ratios in normal ovaries, benign ovarian neoplasms, and epithelial ovarian cancer, J. Soc. Gynecol. Investig. 12 (2005) 539–544.
[13] O. Vasiljeva, M. Korovin, M. Gajda, H. Brodoefel, L. Bojic, A. Kruger, U. Schurigt, L. Sevenich, B. Turk, C. Peters, T. Reinheckel, Reduced tumour cell proliferation and delayed development of high-grade mammary carcinomas in cathepsin B-deficient mice, Oncogene 27 (2008) 4191–4199.
[14] O. Vasiljeva, A. Papazoglou, A. Kruger, H. Brodoefel, M. Korovin, J. Deussing, N. Augustin, B.S. Nielsen, K. Almholt, M. Bogyo, C. Peters, T. Reinheckel, Tumor cell-derived and macrophage-derived cathepsin B promotes progression and lung metastasis of mammary cancer, Cancer Res. 66 (2006) 5242–5250.
[15] D.T. Jane, L. Morvay, L. Dasilva, D. Cavallo-Medved, B.F. Sloane, M.J. Dufresne, Cathepsin B localizes to plasma membrane caveolae of differentiating myoblasts and is secreted in an active form at physiological pH, Biol. Chem. 387 (2006) 223–234.
[16] K. Chandran, N.J. Sullivan, U. Felbor, S.P. Whelan, J.M. Cunningham, Endosomal proteolysis of the ebola virus glycoprotein is necessary for infection science, Science 308 (2005) 1643–1645.
[17] G. Simmons, D.N. Gosalia, A.J. Rennekamp, J.D. Reeves, S.L. Diamond, P. Bates, Inhibitors of cathepsin L prevent severe acute respiratory syndrome coronavirus entry, Proc. Natl. Acad. Sci. U.S.A. 102 (2005) 11876–11881.
[18] V.Y. Hook, M. Kindy, G. Hook, Inhibitors of cathepsin B improve memory and reduce beta-amyloid in transgenic Alzheimer disease mice expressing the wild-type, but not the Swedish mutant, beta-secretase site of the amyloid precursor protein, J. Biol. Chem. 283 (2008) 7745–7753.
[19] M. Hosokawa, K. Kashiwaya, H. Eguchi, H. Ohigashi, O. Ishikawa, M. Furihata, Y. Shinomura, K. Imai, Y. Nakamura, H. Nakagawa, Over-expression of cysteine proteinase inhibitor cystatin 6 promotes pancreatic cancer growth, Cancer Sci. 99 (2008) 1626–1632.
[20] B.S. Parker, D.R. Ciocca, B.N. Bidwell, F.E. Gago, M.A. Fanelli, J. George, J.L. Slavin, A. Moller, R. Steel, N. Pouliot, B.L. Eckhardt, M.A. Henderson, R.L. Anderson, Primary tumour expression of the cysteine cathepsin inhibitor Stefin A inhibits distant metastasis in breast cancer, J. Pathol. 214 (2008) 337–346.
[21] P.D. Greenspan, K.L. Clark, R.A. Tommasi, S.D. Cowen, L.W. McQuire, D.L. Farley, J.H. van Duzer, R.L. Goldberg, H. Zhou, Z. Du, J.J. Fitt, D.E. Coppa, Z. Fang, W. Macchia, L. Zhu, M.P. Capparelli, R. Goldstein, A.M. Wigg, J.R. Doughty, R.S. Bohacek, A.K. Knap, Identification of dipeptidyl nitriles as potent and selective inhibitors of cathepsin B through structure-based drug design, J. Med. Chem. 44 (2001) 4524–4534.
[22] H.-H. Otto, T. Schirmeister, Cysteine proteases and their inhibitors, Chem. Rev. 97 (1997) 133–172.
[23] A. Yamamoto, K. Tomoo, T. Hara, M. Murata, K. Kitamura, T. Ishida, Substrate specificity of bovine cathepsin B and its inhibition by CA074, based on crystal structure refinement of the complex, J. Biochem. 127 (2000) 635–643.
[24] M. Mladenovic, K. Ansorg, R.F. Fink, W. Thiel, T. Schirmeister, B. Engels, Atomistic insights into the inhibition of cysteine proteases: first QM/MM calculations clarifying the stereoselectivity of epoxide-based inhibitors, J. Phys. Chem. B 112 (2008) 11798–11808.
[25] I. Redzynia, A. Ljunggren, M. Abrahamson, J.S. Mort, J.C. Krupa, M. Jaskolski, G. Bujacz, Displacement of the occluding loop by the parasite protein, chagasin, results in efficient inhibition of human cathepsin B, J. Biol. Chem. 283 (2008) 22815–22825.
[26] D. Watanabe, A. Yamamoto, K. Tomoo, K. Matsumoto, M. Murata, K. Kitamura, T. Ishida, Quantitative evaluation of each catalytic subsite of cathepsin B for inhibitory activity based on inhibitory activity-binding mode relationship of epoxysuccinyl inhibitors by X-ray crystal structure analyses of complexes, J. Mol. Biol. 362 (2006) 979–993.
[27] P. Markt, C. McGoohan, B. Walker, J. Kirchmair, C. Feldmann, G.D. Martino, G. Spitzer, S. Distinto, D. Schuster, G. Wolber, C. Laggner, T. Langer, Discovery of novel cathepsin s inhibitors by pharmacophore-based virtual high-throughput screening, J. Chem. Inf. Model. 48 (2008) 1693–1705.
[28] M.P. Beavers, M.C. Myers, P.P. Shah, J.E. Purvis, S.L. Diamond, B.S. Cooperman, D.M. Huryn, A.B. Smith 3rd, Molecular docking of cathepsin L inhibitors in the binding site of papain, J. Chem. Inf. Model. 48 (2008) 1464–1472.
[29] P.P. Shah, M.C. Myers, M.P. Beavers, J.E. Purvis, H. Jing, H.J. Grieser, E.R. Sharlow, A.D. Napper, D.M. Huryn, B.S. Cooperman, I. Amos, B. Smith, S.L. Diamond, Kinetic characterization and molecular docking of a novel, potent, and selective slow-binding inhibitor of human cathepsin L, Mol. Pharmacol. 74 (2008) 34–41.
[30] Z. Zhou, Y. Wang, S.H. Bryant, Computational analysis of the cathepsin b inhibitors activities through LR-MMPBSA binding affinity calculation based on docked complex, J. Comput. Chem. 30 (2009) 2165–2175.
[31] Z. Zhou, M. Bates, J.D. Madura, Structure modeling, ligand binding, and binding affinity calculation (LR-MM-PBSA) of human heparanase for inhibition and drug design, Proteins Struct. Function Bioinform. 65 (2006) 580–592.
[32] M.C. Myers, A.D. Napper, N. Motlekar, P.P. Shah, C.-H. Chiu, M.P. Beavers, S.L. Diamond, D.M. Huryn, A.B. Smith III, Identification and characterization of 3-substituted pyrazolyl esters as alternate substrates for cathepsin B: the confounding effects of DTT and cysteine in biological assays, Bioorg. Med. Chem. Lett. 17 (2007) 4761–4766.
[33] W.L. Jorgensen, J. Tirado-Rives, The OPLS potential function for proteins energy minimizations for crystals of cyclic peptides and crambin, J. Am. Chem. Soc. 110 (1988) 1657–1666.
[34] W. Damm, A. Frontera, J. Tirado-Rives, W.L. Jorgensen, OPLS all-atom force field for carbohydrates, J. Comput. Chem. 18 (1997) 1955–1970.
[35] R.C. Rizzo, W.L. Jorgensen, OPLS all-atom model for amines: resolution of the amine hydration problem, J. Am. Chem. Soc. 121 (1999) 4827–4836.
[36] W.L. Jorgensen, D.S. Maxwell, J. Tirado-Rives, Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids, J. Am. Chem. Soc. 118 (1996) 11225–11236.
[37] D.C. Weis, D.P. Visco Jr., J.-L. Faulon, Data mining PubChem using a support vector machine with the Signature molecular descriptor: classification of factor XIa inhibitors, J. Mol. Graph. Model. 27 (2008) 466–475.
[38] R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, M.P. Repasky, E.H. Knoll, M. Shelley, J.K. Perry, D.E. Shaw, P. Francis, P.S. Shenkin, Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy, J. Med. Chem. 47 (2004) 1739–1749.
[39] T.A. Halgren, R.B. Murphy, R.A. Friesner, H.S. Beard, L.L. Frye, W.T. Pollard, J.L. Banks, Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening, J. Med. Chem. 47 (2004) 1750–1759.
[40] Z. Zhou, M. Khaliq, J.-E. Suk, C. Patkar, L. Li, R.J. Kuhn, C.B. Post, Antiviral compounds discovered by virtual screening of small-molecule libraries against dengue virus E protein, ACS Chem. Biol. 3 (2008) 765–775.
[41] D.R. Lide, CRC Handbook of Chemistry and Physics, CRC Press, Boca Raton, FL, 1994.
[42] P. Labute, Binary QSAR: a new method for the determination of quantitative structure activity relationships, Pacific Symp. Biocomput. 4 (1999) 444–455.
[43] H. Gao, C. Williams, P. Labute, J. Bajorath, Binary quantitative structure–activity relationship (QSAR) analysis of estrogen receptor ligands, J. Chem. Inform. Comput. Sci. 39 (1999) 164–168.