Effects of Variable Docking Conditions and Scoring Functions on Corresponding Protein-Aligned Comparative Molecular Field Analysis Models Constructed from Diverse Human Protein Tyrosine Phosphatase 1B Inhibitors

Mutasem O. Taha*,† and Murad A. AlDamen‡

Department of Pharmaceutical Sciences, Faculty of Pharmacy, University of Jordan, Amman, Jordan, and Department of Chemistry, Faculty of Science, University of Jordan, Amman, Jordan

Received August 26, 2005

The effects of variable docking conditions and scoring functions on corresponding protein-aligned comparative molecular field analysis (CoMFA) models have been assessed. To this end, a group of diverse inhibitors were docked into the active site of human protein tyrosine phosphatase 1B (h-PTP 1B). The docked structures were utilized to construct corresponding protein-aligned CoMFA models by employing probe-based (H^+, OH, CH_3) energy grids and genetic partial least squares (G/PLS) statistical analysis. A total of 48 different docking configurations were evaluated, of which some succeeded in producing self-consistent and predictive CoMFA models. However, the best CoMFA model coincided with docking the un-ionized ligands into the hydrated form of the binding site via the PLP1 scoring function and restricted docking settings (r^2(LOO) = 0.647, r^2(PRESS) against 27 test compounds = 0.617). Interestingly, the most significant CoMFA models were orthogonal and corresponded to significantly different docked conformers/poses. To utilize the predictive potentials of the best CoMFA models collectively, it was decided to combine them in a single quantitative structure−activity relationship (QSAR) model. The combination model illustrated excellent statistical properties (r^2(LOO) = 0.890, r^2(PRESS) against 27 test compounds = 0.750).

Introduction

Comparative molecular field analysis (CoMFA) is a three-dimensional quantitative structure−activity relationship (3D QSAR) approach superior to most other QSAR methods with regard to their predictive capabilities.[1] It started and became widely used with the contributions of R. D. Cramer III et al.[2,3] The idea underlying this technique is that differences in the biological properties of compounds can often be explained by differences in the noncovalent fields surrounding the molecules. CoMFA is based on the Lennard-Jones steric and the Coulombic electrostatic field values computed at the intersections of a lattice within a 3D region surrounding the bioactive molecules. Thus, each CoMFA descriptor is represented by steric or electrostatic field values at a certain grid point. These descriptors serve as independent variables in QSAR analysis.[4−9] The first step of a CoMFA procedure is to obtain the compounds' active conformations responsible for the bioactivity and to align these conformations in space, in accordance with a postulated pharmacophore model, docking results, or crystallographic data, etc.

Structure-based alignment deals with a set of ligands that are superimposed onto a reference molecule. It is usually the most rigid one or the one possessing high affinity to the receptor according to a certain proposed pharmacophore model. Furthermore, the reference molecule can be a certain high-affinity ligand docked within the binding pocket of the target macromolecule. In some cases, especially if all compounds are flexible, this approach can lead to doubtful results because the chosen conformations remain far from the biologically active ones. For example, it was found that the atom-based alignment that yielded the statistically best CoMFA model for certain MMP-3 inhibitors is inconsistent with the crystal structure of the bound conformation of the optimal inhibitor.[10]

On the other hand, protein (or receptor)-based alignment involves a set of ligands docked to the active site of the protein before superimposing them to each other according to their relative positions in the active site. This way optimizes the choice of the biologically active conformations and poses. Several articles dealing with docking−CoMFA combinations were published recently.[10−22]

However, molecular docking, which is basically a conformational sampling procedure in which various docked conformations are explored to identify the right one, can be a very challenging problem given the degree of conformational flexibility at the ligand−macromolecular level.[23−28] Almost all current docking programs perform flexible ligand docking, while they treat the receptor as a rigid entity. The conformational sampling methods that these programs are based upon vary considerably.[29−34] Still, regardless of the applied search technique, conformational sampling must be guided by a scoring function that is used to evaluate the quality of the fit between the protein and the ligand. The final docked conformations are selected according to their scores.[35−37] Current scoring functions can be roughly grouped into three categories: force field methods,[30−32]

* Corresponding author. Telephone: 00962 6 5355000, ext. 2505. Fax: 00962 6 5339649. E-mail: mutasem@ju.edu.jo.
† Faculty of Pharmacy.
‡ Faculty of Science.

10.1021/jm058047o CCC: $30.25
Published on Web 11/15/2005

empirical scoring functions[29,38−44] and knowledge-based potentials.[45−49]

Despite that modern docking methods are able to calculate fairly accurately the position and orientation of a potential ligand within a receptor binding site,[36,37] their major problem is the inability of the scoring functions to evaluate binding free energies correctly to rank different potential ligand−receptor complexes. The main problem in affinity prediction is that the underlying molecular interactions are highly complex and various terms should be taken into account to quantify the free energy of the interaction process.[50−53] Accordingly, the molecular modeler must find the optimal combination of docking/scoring algorithms to predict the correct conformer/pose of a potential ligand docked within a certain binding pocket.

In addition to deciding the optimal docking/scoring combination for a particular docking problem, the molecular modeler must decide whether to leave crystallographically explicit water molecules in the binding site prior to ligand docking.[54−59] Furthermore, the fact that crystallographic structures lack information on hydrogen atoms means that it should be appropriately assumed whether the ligand's ionizable moieties embedded within the binding site exist in their ionized form prior to docking.[54,56] The pK_a values of various groups embedded within a receptor depend on the respective local microenvironment. For example, if a carboxylic acid group is in a nonpolar local environment, its pK_a will be raised because the anionic form is destabilized.[56]

The recent interest in employing protein-based alignment techniques in CoMFA studies prompted us to evaluate the effects of different docking approaches and scoring functions on the corresponding protein-based CoMFA models. Furthermore, we were encouraged to evaluate the effects of ligand-related ionization assumptions, as well as the presence or absence of crystallographically explicit water molecules within the binding site, on the qualities of the corresponding protein-based CoMFA models.

The current study was conducted by docking 137 diverse human protein tyrosine phosphatase (h-PTP 1B) inhibitors (Table 1 and Figure 1) into a selected crystallographic structure of this interesting diabetes-related target.[60−62] We decided to conduct the docking part utilizing the recently introduced software LigandFit, which utilizes Monte Carlo simulations for conformational sampling.[34,63] This docking engine was recently reported to illustrate good overall performance, particularly in virtual high-throughput screening experiments.[28,36,37,64] The software was instructed to select a maximum of 10 distinct optimal conformers/poses for each docked inhibitor (i.e., of docking energy ≤20 kcal/mol). Subsequently, six scoring functions representing the three major scoring categories (i.e., PLP1,[39] PLP2,[40] LigScore1,[34,63,65] LigScore2,[34,63,65] PMF,[45−47] and LUDI[41,42]) were separately employed to rank the optimal docked structures of each inhibitor. The highest-ranking conformers/poses, according to each scoring function, were aligned together to construct corresponding CoMFA models that were subsequently appropriately validated. The cycle of docking, scoring, and CoMFA modeling was repeated to cover all possible docking combinations resulting from the presence (or absence) of crystallographically explicit water molecules within the binding site and the ionization states of the ligands (ionized or un-ionized).

Our results illustrate the ability of certain docking/scoring conditions to access self-consistent CoMFA models that are comparable and even superior to reported atom-aligned (i.e., structure-based) CoMFA models developed for h-PTP 1B inhibitors, which required the incorporation of log *P* (logarithm of partition coefficient) as additional descriptor to achieve satisfactory statistical significance.[66]

Despite the presence of several comparative studies of various scoring functions on a number of docking programs, the scoring results in these studies were judged from the similarity of the docked compounds to corresponding crystallographic structures or the ability to identify known active compounds from a random pool.[28,35,36,64,67−69] However, we believe such success criteria suffer from the implicit assumption that crystallographic structures of bound ligands are sufficiently realistic to be used as reference standards. Although crystallographic data are considered the most reliable structural information that can be used for drug design, they are associated with some serious problems such as inadequate resolution[70] and crystallization-related distortions inflected upon the structure of the ligand−protein complex.[71] Accordingly, we believe that judging the success of a particular docking-scoring combination from the statistical qualities of the corresponding protein-aligned CoMFA model provides an interesting additional validation of the docking approach.

Results

As described above, we evaluated the effects different docking approaches (i.e., scoring functions, ligand ionization state, and binding site hydration) on the statistical qualities of the corresponding protein-aligned CoMFA models constructed from diverse h-PTP 1B inhibitors. The compounds were docked into the binding site employing two separate docking configurations, that is, wide and restricted. In the wide docking configuration, LigandFit was instructed to explore the stabilities of wider diversity of potential docked conformers/poses (i.e., compared to the restricted docking settings) and to tightly fit promising conformers/poses into the binding pocket via an extended number of energy minimization iterations (see Docking Simulations under Methods). Subsequently, the docked structures were aligned together for molecular field analysis (MFA) utilizing genetic partial least squares (G/PLS) for statistical modeling.

The inhibitors were divided into two groups: a training subset of 110 compounds and a randomly selected test subset of 27 compounds. The test compounds were selected to represent a range of biological activities similar to that of the training set (see Dataset under Methods).

The qualities of each CoMFA model were assessed via five statistical criteria: (i) conventional regression coefficients against 110 training compounds (r_{110}^2), (ii) leave-one-out regression coefficients (r^2(LOO)), (iii) bootstrapping regression coefficients (r_{BS}^2), (iv) predictive regression coefficients against the external set of 27 test compounds (r^2(PRESS)), and (v) the sum of squared deviations between predicted and experimental bioac-

Table 1. The Structures of h-PTP 1B Compounds Utilized in Modeling[a]

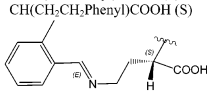
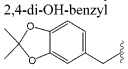
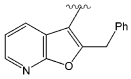
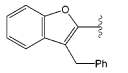
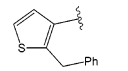
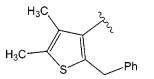
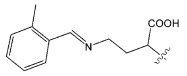
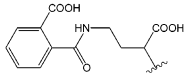
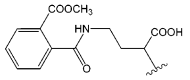
Compound	R1	R2	R3	R4	R5	X	IC50 µmol
1	butyl	H	H	---	---	O	0.740
2	benzyl	H	H	---	---	O	0.920
3	benzoyl	H	H	---	---	O	0.740
4	butyl	H	H	---	---	S	0.700
5	4-OH-benzyl	H	H	---	---	S	1.080
6	2,4-di-OH-benzyl	H	H	---	---	S	0.580
7	butyl	CH_2COOH	H	---	---	O	2.190
8	butyl	$CH(CH_2Phenyl)COOH$	H	---	---	O	0.440
9	benzyl	$CH(CH_2Phenyl)COOH$	H	---	---	O	0.270
10	benzyl	$CH(CH_2Phenyl)COOH$ (R)	H	---	---	O	0.350
11	benzyl	$CH(CH_2Phenyl)COOH$ (S)	H	---	---	O	0.320
12	benzyl	$CH(CH_2CH_2Phenyl)COOH$ (S)	H	---	---	O	0.220
13	benzyl		H	---	---	O	0.340
14	butyl	CH_3	H	---	---	O	2.500
15	ethyl	H	H	---	---	O	2.500
16	H	$CH(CH_2Phenyl)COOH$	H	---	---	O	2.500
17	benzyl	$CCH_3(CH_2Phenyl)COOH$	H	---	---	O	0.290
18	benzyl	CH(Phenyl)COOH (R)	H	---	---	O	0.400
19	benzyl	$CH(CH_3)COOH$ (R)	H	---	---	O	1.320
20	benzoyl	$CH(CH_2Phenyl)COOH$ (R)	H	---	---	O	0.680
21	CH(OH)phenyl	$CH(CH_2Phenyl)COOH$ (R)	H	---	---	O	0.110
22	benzyl	$CH_2Phenyl$-4-COOH	H	---	---	O	0.360
23	butyl	$CH(CH_2Phenyl)COOH$ (R)	H	---	---	S	0.170
24	benzyl	$CH(CH_2Phenyl)COOH$ (R)	H	---	---	S	0.095
25	butyl	CH(Phenyl)COOH (R)	H	---	---	S	0.110
26	4-F-benzyl	$CH(CH_2Phenyl)COOH$ (R)	H	---	---	S	0.120
27	4-OCH_3-benzyl	$CH(CH_2Phenyl)COOH$ (R)	H	---	---	S	0.077
28	3,4-di-OCH_3-benzyl	$CH(CH_2Phenyl)COOH$ (R)	H	---	---	S	0.120
29	2,4-di-OCH_3-benzyl	$CH(CH_2Phenyl)COOH$ (R)	H	---	---	S	0.085
30	2,4-di-OH-benzyl	$CH(CH_2Phenyl)COOH$ (R)	H	---	---	S	0.120
31		$CH(CH_2Phenyl)COOH$ (R)	H	---	---	S	0.077
32	2-Thiazolinyl-CH_2-	$CH(CH_2Phenyl)COOH$ (R)	H	---	---	S	1.160
33	2-Pyridinyl-CH_2-	$CH(CH_2Phenyl)COOH$ (R)	H	---	---	S	1.550
34	benzyl	$CH(CH_2Phenyl)COOH$(S)	F	---	---	O	0.130
35	benzyl	$CH(CH_2Phenyl)COOH$ (R)	CH_3	---	---	O	0.410
36		$CH(CH_2Phenyl)COOH$ (S)	---	---	---	---	0.590
37		$CH(CH_2Phenyl)COOH$ (R)	---	---	---	---	0.350
38		$CH(CH_2Phenyl)COOH$ (R)	---	---	---	---	0.970
39		$CH(CH_2Phenyl)COOH$ (R)	---	---	---	---	0.510
40	Br	H	H	H	H	S	1.070
41	Br	Br	H	H	H	S	0.450
42	I	I	H	H	H	S	0.520
43	Br	H	$CH(CH_2Phenyl)COOH$ (R)	H	H	S	0.058
44	Br	Br	$CH(CH_2Phenyl)COOH$ (R)	H	H	S	0.025
45	4-OCH_3-Phenyl	H	$CH(CH_2Phenyl)COOH$ (R)	H	H	S	0.053
46	4-Cl-Phenyl	H	$CH(CH_2Phenyl)COOH$ (R)	H	H	S	0.052
47	Br	Br	$CH(CH_2CH_2Phenyl)COOH$ (S)	H	H	S	0.290
48	Br	Br		H	H	S	0.044
49	Br	Br		H	H	S	0.180
50	Br	Br		H	H	S	0.054
51	Br	H	CH_2COOH	H	H	S	0.360
52	Br	Br	CH_2COOH	H	H	S	0.100
53	Phenyl	H	CH_2COOH	H	H	S	0.100
54	4-OCH_3-Phenyl	H	CH_2COOH	H	H	S	0.080
55	4-OC_2H_5-Phenyl	H	CH_2COOH	H	H	S	0.052
56	2,3-di-OCH_3-Phenyl	H	CH_2COOH	H	H	S	0.071
57	3,4,5-tri-OCH_3-Phenyl	H	CH_2COOH	H	H	S	0.100
58	4-OCH_3-Phenyl	Br	CH_2COOH	H	H	S	0.029
59	3-OCH_3-Phenyl	Br	CH_2COOH	H	H	S	0.028
60	2,4-di-OCH_3-Phenyl	Br	CH_2COOH	H	H	S	0.047
61	4-OCH_3-Phenyl	4-OCH_3-Phenyl	CH_2COOH	H	H	S	0.025
62	3-OCH_3-Phenyl	3-OCH_3-Phenyl	CH_2COOH	H	H	S	0.025
63	Br	H	$CH_2CH_2CH_2COOH$	H	H	S	0.170
64	Br	H	$CH(CH_2Phenyl)COOH$ (S)	H	H	O	0.056
65	Br	Br	$CH(CH_2Phenyl)COOH$ (S)	H	H	O	0.038
66	4-OCH_3-Phenyl	H	$CH(CH_2Phenyl)COOH$ (S)	H	H	O	0.043
67	NO_2	H	$CH(CH_2Phenyl)COOH$ (R)	H	H	O	0.230
68	Br	Br	$CH(CH_2CH_3)COOH$ (S)	H	H	O	0.130
69	Br	Br	$CH[CH_2CH(CH_3)_2]COOH$ (R)	H	H	O	0.054

Table 1. (Continued)

Compound	R1	R2	R3	R4	R5	X	IC50 μmol
70	Br	Br	$CH[(CH_2)_3CH_3]COOH$	H	H	O	0.052
71	Br	Br	$CH[(CH_2)_5CH_3]COOH$	H	H	O	0.023
72	CH_3	CH_3	$CH(CH_2Phenyl)COOH$ (R)	H	H	O	0.074
73	cyclopentyl	H	$CH(CH_2Phenyl)COOH$ (S)	H	H	O	0.055
74	cyclopentyl	H	CH_2COOH	H	H	O	0.170
75	$NHCH_2COOH$	H	$CH_2CH_2Phenyl$	H	H	O	0.082
76	$NHCH_2CH_2COOH$	H	$CH_2CH_2Phenyl$	H	H	O	0.140
77	$NHCO\text{-}CH_2CH_2\text{-}COOH$	H	H	H	H	O	0.920
78	(*E*)NHCO-CH=CH-COOH	H	H	H	H	O	0.460
79	$NHCO\text{-}C_6H_4\text{-}2\text{-}COOH$	H	H	H	H	O	0.160
80	$4\text{-}OCH_3\text{-}Phenyl$	$4\text{-}OCH_3\text{-}Phenyl$	CH_2COOH	F	H	O	0.048
81	$4\text{-}OCH_3\text{-}Phenyl$	$4\text{-}OCH_3\text{-}Phenyl$	CH_2COOH	H	F	O	0.031
82	H	H	H	---	---	CH_2	1.190
83	H	H	H	---	---	C=O	2.500
84	H	H	H	---	---	CH(OH)	0.230
85	H	Br	Br	---	---	CH(OH)	1.400
86	CH_2COOH	H	H	---	---	CH_2	1.150
87	CH_2COOH	H	H	---	---	CH(OH)	0.540
88	CH_2-tetrazole	H	H	---	---	CH_2	0.510
89	H	H	H	---	---	---	2.260
90	CH_2COOH	H	H	---	---	---	0.800
91	$CH(CH_2Phenyl)COOH$	H	H	---	---	---	1.300
92		H	H	---	---	---	0.900
93	H	Br	Br	---	---	---	0.650
94	CH_2COOH	Br	Br	---	---	---	0.470
95	$CH(CH_2Phenyl)COOH$	Br	Br	---	---	---	0.130
96	---	---	---	---	---	---	1.600
97	H	H	---	---	---	CH_2	1.300
98	H	H	---	---	---	CH(OH)	1.100
99	H	Br	---	---	---	CH(OH)	0.480
100	H	Br	---	---	---	CH_2	0.330
101	H	I	---	---	---	CH_2	0.380
102	CH_2COOH	Br	---	---	---	CH_2	1.400
103	$CH(CH_2Phenyl)COOH$	Br	---	---	---	CH_2	0.370
104	$CH(CH_2Phenyl)COOH$	Br	---	---	---	C=O	1.200
105	$CH(CH_2Phenyl)COOH$	I	---	---	---	CH_2	0.320
106	CH_2-tetrazole	Br	---	---	---	CH_2	0.700
107	CH_2-tetrazole	Br	---	---	---	C=O	1.100
108	H	H	---	---	---	C=O	3.050
109	CH_2COOH	Br	---	---	---	---	1.300
110	$CH(CH_2Phenyl)COOH$	H	---	---	---	---	3.050
111	H	COOH	H	H	---	O	0.075
112	COOH	H	H	H	---	O	0.106
113	OH	COOH	H	H	---	O	0.039
114	COOH	OH	H	H	---	O	0.026
115	OH	COOH	CH_3	CH_3	---	O	0.034
116	OH	COOH	H	NO_2	---	O	0.029
117	OH	COOH	H	cyclopentyl	---	O	0.028
118	OH	COOH	H	H	---	S	0.028
119	OH	COOH	H	Br	---	S	0.024
120	OH	COOH	Br	Br	---	S	0.030
121	---	---	---	---	---	---	0.032
122	OH	COOH	Cyclopentyl	Benzoyl	---	---	0.040
123	OH	COOH	H	H	---	---	0.354
124	OAcetyl	COOH	H	H	---	---	1.160
125	OH	COOH	NO_2	Benzoyl	---	---	0.178
126	Phenyl	CH_3	H	---	---	---	0.300
127	CH_3	Propyl	H	---	---	---	1.900
128	CH_3	Butyl	H	---	---	---	1.400
129	CH_3	Octyl	H	---	---	---	0.300
130	Phenyl	Butyl	H	---	---	---	0.370
131	CH_3	Phenyl	CH_2COOH	---	---	---	0.850
132	CH_3	(*E*) Octyl	CH_2COOH	---	---	---	0.160
133	CH_3	(*Z*) Octyl	CH_2COOH	---	---	---	0.120
134	3,5-dichlorophenyl	---	---	---	---	---	1.300
135	3,5-dichlorobenzyl	---	---	---	---	---	48.000
136	---	---	---	---	---	---	13.000
137	---	---	---	---	---	---	3.050

[a] The corresponding scaffolds are as in Figure 1.

tivities for molecules in the test set (PRESS). Table 2 shows the statistical criteria of the resulting CoMFA models.

Despite that traditional cross-validation tests such as r^2(LOO) are very useful,[9] they do not always pick up poor equations.[72,73] Accordingly, we decided to further validate superior models that combine $r^2(\text{LOO}) \geq 0.500$ and $r^2(\text{PRESS}) \geq 0.45$ (Table 2) by calculating their leave-20%-out (r^2(L-20%-O)) and randomization correlation coefficients (r^2(random)). These tests ensure that the generated regression models were not produced by chance.[73] G/PLS is a powerful search technique that might yield apparently significant chance correlations necessitating these additional validation tests.[72,73]

It is clearly evident from Table 2 that the best CoMFA model, that is, with the highest combination of r^2(LOO) and r^2(PRESS), was generated by docking the ionized ligands into the hydrated form of the binding site via restricted docking settings and PLP1 scoring function (code Res1; $r^2(\text{LOO}) = 0.647$, $r^2(\text{PRESS}) = 0.617$). Interestingly, this combination of docking conditions maintained high-quality statistical models regardless to the settings of the G/PLS analysis, as shown in Table 3. On the other hand, CoMFA models resulting from other docking−scoring configurations were quite sensitive to alterations in G/PLS settings, particularly to the number of latent variables and genetic iterations. Figure 2 illustrates the experimental bioactivities versus fitted (110 training set) and predicted (27 testing compounds) values produced by the best Res1-aligned CoMFA model.

Undoubtedly, the apparent robustness of Res1-based 3D QSAR models reflects the realistic qualities of the corresponding docked conformers/poses. Figure 3 shows

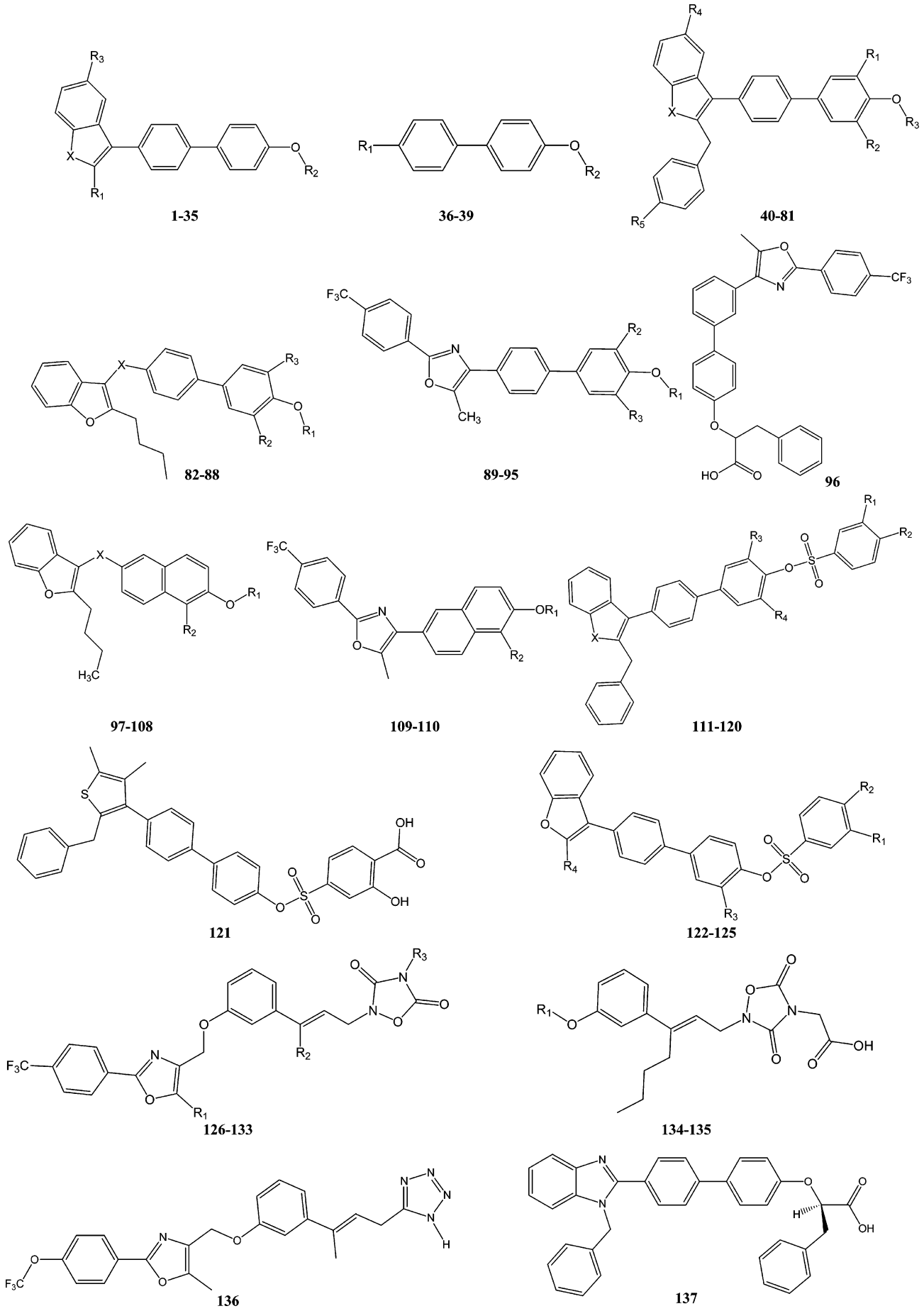


Figure 1. The chemical scaffolds of different h-PTP 1B inhibitors shown in Table 1.

Table 2. The Statistical Results of the Best CoMFA Models Obtained via Various Docking−Scoring Combinations

			restricted docking exploration[h]							wide docking exploration[h]						
ligand ionization state	explicit water	scoring function	code[a]	LV[b]	r^2(110)[c]	r^2(LOO)[d]	r^2(BS)[e]	r^2(PRESS)[f]	PRESS[g]	code[a]	LV[b]	r^2(110)[c]	r^2(LOO)[d]	r^2(BS)[e]	r^2(PRESS)[f]	PRESS[g]
un-ionized	present	PLP1	Res1	5	**0.807**	**0.647**	**0.751**	**0.617**	**3.821**	Wd1	3	0.805	0.714	0.774	0.113	8.839
		PLP2	Res2	3	0.728	0.589	0.693	0.264	7.338	Wd2	2	**0.750**	**0.687**	**0.749**	**0.467**	**5.310**
		LigScore1	Res3	3	0.743	0.613	0.701	0.155	8.418	Wd3	3	0.736	0.566	0.711	−0.489	14.834
		LigScore2	Res4	4	**0.756**	**0.536**	**0.668**	**0.451**	**5.475**	Wd4	2	**0.754**	**0.624**	**0.736**	**0.591**	**4.075**
		LUDI	Res5	4	0.760	0.530	0.637	0.217	7.804	Wd5	2	0.755	0.675	0.755	0.088	9.086
		PMF	Res6	3	0.753	0.611	0.715	0.409	5.894	Wd6	2	0.799	0.705	0.760	0.150	8.468
	absent	PLP1	Res7	4	0.763	0.594	0.680	0.199	7.981	Wd7	4	0.767	0.577	0.684	−0.009	10.059
		PLP2	Res8	4	0.759	0.592	0.697	−0.309	13.049	Wd8	2	0.747	0.674	0.737	0.302	6.960
		LigScore1	Res9	3	0.751	0.636	0.732	−0.201	11.963	Wd9	2	0.690	0.583	0.630	0.263	7.343
		LigScore2	Res10	3	0.810	0.735	0.772	0.017	9.792	Wd10	3	0.713	0.613	0.713	0.241	7.567
		LUDI	Res11	3	0.737	0.636	0.714	0.209	7.880	Wd11	2	0.747	0.693	0.747	0.049	9.475
		PMF	Res12	5	0.762	0.596	0.622	0.274	7.239	Wd12	3	**0.752**	**0.651**	**0.727**	**0.516**	**4.826**
ionized	present	PLP1	Res13	5	**0.791**	**0.622**	**0.705**	**0.452**	**5.459**	Wd13	2	0.779	0.699	0.779	0.378	6.201
		PLP2	Res14	5	0.797	0.492	0.663	0.281	7.164	Wd14	3	**0.776**	**0.677**	**0.775**	**0.498**	**4.999**
		LigScore1	Res15	5	0.732	0.586	0.681	0.323	6.749	Wd15	6	0.803	0.682	0.779	−0.320	13.155
		LigScore2	Res16	4	**0.772**	**0.572**	**0.772**	**0.448**	**5.499**	Wd16	5	0.799	0.666	0.746	0.292	7.058
		LUDI	Res17	5	0.781	0.605	0.717	0.209	7.887	Wd17	5	0.739	0.515	0.683	0.019	9.777
		PMF	Res18	5	0.785	0.577	0.716	0.185	8.118	Wd18	2	0.742	0.673	0.742	0.301	6.965
	absent	PLP1	Res19	4	0.798	0.648	0.766	−0.422	14.174	Wd19	4	0.720	0.599	0.686	−0.053	10.493
		PLP2	Res20	5	0.790	0.662	0.740	−0.272	12.680	Wd20	4	0.704	0.612	0.704	−0.443	14.379
		LigScore1	Res21	5	0.780	0.584	0.703	0.194	8.032	Wd21	7	0.769	0.626	0.744	−0.102	10.982
		LigScore2	Res22	3	0.784	0.687	0.748	0.366	6.314	Wd22	2	0.749	0.631	0.742	0.134	8.628
		LUDI	Res23	5	0.777	0.632	0.741	0.299	6.989	Wd23	4	0.779	0.652	0.742	0.369	6.285
		PMF	Res24	3	**0.738**	**0.637**	**0.705**	**0.517**	**4.814**	Wd24	2	0.796	0.656	0.765	−0.420	12.547
				(5)	**(0.789)**	**(0.618)**	**(0.727)**	**(0.450)**	**(5.480)**							

[a] These codes stand for restricted (Res) or wide (Wd) docking configurations, while the associated serial numbers indicate the corresponding docking−scoring conditions, that is, ligand ionization, presence of explicit water and the corresponding scoring function (these codes are used in captions of tables and figures and in the text to indicate the corresponding docking−scoring conditions). [b] Optimal number of latent variables at 20 000 generations of G/PLS. [c] Non-cross-validated correlation coefficient for 110 training compounds. [d] Cross-validation correlation coefficients determined by the leave-one-out technique. [e] Bootstrapping correlation coefficient. [f] Predictive r^2 determined for the 27 test compounds. [g] The sum of squared deviations between predicted and actual activity values for every molecule in the test set of 27 compounds. [h] Bold statistical parameters correspond to the best docking/scoring combinations.

Table 3. The Effects of Variable G/PLS Settings on the Statistical Criteria of CoMFA Models Obtained for the Training Compounds Aligned by Res1 Docking−Scoring Conditions (See Table 2)

G/PLS settings			statistical parameters				
LV[a]	no. of generations	no. of terms in model	r^2(110)[b]	r^2(LOO)[c]	r^2(BS)[d]	r^2(PRESS)[e]	PRESS[f]
3	10 000	18	0.744	0.586	0.658	0.502	4.965
3	20 000	18	0.775	0.693	0.759	0.422	5.762
4	10 000	18	0.735	0.595	0.686	0.607	3.914
4	20 000	18	0.767	0.586	0.685	0.577	4.213
4	20 000	15	0.776	0.666	0.748	0.575	4.237
5	10 000	18	0.755	0.521	0.661	0.681	3.183
5	20 000	18	0.807	0.647	0.751	0.617	3.821
5	20 000	21	0.806	0.642	0.757	0.403	5.950

[a] Number of PLS latent variables. [b] Non-cross-validated correlation coefficient for 110 training compounds. [c] Cross-validation correlation coefficients determined by the leave-one-out technique. [d] Bootstrapping correlation coefficient. [e] Predictive r^2 determined for the 27 test compounds. [f] The sum of squared deviations between predicted and actual activity values for every molecule in the test set of 27 compounds.

the alignment of the inhibitors into the binding pocket of h-PTP 1B via the Res1 docking−scoring configuration. Figures 7a and 9a illustrate the docked conformers/poses of two potent inhibitors (**61** and **119**, $IC_{50} = 0.025$ and 0.024 μM, respectively) as generated by the same docking conditions. The three figures (Figures 3, 7a, and 9a) suggest a significant role played by the water molecules within the binding site leading to the superior docking/CoMFA results of Res1 conditions. This conclusion is supported by the fact that most of the other high-ranking CoMFA models (of bold statistical parameters in Table 2) coincide with docking experiments involving the hydrated form of the binding site regardless of the docking configuration (wide or restricted) or ligand ionization state (ionized or un-ionized), albeit in conjunction with PLP and LigScore scoring functions (Table 2), that is, Res1 (r^2(LOO) = 0.647, r^2(PRESS) = 0.617), Res4 (r^2(LOO) = 0.536, r^2(PRESS) = 0.451), Res13 (r^2(LOO) = 0.622, r^2(PRESS) = 0.452), Res16 (r^2(LOO) = 0.572, r^2(PRESS) = 0.448), Wd2 (r^2(LOO) = 0.687, r^2(PRESS) = 0.467), Wd4 (r^2(LOO) = 0.624, r^2(PRESS) = 0.591), and Wd14 (r^2(LOO) = 0.677, r^2(PRESS) = 0.498). Incidentally, LigandFit/PLP and LigandFit/Ligscore were reported to yield superior docking accuracies when compared to other docking−scoring combinations.[28,36,64]

However, surprisingly, two of the promising 3D QSAR models coincided with docking into the anhydrous form of the binding pocket, albeit in conjunction with PMF scoring function (Table 2), that is, Wd12 (r^2(LOO) = 0.651, r^2(PRESS) = 0.516) and Res24 (r^2(LOO) = 0.637, r^2(PRESS) = 0.517). PMF was reported to yield self-consistent CoMFA models when combined with DOCK4 to align 51 biphenyl carboxylic acid MMP-3 inhibitors.[10]

Further assessment of high-ranking CoMFAs by looking at their r^2(PRESS) values after removing the worst-predicted test compound from their testing lists (r^2(PRESS−26), Tables 4 and 5) emphasized specifically the significance of QSAR models corresponding to Res1 (r^2(PRESS−26) = 0.634), Res24 (r^2(PRESS−26) = 0.627),

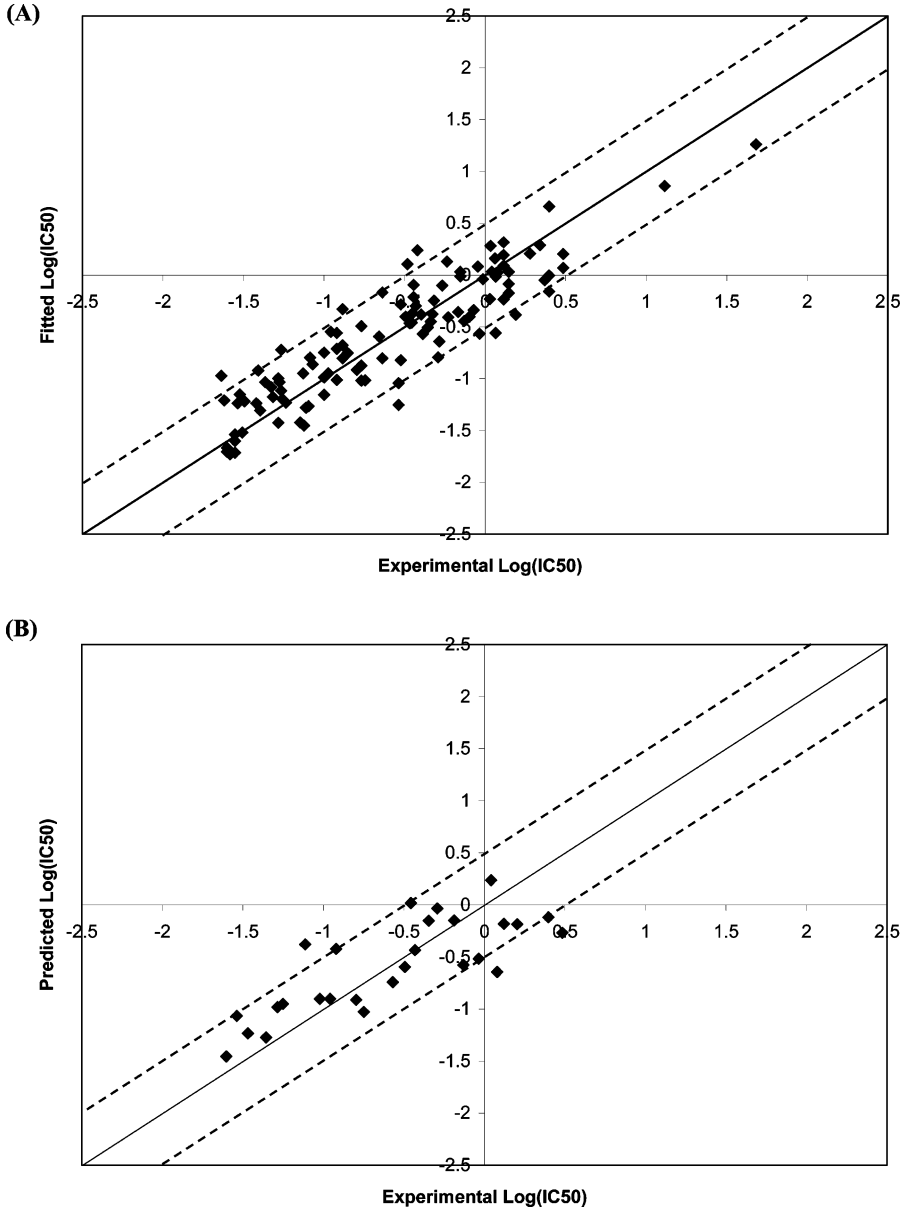


Figure 2. Experimental versus fitted (A, 110 compounds, r^2(LOO) = 0.647) and predicted (B, 27 compounds, r^2(PRESS) = 0.617) bioactivities calculated from the best CoMFA model obtained after 20 000 iterations of G/PLS (five latent variables, LVs) performed on the training compounds aligned by Res1 docking−scoring conditions (code as in Table 2, restricted docking configuration, hydrated binding pocket, un-ionized ligands, and PLP1 scoring). The solid lines are the regression lines for the fitted and predicted bioactivities of training and test compounds, respectively, whereas the dotted lines indicate the ±0.5 log point error margins.

Wd4 (r^2(PRESS−26) = 0.636), and Wd12 (r^2(PRESS−26) = 0.659) because they exhibit r^2(PRESS−26) > 0.60. A particularly interesting case is the CoMFA model related to Wd12 settings (Tables 2 and 5), which illustrated an impressive jump in r^2(PRESS) upon removing inhibitor **116** from the test set, that is, from 0.516 to 0.659. Figures 4−6 illustrate the experimental versus fitted (110 compounds) and predicted (27 compounds) bioactivities employing Res24, Wd4, and Wd12 docking−scoring conditions.

Discussion

By evaluating the binding interactions proposed by the highest-performing docking approaches (i.e., Res1, Res24, Wd4, and Wd12), one can identify three primary regions within the binding site involved in hydrogen bonding with the docked ligands (Figures 7−10). The first region is comprised of Tyr46, Lys116, Lys120, and Ser216 combined with two hydrogen-bonded water molecules. The second region is composed of Cys215, Ala217, Arg221, and Gln266, while the third region is comprised of Arg24, Arg254, Met258, Gln262, and two hydrogen-bonded water molecules. Additionally, Arg47 and Asp48 contribute to ligand−protein binding, albeit not through hydrogen bonding. All these binding regions were implicated in substrate recognition and the overall catalytic activities of h-PTP 1B.[62] However, Cys215, in particular, provides the necessary nucleophile required for cleaving the phosphate group of phosphotyrosine moieties.[62]

Moreover, through careful evaluation of the docked conformers/poses produced by the top four docking settings, it is possible to classify the ligands into two general categories based on their binding profiles: a

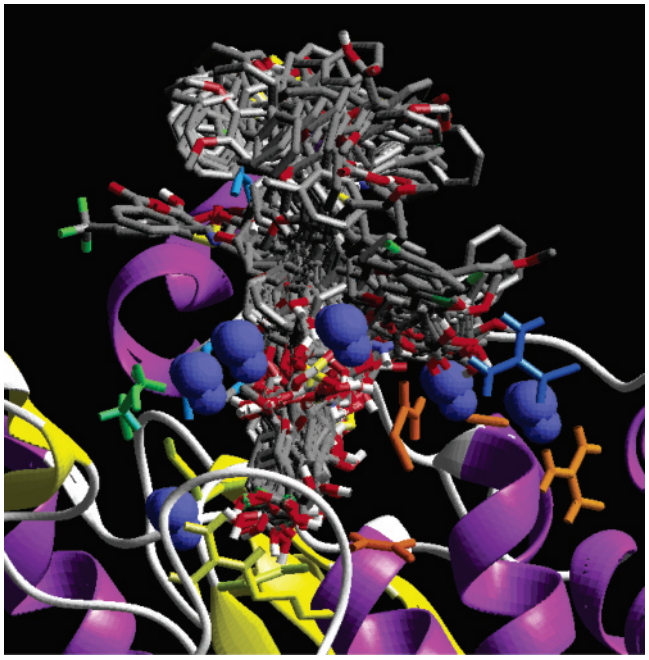


Figure 3. Alignment of the docked inhibitors within the binding pocket of h-PTP 1B as proposed by Res1 docking−scoring conditions. The structures were docked into the binding site in the presence of crystallographically explicit water molecules (shown as blue spheres). The image also shows some essential amino acid moieties involved in the binding of most of the docked inhibitors (more details are in Figures 7−10 and associated text).

major "linear" group characterized by small-sized substituents on their middle scaffolds (i.e., the bisphenyl, naphthyl, or phenoxyalkenyl moieties) and a minor bulkier "branched" group with larger aromatic or alicyclic substituents on their middle scaffolds. To discuss the docking details of the inhibitors, we selected compounds **119** and **61** (IC_{50} = 0.024 and 0.025 μM, respectively) to represent the linear and branched sets, respectively.

Figures 7a and 9a illustrate the bound conformers/poses of **61** and **119** as proposed by Res1 docking−scoring conditions. Evidently, this approach orients the carboxylic acid group of **61** (Figure 7a) toward the third binding region within the binding site where it forms direct hydrogen bonds with Arg24 and Gln262 and water-mediated hydrogen bonds with Arg254 and Met258 (Figures 7a and 8a). Furthermore, additional stabilization seems to result from the interaction of the benzyl moiety of the benzyl-benzothiophene in **61** with the guanidine side chain of Arg47.

On the other hand, Res1 conditions direct the "linear" analogues toward the first and second binding regions of the binding pocket as exemplified by **119** in Figure 9a. Clearly from the figure, the salicylic acid moiety of **119** is involved in hydrogen-bonding interactions with the side chains of Gln266 and Cys215 and the amidic hydrogen of Arg221. These interactions are reminiscent of the hydrogen bonds that anchor phosphotyrosine substrates within h-PTP 1B.[62] Furthermore, Figure 9a shows the involvement of the sulfonic acid ester of **119** in hydrogen-bonding interactions with Lys120, Tyr46, and Ser216, albeit via two hydrogen-bonded water molecules. Additional stabilizing interactions seem to come from Arg47 and Asp48 that occur at close proximities to the benzothiophene and bromo-benzene moieties of **119**, respectively, which allow effective attractive interactions between the charged side chains of these two amino acids and their electronically complementary aromatic neighbors from the ligand (Figures 9a and 10a).

Interestingly, Wd4-based docking produced generally similar molecular poses to those of Res1 conditions despite some noticeable conformational differences, as shown in Figures 7b and 9b. Their major difference is the flipped conformation of the benzyl-benzothiophene fragments (of both **61** and **119** and related analogues). Furthermore, under Wd4 conditions, some members of the "linear" group, for example, **119**, lost most of their hydrogen-bonding interactions with the first (Tyr46, Lys116, Lys120, and Ser216) and second (Cys215, Ala217, Arg221, and Gln266) binding regions. Apparently, Wd4 flipped the benzyl-benzothiophene groups of **61** and **119** and shifted the sulfonic ester and salicylic acid moieties of **119** away from the first and second binding regions, respectively, to optimally align the benzothiophene fragment of the ligands with the guanidine group of Arg47 to promote their maximal mutual attractive interactions. Moreover, this setting seems to optimize the orientation of the bromo-benzene of **119** relative to the carboxylic acid side chain of Asp48 to allow their optimal attraction.

Finally, regarding Wd12 and Res24 docking settings, dehydrating h-PTP 1B under these conditions, caused dramatic changes in the steric dimensions of the binding pocket, which seem to clear the way for the carboxylic acid moieties of **61** and **119** to access and interact directly with the positively charged side chains of Lys120 and Lys116 (i.e., the first binding region), as evident in Figures 8c,d and 10c. This trend is particularly evident in the case of Res24 (ionized ligands, Figures 7c and 9c), which is probably due to the strong electrostatic attractive interactions between the carboxylate moieties of the ligands and the quaternary ammonium groups of Lys120 and Lys116. However, additional stabilizing interactions contribute to the orientations and conformations of Res24-docked ligands. Accordingly, this approach aligns the electron-rich benzothiophene moiety of **61** (and related analogues) with the positive guanidine group of Arg24 for optimal mutual attraction (Figures 7c and 8c). On the other hand, Res24 imposes a certain odd conformation/pose on **119** and related analogues (Figures 9c and 10c), in which the ligand's salicylic carboxylate is electrostatically bound to Lys120 and Lys116, while one of the sulfonic acid oxygens is hydrogen-bonded to Ala217. However, the rest of the molecule is erected nearly perpendicular to the salicylic acid moiety in such a way to allow the benzyl group of the benzyl-benzothiophene and the phenyl ring of the bisphenyl fragment to approach and interact with the charged side chains of Arg47 and Asp48, respectively.

Interestingly, the results of Wd12 settings mimic closely those of Res24 regarding inhibitor **61** (and other close analogues), as shown in Figures 7d and 8d. However, their effect on **119** and other "linear" analogues is quite unique, as they orient the benzyl-benzothiophene of **119** (or benzyl benzofuran in other analogues) moiety toward Lys120 for hydrogen bonding

Table 4. Additional Cross-Validation Statistics Calculated for the Highest-Ranking CoMFA Models Obtained via Restricted Docking Settings (That Is, Models of bolded PRESS and r^2(PRESS) Values in Table 2)

								predictive statistics using the test set without one outlier (26 inhibitors only)		
docking conditions		scoring function	code[a]	LV[b]	r^2(L-20%-O)[c]	r^2(random)[d]	F_{test}	r^2(PRESS)[e]	PRESS[e]	outlier[f]
with explicit water molecules	un-ionized ligands	PLP1	Res1	5	0.634	0.146	24.500	0.631	3.257	**48**
		LigScore2	Res4	4	0.674	0.145	16.500	0.520	4.341	**116**
	ionized ligands	PLP1	Res13	5	0.693	0.147	21.613	0.540	4.571	**125**
		LigScore2	Res16	4	0.724	0.155	19.766	0.550	3.972	**137**
without explicit water molecules	ionized ligands	PMF	Res24	3	0.561	0.133	14.881	0.627	3.552	**104**

[a] Codes correspond to the "restricted" docking−scoring conditions, as in Table 2. [b] Optimal number of latent variables for 20 000 generations of G/PLS and 18 explanatory variables. [c] Cross-validation correlation coefficients determined by the leave-20%-out technique repeated 10 times. [d] The average randomization correlation coefficients: the biological activities were randomized 99 times, and the mean randomization r^2 was calculated. [e] Predictive r^2(PRESS) and PRESS determined for 26 test compounds. [f] Worst predicted test compounds of highest predicted-minus-actual [log(IC_{50})] absolute difference. Compound numbers are as in Table 1.

Table 5. Additional Cross-Validation Statistics Calculated for the Highest-Ranking CoMFA Models Obtained via Wide Docking Settings (That Is, Models of Bolded PRESS and r^2(PRESS) Values in Table 2)

								predictive statistics using the test set without one outlier (26 inhibitors only)		
docking conditions		scoring function	code[a]	LV[b]	r^2(L-20%-O)[c]	r^2(random)[d]	F_{test}	r^2(PRESS)[e]	PRESS[e]	outlier[f]
with explicit water molecules	un-ionized ligands	PLP2	Wd2	2	0.706	0.151	16.814	0.546	4.250	**48**
		Ligscore2	Wd4	2	0.754	0.149	18.963	0.636	3.521	**27**
	ionized ligands	PLP2	Wd14	3	0.714	0.153	15.416	0.555	4.158	**93**
without explicit water molecules	un-ionized ligands	PMF	Wd12	3	0.641	0.153	19.198	0.659	3.088	**116**

[a] Codes correspond to the "wide" docking−scoring conditions, as in Table 2. [b] Optimal number of latent variables for 20 000 generations of G/PLS and 18 explanatory variables. [c] Cross-validation correlation coefficients determined by the leave-20%-out technique repeated 10 times. [d] The average randomization correlation coefficients: the biological activities were randomized 99 times, and the mean randomization r^2 was calculated. [e] Predictive r^2(PRESS) and PRESS determined for 26 test compounds. [f] Worst predicted test compounds of highest predicted-minus-actual [log(IC_{50})] absolute difference. Compound numbers are as in Table 1.

(Figures 9d and 10d). We believe the major driving force for this pose is the tendency of Wd12 to orient the hydrophilic groups of the ligands toward the aqueous exterior; moreover, this pose seems to be stabilized by additional attraction resulting from the close proximity between the bromine atom of the brominated bisphenyl ring and the guanidine group of Arg24.

The fact that several docking−scoring configurations yielded self-consistent CoMFA models, despite their orthogonality (their cross-correlation regression coefficients $r^2 \leq 0.70$, Table 6) and the apparent differences among their respective conformers/poses, hints to the existence of multiple binding modes adopted by various inhibitors within the binding pocket of h-PTP 1B. However, this conclusion requires further experimental evidence to be substantiated, particularly from crystallographic studies.

Still, the emergence of several satisfactory orthogonal CoMFA models prompted us to envisage the possibility of merging them in a single multiple linear regression QSAR equation that combines their predictive potentials. Therefore, the bioactivity estimates (fitted values) produced by each of the four models were treated as independent explanatory descriptors, while the corresponding experimental bioactivities were considered as the response variable. Equation 1 shows the resulting correlation. However, Figure 11 illustrates experimental bioactivities versus fitted (110 training set) and predicted (27 testing compounds) values determined by employing the combination regression model.

$$\log(IC_{50}) = 0.097 + (0.440 \pm 0.135)\log(IC_{50})_{Res1} + (0.283 \pm 0.134)\log(IC_{50})_{Res24} + (0.211 \pm 0.150) \times \log(IC_{50})_{Wd4} + (0.233 \pm 0.149)\log(IC_{50})_{Wd12}$$

$$r^2 = 0.90;\quad r^2(\mathrm{LOO}) = 0.890;\quad n = 110;\quad F = 232.63;\quad r^2(\mathrm{PRESS-27}) = 0.75 \quad (1)$$

where $\log(IC_{50})_{Res1}$, $\log(IC_{50})_{Res24}$, $\log(IC_{50})_{Wd4}$, and $\log(IC_{50})_{Wd12}$ are bioactivity estimates produced by CoMFA models corresponding to Res1, Res24, Wd4, and Wd12 docking−scoring settings, respectively. The 95% confidence limits (CL) of different regression coefficients are shown (±CL).

It is clearly evident from Figure 11 and the statistical criteria of eq 1 that the combination QSAR model is of excellent fitting and prediction potentials. The significance of such a model in drug discovery can be summarized in two points: (i) In virtual high-throughput screening, molecules can be docked into the binding pocket of a targeted biological macromolecule employing high-performing docking−scoring conditions deduced from such a study. Subsequently, the associated optimal CoMFA models can be utilized to prioritize hit compounds. (ii) Within the context of structure-based de-

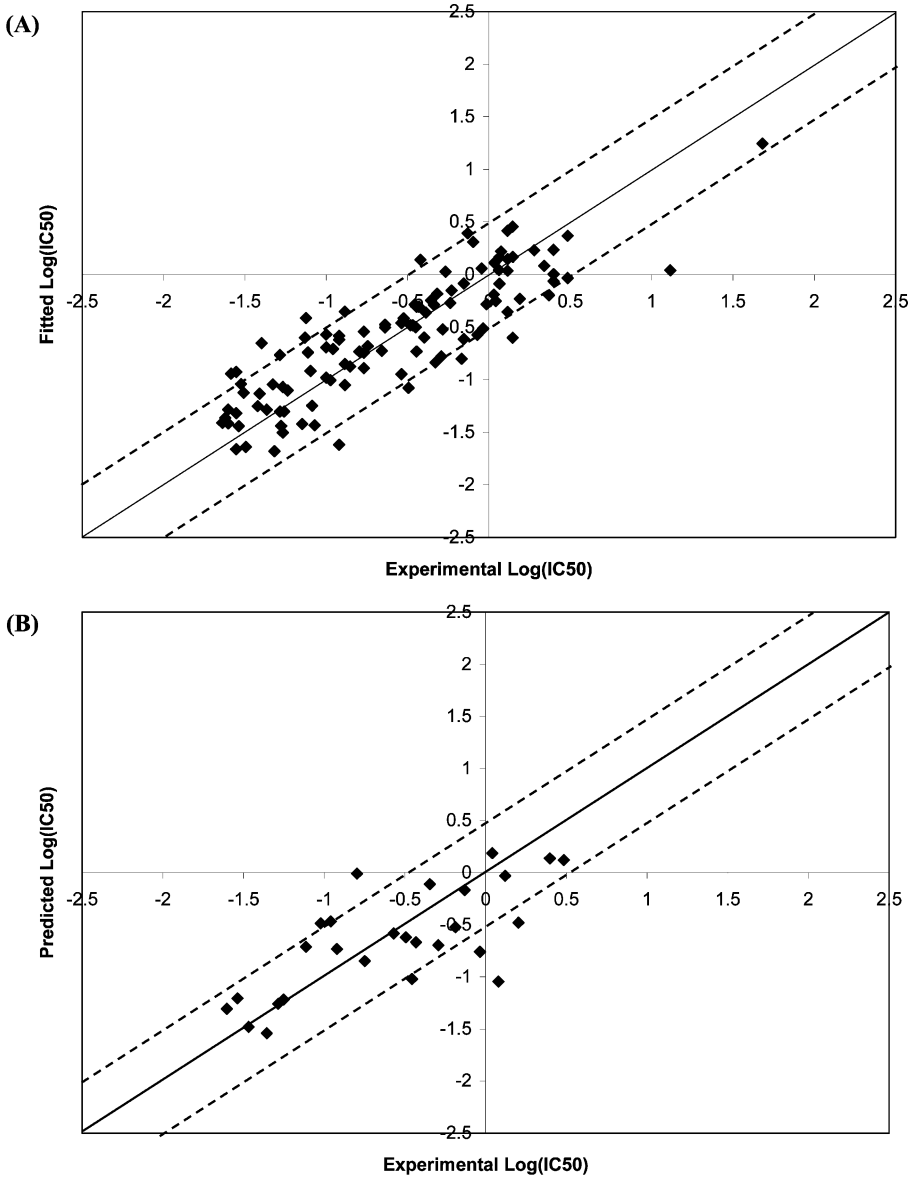


Figure 4. Experimental versus fitted (A, 110 compounds, r^2(LOO) = 0.637) and predicted (B, 27 compounds, r^2(PRESS) = 0.517) bioactivities calculated from the best CoMFA model obtained after 20 000 iterations of G/PLS (three LVs) performed on the training compounds aligned utilizing Res24 conditions (code as in Table 2, restricted docking configuration, anhydrous binding pocket, ionized ligands, and PMF scoring). The solid lines are the regression lines for the fitted and predicted bioactivities of training and test compounds, respectively, whereas the dotted lines indicate the ±0.5 log point error margins.

sign, the docked confomers/poses of potent inhibitors, generated by the high-performing docking−scoring combinations, can be utilized to identify essential ligand−protein interactions, which can be subsequently used to design more potent inhibitors of extended attractive interactions within the binding pocket.

Conclusions

A group of diverse inhibitors were docked into the active site of h-PTP 1B employing the LigandFit docking engine. The docked conformers/poses were utilized to construct corresponding protein-aligned CoMFA models. We evaluated the effects of a variety of docking−scoring factors on the statistical properties of the corresponding CoMFA models. Few docking configurations succeeded in producing self-consistent CoMFA models. However, the best model coincided with docking the un-ionized ligands into the hydrated form of the binding site via the PLP1 scoring function and restricted docking settings (r^2(LOO) = 0.647, r^2(PRESS−27) = 0.617). Interestingly, the best-performing docking approaches (i.e., those that yielded the most significant 3D QSAR models) generated significantly different binding conformers/poses. To utilize the predictive potentials of the highest-ranking CoMFA models collectively, we decided to combine them in a single QSAR equation. The combination model illustrated excellent predictive properties against a 27-membered external test set of inhibitors (r^2(PRESS) = 0.75). This approach should enhance the predictive potential of protein-aligned CoMFA modeling; furthermore, it provides an interesting way to benefit from the differences among various docking−scoring functions.

Methods

Hardware and Software. Docking, scoring, and molecular field analysis studies were performed using the CERIUS2 suite of programs (version 4.9) from Accelrys Inc. (San Diego, CA,

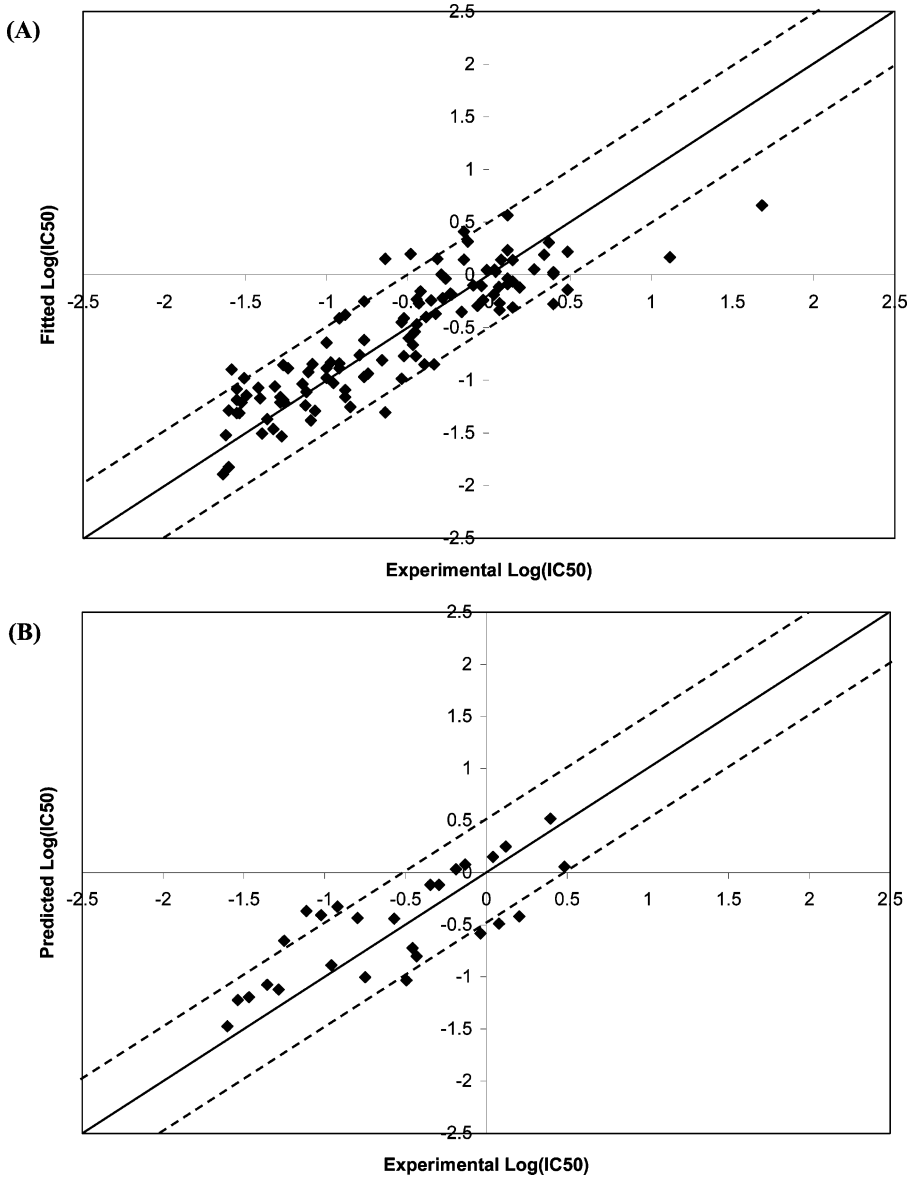


Figure 5. Experimental versus fitted (A, 110 compounds, r^2(LOO) = 0.624) and predicted (B, 27 compounds, r^2(PRESS) = 0.591) bioactivities calculated from the best CoMFA model obtained after 20 000 iterations of G/PLS (two LVs) performed on the training compounds aligned by Wd4 conditions (as in Table 2, wide docking configuration, hydrated binding pocket, un-ionized ligands, and LigScore2 scoring). The solid lines are the regression lines for the fitted and predicted bioactivities of training and test compounds, respectively, whereas the dotted lines indicate the ±0.5 log point error margins.

www.accelrys.com) installed on a Silicon Graphics Octane2 desktop workstation equipped with a 600 MHz MIPS R14000 processor (1.0 GB RAM) running the Irix 6.5 operating system.

Dataset. A set of 137 h-PTP 1B inhibitors belonging to benzofuran/benzothiophene biphenyls[60] (compounds **1**−**125** and **137** in Figure 1 and Table 1) and azolidinediones[61] (compounds **126**−**136** in Figure 1 and Table 1) were used for modeling. The in vitro bioactivities of most of the collected inhibitors were expressed as the concentration of the test compound that inhibited recombinant h-PTP 1B activity by 50% (IC_{50}). However, the activities of a few benzofuran/benzothiophene biphenyls were expressed as the average percent inhibition of h-PTP 1B at 50, 10, 2.5, 1.0, 0.25, or 0.10 μM inhibitor concentrations.[60] Generally, in such cases, the respective inhibitors were either excluded from modeling if their reported percent inhibition values were outside a 45%−55% interval or included in modeling by estimating their IC_{50} values provided they were within the 45%−55% inhibition interval at the particular reported inhibitory concentration. Table 1 and Figure 1 show the structures and IC_{50} values of the considered inhibitors.

The logarithm of measured IC_{50} (μM) values was used in 3D QSAR, thus correlating the data linearly to the free energy change. A training subset of 110 molecules was selected. However, since it is essential to access the predictive power of the resulting CoMFA models on an external set of inhibitors, the remaining 27 molecules (ca. 20% of the dataset) were employed as an external test subset for validating the 3D QSAR models. The test molecules were selected as follows: the 137 inhibitors were ranked according to their IC_{50} values, then every fifth compound was selected for the test set starting from the high-potency end. This selection considers the fact that the test molecules must represent a range of biological activities similar to that of the training set. The selected test compounds are **3**, **9**, **15**, **19**, **21**, **24**, **27**, **30**, **37**, **41**, **48**, **62**, **64**, **70**, **77**, **88**, **93**, **94**, **98**, **104**, **105**, **115**, **116**, **125**, **130**, **132**, and **137** (numbers are as in Table 1 and Figure 1).

Preparation of the h-PTP 1B Inhibitors. The three-dimensional structures of the inhibitors (**1**−**137**) were sketched in CERIUS2. Two protonation states were assumed for each inhibitor, ionized and un-ionized. In the ionized forms, the carboxylic acids ($pK_a \approx 4.0-4.5$) and tetrazole groups ($pK_a \approx$

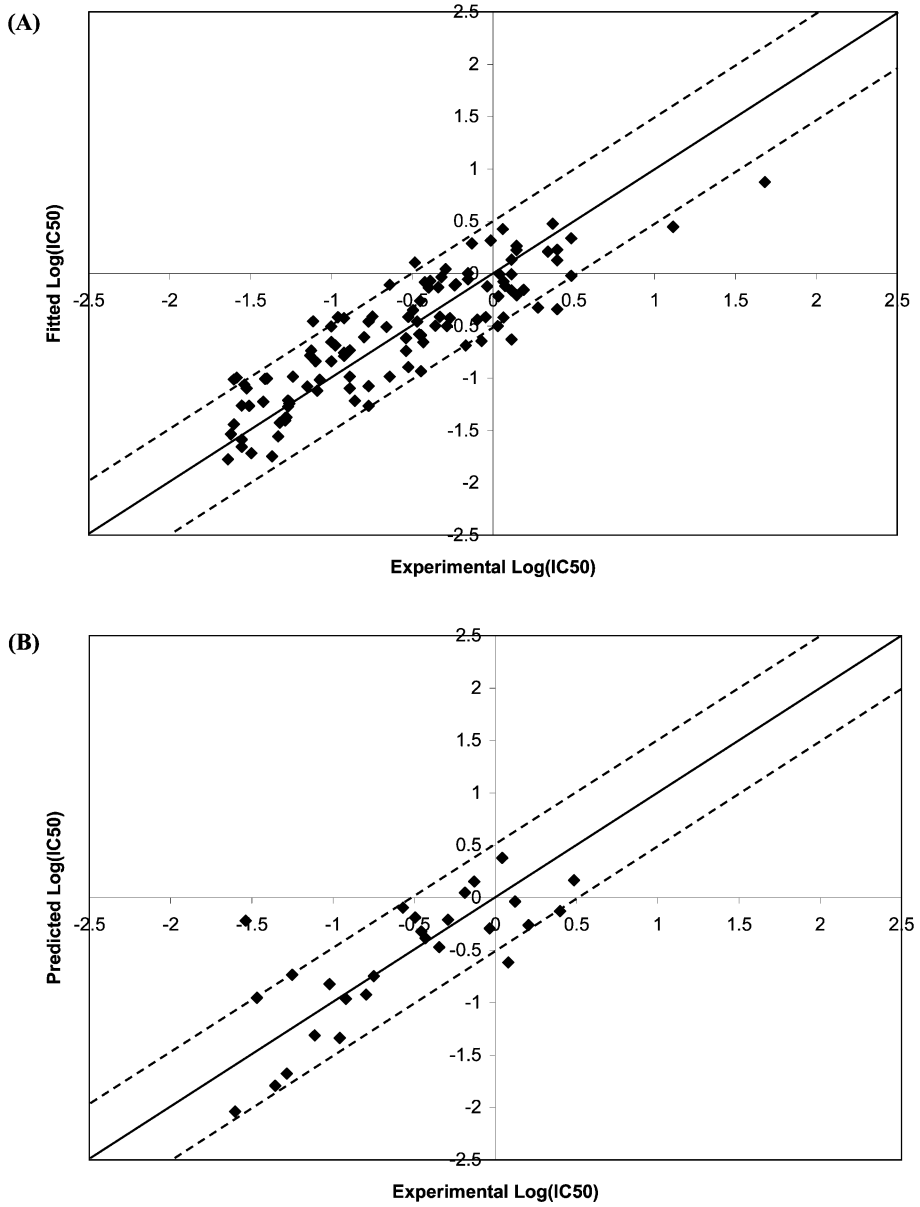


Figure 6. Experimental versus fitted (A, 110 compounds, r^2(LOO) = 0.651) and predicted (B, 27 compounds, r^2(PRESS) = 0.516) bioactivities calculated from the best CoMFA model obtained after 20 000 iterations of G/PLS (three LVs) performed on the training compounds aligned by Wd12 docking−scoring conditions (as in Table 2, wide docking configuration, anhydrous binding pocket, un-ionized ligands, and PMF scoring). The solid lines are the regression lines for the fitted and predicted bioactivities of training and test compounds, respectively, whereas the dotted lines indicate the ±0.5 log point error margins.

5.0−5.5) were deprotonated and given formal negative charges at the appropriate atoms. However, none of the inhibitors included amino functionalities, which excluded the possibility of formal positive ionization. The remaining atoms within the inhibitor structures were assigned partial charges using the default Gasteiger method[74] implemented in CERIUS2. The structures were subsequently energy-minimized employing the UNIVERSAL force field (version 1.02, default settings) implemented within CERIUS2.[75]

Racemic inhibitors (i.e., **8**, **9**, **16**, **17**, **21**, **48**, **49**, **50**, **70**, **71**, **84**, **85**, **87**, **91**, **95**, **96**, **103**, **104**, **105**, and **110** in Table 1 and Figure 1) were assigned arbitrary absolute chiral configurations for subsequent modeling. This assumption is based on the fact that the enantiomeric pairs of most of these ligands behaved similarly under different docking−scoring approaches, that is, they yielded visually similar docked conformers/poses.

h-PTP 1B Crystal Structure. The 3D coordinates of h-PTP 1B were retrieved from the Protein Data Bank (PDB code 1g7f). The selected structure is that exhibiting the best 3D resolution (1.80 Å) compared to other available h-PTP 1B structures. Hydrogen atoms were added to the protein utilizing CERIUS2 templates for protein residues. Gasteiger charges were assigned to the protein atoms as implemented within LigandFit.[34,63] The protein structure was utilized in subsequent docking experiments without energy minimization. Explicit water molecules were either kept or removed according to the required docking conditions, that is, docking in the presence or absence of explicit water molecules.

Docking Simulations. LigandFit considers the flexibility of the ligand and treats the receptor as rigid. There are two steps implemented in the LigandFit process: (1) Defining the location(s) of potential binding site(s) by shape-based search for cavities in the protein. The algorithm for cavity detection calculates a rectangular grid enclosing the protein, cavity regions, and explicit water molecules around the complex. The protein is mapped on the grid. All grid points occupied by the protein (or crystallographically explicit water molecules) are not available in the site search. The unoccupied grid points inside the protein are potential binding sites. However, if a certain ligand is cocrystallized within the targeted protein then

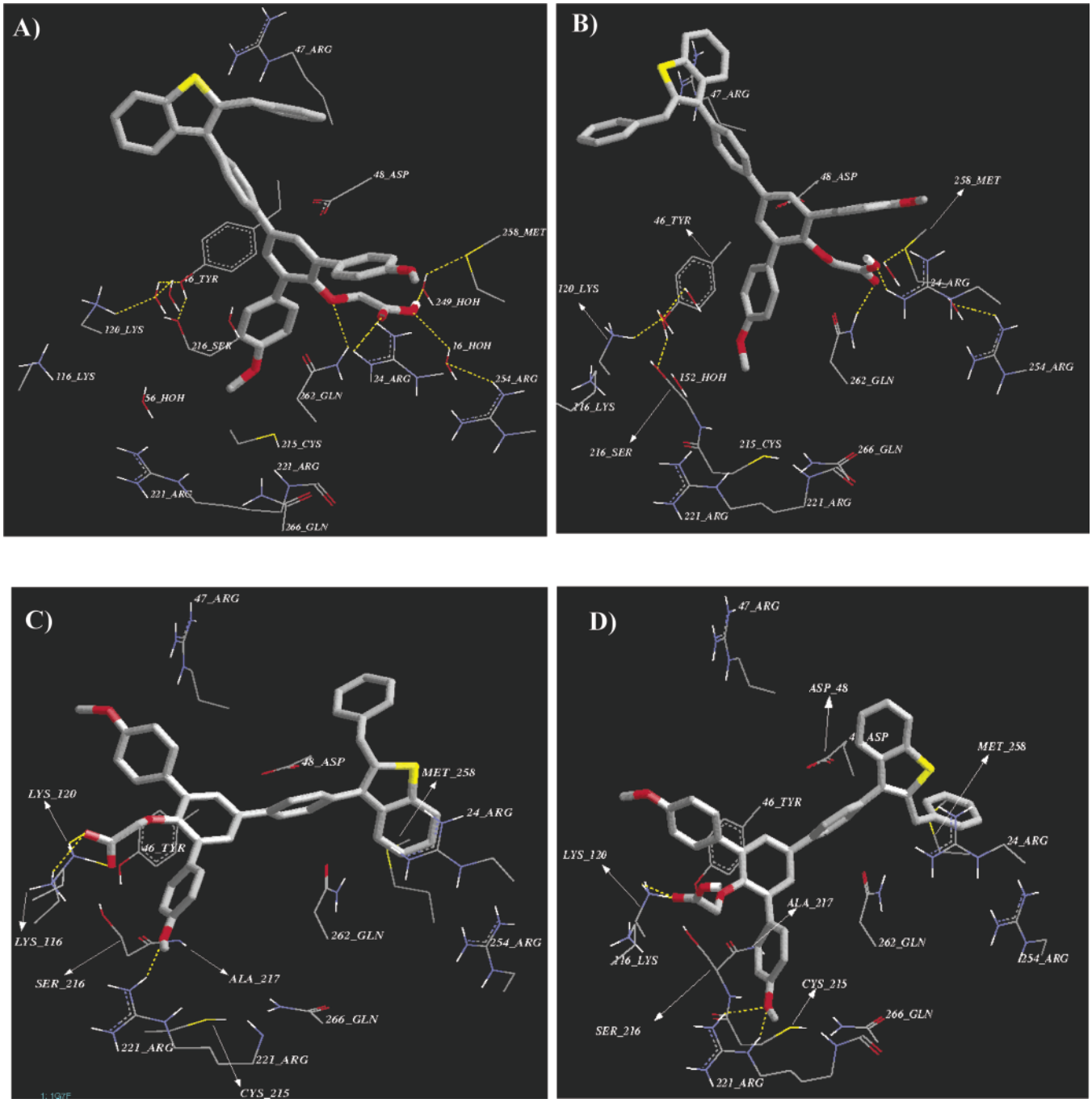


Figure 7. The docked conformers/poses of inhibitor **61** (IC_{50} = 0.025 μmol) as generated via the best docking−scoring combinations: (A) Res1; (B) Wd4; (C) Res24; (D) Wd12 (codes are as in Table 2). The dotted yellow lines illustrate the positions of probable hydrogen-bonding interactions as calculated by the H-Bond calculator imbedded in CERIUS2.

it is possible to generate the binding site from the docked ligand by collecting all grid points that lie within the radius of any atom of the ligand to form the binding site.[34,63] In the current docking experiments the binding site was generated from the cocrystallized ligand within the targeted protein. The grid resolution was set to 0.5 Å, the radius of hydrogen atoms in the cocrystallized ligand and the protein was set to 2.0 Å, while the radius of heavy atoms (carbon, nitrogen, oxygen, and sulfur) in the cocrystallized ligand and the protein was set to 2.5 Å. (2) Docking the ligands in the binding site. In LigandFit, docking is composed of few major substeps:[34,63] (i) Conformational search of the flexible ligand employing Monte Carlo randomized process is performed. (ii) Pose/conformation selection based on shape similarity with the binding site is made. (iii) Candidate conformers/poses exhibiting low shape discrepancy are further enrolled in calculation of the dock and interaction energies. The dock energy is composed of two terms, namely, the internal energy of the ligand and the interaction energy with the receptor, summarized by van der Waals and electrostatic energy terms. To improve the time-consuming computation of the interaction energy, an approximation by grid-based interpolation is employed in LigandFit. A grid encloses the site, and at each point of the grid, the potentials are computed for the active site. The potentials at the ligand atom locations are subsequently interpolated. (iv) Each docked conformation/pose is further fitted into the binding pocket through a number of rigid-body minimization iterations, that is, minimization of the interaction energy via molecular rotations and translations of the docked ligand. (v) Docked conformers/poses that have docking energies below a certain user-defined threshold are subsequently clustered according to their RMS similarities. Representative conform-

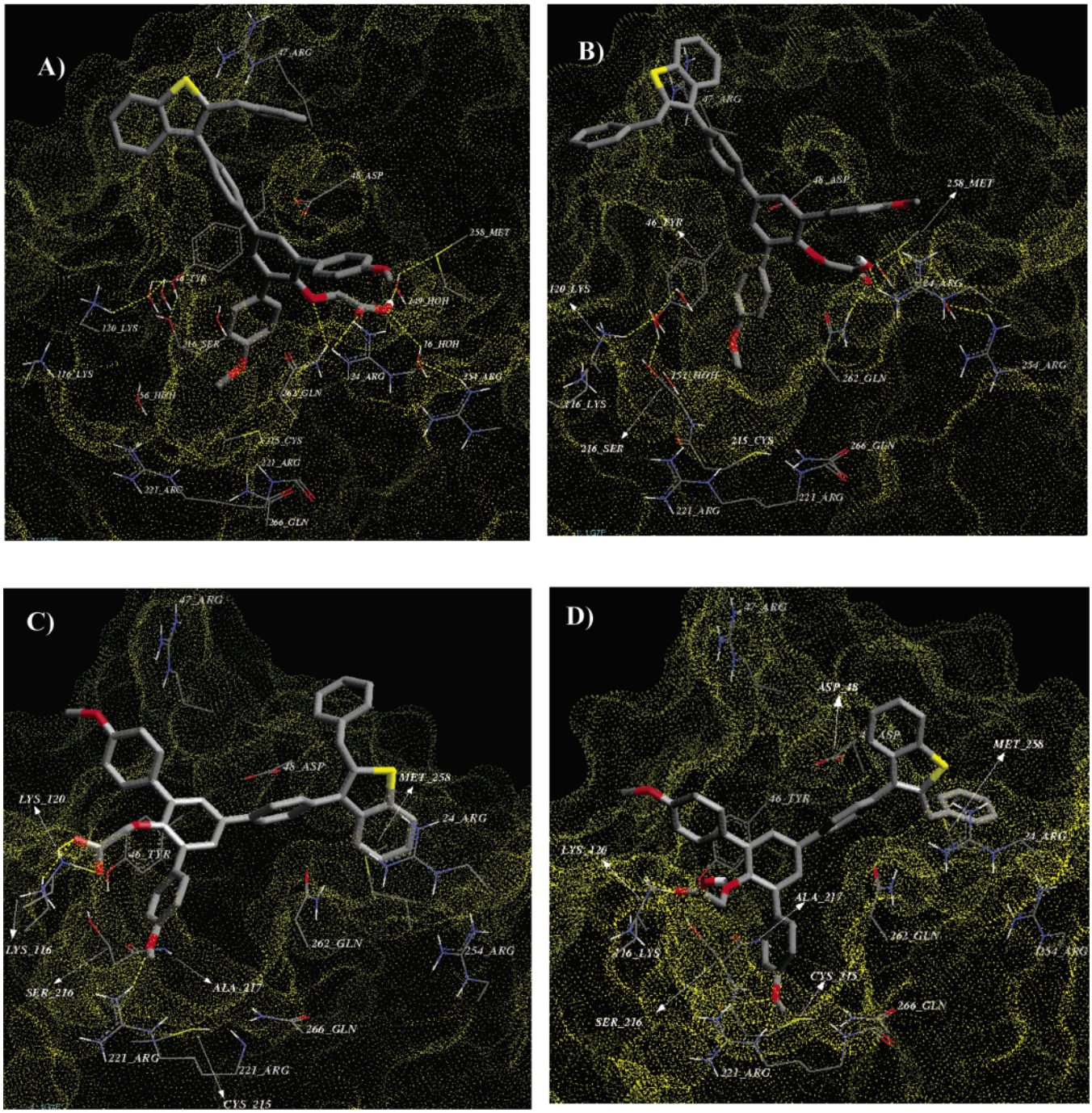


Figure 8. Illustrations of Connolly's water-accessible surface (dotted yellow surface) calculated for h-PTP 1B employing CERIUS2 and its spatial relationships to the docked conformers/poses of inhibitor **61** as generated by (A) Res1, (B) Wd4, (C) Res24, and (D) Wd12 docking−scoring conditions (codes are as in Table 2).

ers/poses are then selected, further energy-minimized within the binding site, and saved for subsequent scoring.

In the current docking experiments, all 137 ligands (training and testing subsets) in their ionized and un-ionized forms were docked into the binding site in the presence and absence of explicit water molecules, employing two separate docking configurations.

(I) Restricted Exploration Settings. (i) Monte Carlo search parameters were number of trials = 10 000 and search step for torsions with polar hydrogens = 30.0°. (ii) The RMS threshold for ligand-to-binding site shape match was set to 2.0 employing a maximum of 1.0 binding site partitions. (iii) Interaction energy parameters were determined as follows: The interaction energies were assessed employing Drieding force field (version 2.21) with a nonbonded cutoff distance of 10.0 Å and distance dependent dielectric. An energy grid extending 3.0 Å from the binding site was implemented. The interaction energy was estimated by a trilinear interpolation value using soft potential energy approximations.[34] (iv) Rigid-body ligand minimization parameters were determined as follows: 10 iterations of rigid-body minimization (molecular translational and rotational movements) were applied to every orientation of the docked ligand. (v) The docked conformations/poses of calculated interaction energies ≤20.0 kcal/mol were clustered using the complete linkage algorithm in CERIUS2 with RMS similarity threshold of 1.5 Å. The best member within the cluster was selected and was further energy-minimized within the binding site for a maximum of 100 rigid-body iterations and 250 flexible conformation iterations. Eventually, a maximum of 10 optimal conformers/poses were saved for each molecule for subsequent scoring.

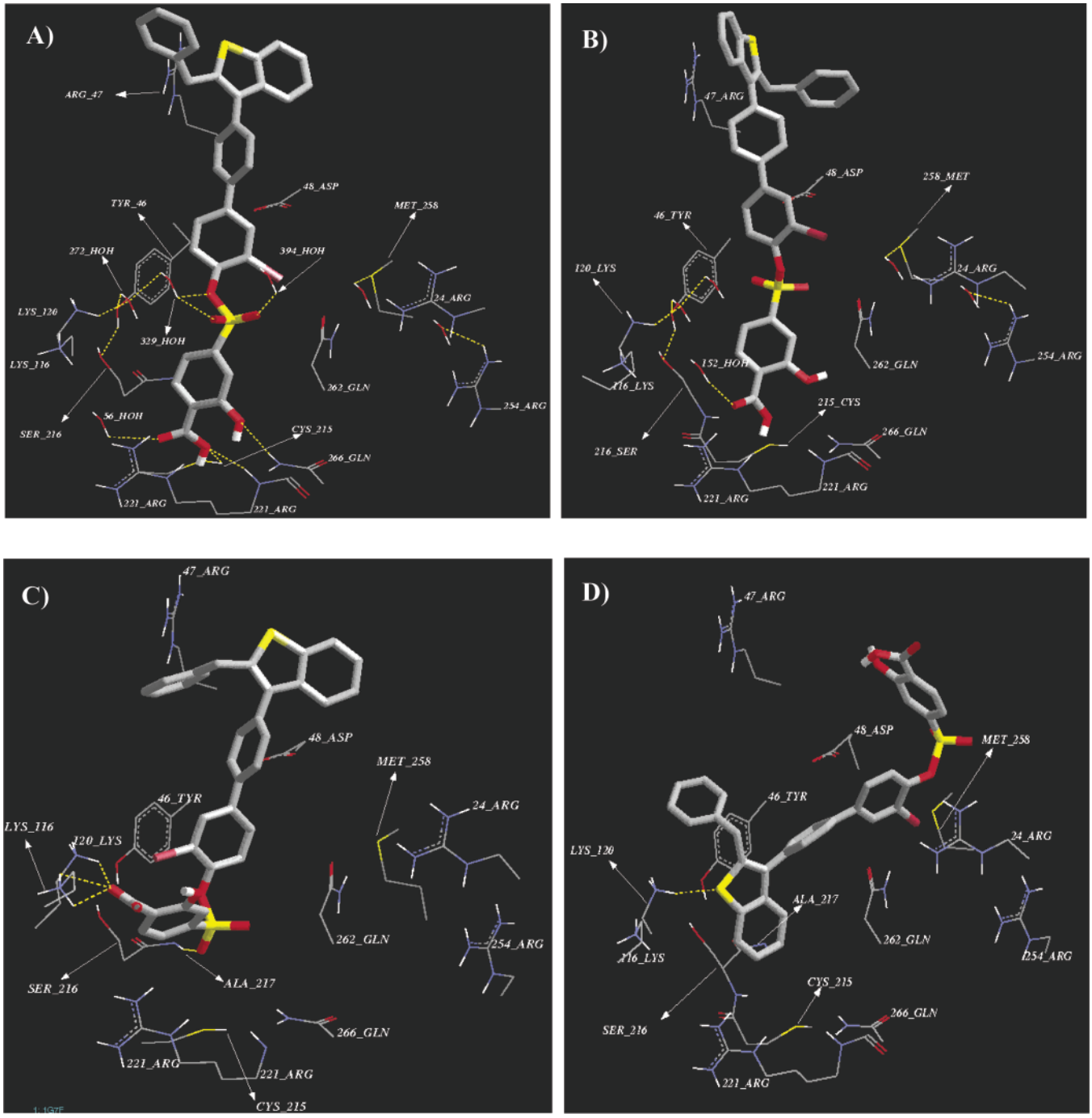


Figure 9. The docked conformers/poses of inhibitor **119** (IC_{50} = 0.024 μmol) produced by the most successful docking−scoring combinations: (a) Res1; (b) Wd4; (c) Res24; (d) Wd12 (codes are as in Table 2). The dotted yellow lines illustrate the positions of probable hydrogen-bonding interactions as calculated by the H-Bond calculator imbedded in CERIUS2.

(II) Wide Exploration Settings. (i) Monte Carlo search parameters were number of trials = 20 000 and search step for torsions with polar hydrogens = 10.0°. (ii) The RMS threshold for ligand-to-site shape match was set to 2.0 employing a maximum of 5.0 binding site partitions. (iii) Interaction energy parameters were determined as follows: An energy grid was implemented using Drieding force field (version 2.21) with a nonbonded cutoff distance of 10.0 Å and distance dependent dielectric constant. The energy grid was extended 3.0 Å from the binding site. The interaction energy is approximated by a trilinear interpolation value using soft potential energy approximations.[34] (iv) Rigid-body ligand minimization parameters were determined as follows: 50 iterations of rigid body minimization were applied to every orientation of the docked ligand. (v) Docked conformations/poses of calculated interaction energies ≤20.0 kcal/mol were clustered using the complete linkage algorithm in CERIUS2 with RMS similarity threshold of 1.5 Å. The best member within the cluster was selected and was further energy-minimized within the binding site for a maximum of 250 rigid-body iterations and 500 flexible conformation iterations. A maximum of 10 optimal conformers/poses were saved for each molecule for subsequent scoring.

Scoring of Docked Conformers/Poses. For all optimal docked conformers/poses, we computed scores using the scoring functions LigScore1[34,63,65] LigScore2,[34,63,65] LUDI,[41,42] PLP1,[39] PLP2,[40] and PMF.[45−47] Considering each scoring function in turn, the highest scoring docked conformer/pose was selected for each inhibitor for subsequent 3D QSAR modeling. This resulted in six sets of 137 docked molecules with scores corresponding to each scoring function. However, the docking and scoring cycle was repeated 8 times (2 × 2 × 2) to cover the different combinations of docking conditions, that is, ligand

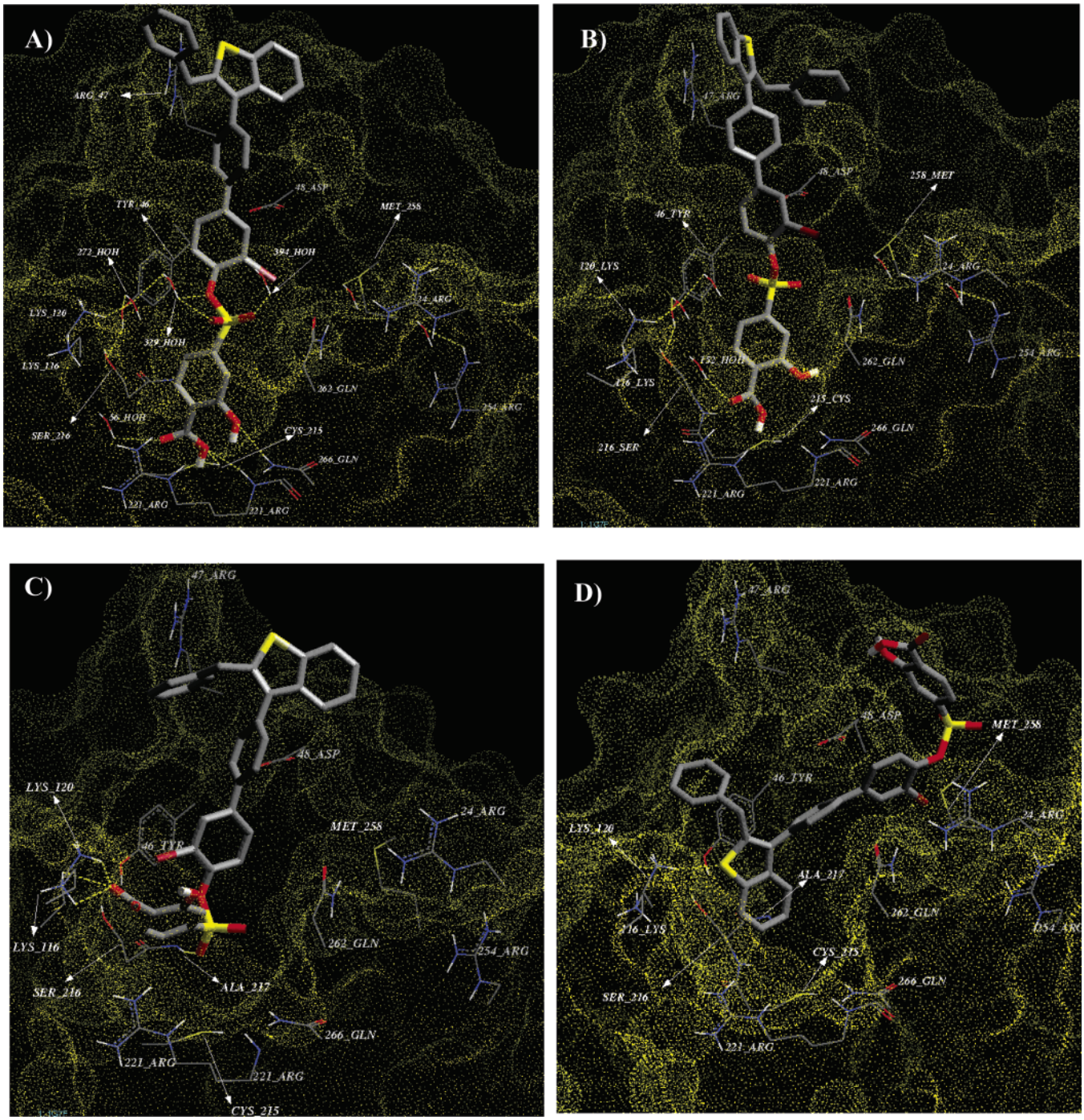


Figure 10. Illustrations of Connolly's water-accessible surface (dotted yellow surface) calculated for h-PTP 1B and its spatial relationships to the docked conformers/poses of inhibitor **119** as generated by (A) Res1, (B) Wd4, (C) Res24, and (D) Wd12 docking–scoring conditions (codes are as in Table 2).

Table 6. The Cross-Correlation Coefficients (r^2 Values) among Bioactivity Estimates of the Training Compounds as Produced by the Top CoMFA Models Corresponding to Res1, Wd4, Res24, and Wd12 Docking Configurations

	Res1	Wd4	Res24	Wd12
Res1	1.00			
Wd4	0.68	1.00		
Res24	0.62	0.60	1.00	
Wd12	0.63	0.70	0.65	1.00

ionization state (two possibilities), existence of crystallographic explicit water molecules (two possibilities), and LigandFit docking configurations (two possibilities).

LigScore1 and LigScore2 scores were calculated employing Dreiding force field (version 2.21) and using grid-based energies with a grid extension of 7.5 Å across the binding site. PMF scores were calculated employing cutoff distances for carbon–carbon interactions and other interactions of 12.0 Å.

Molecular Field Analysis. The molecular field analysis (MFA) and G/PLS modules within CERIUS2 were used to perform 3D QSAR analyses.[76] The alignments of different inhibitors came directly from the top-scoring conformers/orientations according to each considered docking/scoring combination, as mentioned earlier. For each alignment, the interaction fields between the ligands and proton (positively charged), hydrogen-bond donor/acceptor, and methyl (neutral) probes were calculated employing a regularly spaced rectangular grid of 1.0 Å spacing. The spatial limits of the molecular field were defined automatically and were extended past the van der Waals volume of all the molecules in the *X*, *Y*, and *Z*

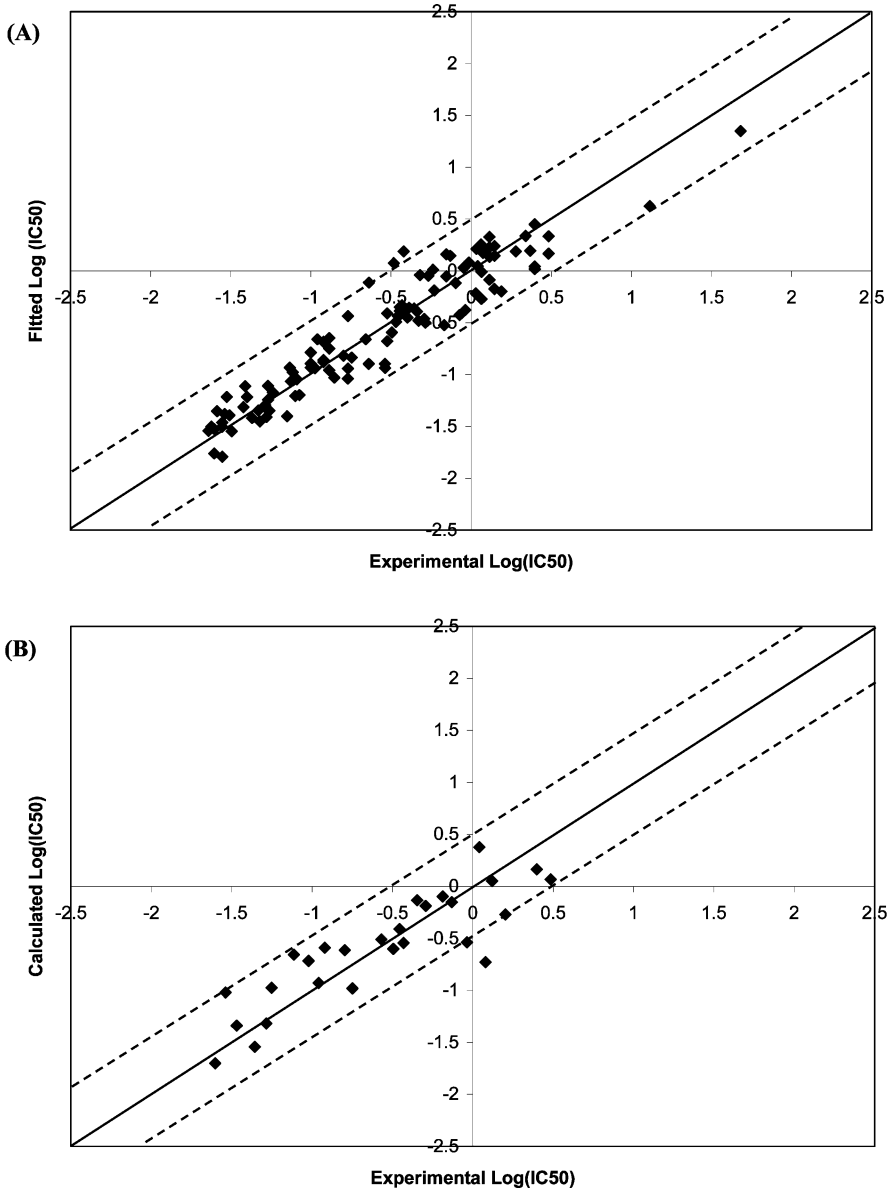


Figure 11. Experimental versus fitted (A, 110 compounds, r^2(LOO) = 0.890) and predicted (B, 27 compounds, r^2(PRESS) = 0.752) bioactivities calculated from the combination QSAR model (eq 1). The solid lines are the regression lines for the fitted and predicted bioactivities of training and test compounds, respectively, whereas the dotted lines indicate the ±0.5 log point error margins.

directions. The ligands were assigned partial charges using the Gasteiger method implemented within CERIUS2. The energy fields were calculated employing the default UNIVERSAL force field (version 1.02) implemented within CERIUS2[75] and were truncated to ±50 kcal/mol. The calculation gave nearly 4290 variables for each compound (1430 variable/probe).

To derive the best possible 3D QSAR statistical model for each docking/scoring combination, we used genetic partial least squares (G/PLS) analysis to search for optimal regression equations capable of correlating the variations in biological activities of the training compounds with variations in the corresponding interaction fields.[77] G/PLS is derived from two methods: genetic function approximation (GFA) and partial least squares (PLS). GFA techniques rely on the evolutionary operations of "crossover and mutation" to select optimal combinations of descriptors (i.e., chromosomes) capable of explaining bioactivity variation among training compounds from a large pool of possible descriptor combinations (i.e., chromosomes population). Each chromosome is associated with a fitness value that reflects how good it is compared to other solutions. The fitness function employed herein is based on Friedman's "lack-of-fit" (LOF).[77]

G/PLS algorithm uses GFA to select appropriate basis functions to be used in a model of the data and PLS regression as the fitting technique to weigh the basis functions' relative contributions in the final model. Application of G/PLS allows the construction of larger QSAR equations while avoiding overfitting and eliminating most variables.[77]

Our preliminary diagnostic trials suggested the following optimal G/PLS parameters: explore linear equations of 18 terms at mating and mutation probabilities of 50%; population size = 500; number of generations (iterations) = 20000 and LOF smoothness parameter = 1.0. However, the optimal number of PLS latent variables (or principle components) was determined for each CoMFA model through assessing the corresponding predictive r^2 (r^2(PRESS)) calculated from the test set of 27 inhibitors that were not included in the training set. These molecules were aligned according to the particular docking/scoring configuration, and their activities were predicted by corresponding G/PLS models generated from the training set (110 compounds) and employing a range of two to seven latent variables. The optimum number of principle components was defined as the one leading to the highest predictive r^2(PRESS) and lowest sum of squared deviations between predicted and actual activity values for every molecule

in the test set (PRESS). Predictive r^2(PRESS) is defined as[77]

$$r^2(\text{PRESS}) = (\text{SD} - \text{PRESS})/\text{SD} \quad (2)$$

where SD is the sum of the squared deviations between the biological activities of the test set and the mean activity of the training set molecules.

All 3D QSAR models were cross-validated employing leave-one-out (LOO) cross-validation and bootstrapping.[73,77] However, high-ranking 3D QSAR models, that is, those corresponding to the best predictive r^2(PRESS) values, were further cross-validated through dividing the training set into five groups of approximately the same size in which the objects were assigned randomly. Subsequently, 80% of the training compounds were randomly selected, and a model was generated, which was then used to predict the remaining compounds within the training set (leave-20%-out). This process was repeated 10 times, and the average predictive r^2 (r^2(L-20%-O)) is determined. This cross-validation technique has been shown to yield better indices for the robustness of a model than the normal LOO procedure.[55,78] An additional validation was also performed for high-ranking 3D QSAR models to rule out the possibility of chance correlations: all biological activities were randomized 99 times (confidence level 99%) and were subjected to regression analysis, and the mean randomization r^2 was calculated.[79]

Acknowledgment. This project was sponsored by the Hamdi-Mango Centre for Scientific Research (Grant Number 04200010). The authors wish to thank the Deanship of Scientific Research and Hamdi-Mango Centre for Scientific Research at the University of Jordan for providing funds toward purchasing O2 and Octane2 Sgi workstations and CERIUS2 software package. The authors are also grateful to Professor M. B. Zughul for his continuous support.

Supporting Information Available: Figures and associated QSAR equations illustrating the optimal CoMFA field points and their statistical relationships to log(IC_{50}) as determined for the 110 training h-PTP 1B inhibitors and showing the relative distribution of optimal field points around inhibitors **119** and **61** in the CoMFA models corresponding to Res1, Wd4, Res24, and Wd12 docking−scoring approaches. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Akamatsu, M. Current state and perspectives of 3D-QSAR. *Curr. Top. Med. Chem.* **2002**, *12*, 1381−1394
(2) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(3) Marshall, G. R.; Cramer, R. D., III. Three-dimensional structure−activity relationships. *Trends Pharmacol. Sci.* **1988**, *9*, 285−289.
(4) Clementi, S.; Wold, S. In *Chemometric Methods in Molecular Design*; Waterbeemd, H., Ed.; VCH: Weinheim, Germany, 1995; pp 319−338.
(5) Wold, S.; Eriksson, L. In *Chemometric Methods in Molecular Design*; Waterbeemd, H., Ed.; VCH: Weinheim, Germany, 1995; pp 309−318.
(6) Kubinyi, H. Variable selection in QSAR studies. I. An evolutionary algorithm. *Quant. Struct.-Act. Relat.* **1994**, *13*, 285−294.
(7) Wold, H. In *Perspectives in Probability and Statistics*; Gani, J., Ed.; Academic Press: London, 1975.
(8) Wold, S. In *Chemometric Methods in Molecular Design*; Waterbeemd, H., Ed.; VCH: Weinheim, Germany, 1995; pp 195−218.
(9) Cramer, R. D., III; Bunce, J. D.; Patterson, D. E. Crossvalidation, bootstrapping and PLS compared with multiple linear regression in conventional QSAR studies. *Quant. Struct.-Act. Relat.* **1988**, *7*, 18−28.
(10) Muegge, I.; Podlogar, B. L. 3D-quantitative structure activity relationships of biphenyl carboxylic acid MMP-3 inhibitors: exploring automated docking as alignment method. *Quant. Struct.-Act. Relat.* **2001**, *20*, 215−222.
(11) Bernard, P.; Kireev, D. B.; Chrétien, J. R.; Fortier, P. L.; Coppet, L. Automated docking of 82 N-benzylpiperidine derivatives to mouse acetylcholinesterase and comparative molecular field analysis with "natural" alignment. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 355−371.
(12) Cho, S. J.; Garsia, M. L.; Bier, J.; Tropsha, A. Structure-based alignment and comparative molecular field analysis of acetylcholinesterase inhibitors. *J. Med. Chem.* **1996**, *39*, 5064−5071.
(13) Vaz, R. J.; McLean, L. R.; Pelton, J. T. Evaluation of proposed modes of binding of (2*S*)-2-[4-[[(3*S*)-1-acetimidoyl-3-pyrrolidinyl]-oxy]phenyl]-3-(7-am idino-2-naphthyl)propanoic acid hydrochloride and some analogues to factor Xa using a comparative molecular field analysis. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 99−110.
(14) Lozano, J. J.; Pastor, M.; Cruciani, G.; Gaedt, K.; Centeno, N. B.; Gago, F.; Sanz, F. 3D-QSAR methods on the basis of ligand−receptor complexes. Application of COMBINE and GRID/GOLPE methodologies to a series of CYP1A2 ligands. *J. Comput.-Aided Mol. Des.* **2000**, *13*, 341−353.
(15) Bernard, P.; Pintore, M.; Berthon, J.-Y.; Chretien, J. R. A molecular modeling and 3D QSAR study of a large series of indole inhibitors of human nonpancreatic secretory phospholipase A2. *Eur. J. Med. Chem.* **2001**, *36*, 1−19.
(16) Pintore, M.; Bernard, P.; Berthon, J.-Y.; Chretien, J. R. Protein-based alignment in 3D QSAR of 26 indole inhibitors of human pancreatic phospholipase A2. *Eur. J. Med. Chem.* **2001**, *36*, 21−30.
(17) Golbraikh, A.; Bernard, P.; Chretien, J. R. Validation of protein-based alignment in 3D quantitative structure−activity relationships with CoMFA models. *Eur. J. Med. Chem.* **2000**, *35*, 123−136.
(18) Wolohan, P.; Reichert, D. E. CoMFA and docking study of novel estrogen receptor subtype selective ligands. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 313−328.
(19) Sippl, W. Development of biologically active compounds by combining 3D QSAR and structure-based design methods. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 825−830.
(20) Hu, X.; Stebbins, C. E. Molecular docking and 3D-QSAR studies of Yersinia protein tyrosine phosphatase YopH inhibitors. *Bioorg. Med. Chem.* **2005**, *13*, 1101−1109.
(21) Moro, S.; Braiuca, P.; Deflorian, F.; Ferrari, C.; Pastorin, G.; Cacciari, B.; Baraldi, P. G.; Varani, K.; Borea, P. A.; Spalluto, G. Combined target-based and ligand-based drug design approach as a tool to define a novel 3D-pharmaco-phore model of human A3 adenosine receptor antagonists: pyrazolo[4,3-e]1,2,4-triazolo[1,5-c] pyrimidine derivatives as a key study. *J. Med. Chem.* **2005**, *48*, 152−162.
(22) Datar P. A.; Coutinho, E. C. A CoMFA study of COX-2 inhibitors with receptor based alignment. *J. Mol. Graphics Modell.* **2004**, *23*, 239−251.
(23) Morris, G. M.; Olson, A. J.; Goodsell, D. S. Protein−Ligand docking. *Methods Princ. Med. Chem.* **2000**, *8*, 31−48.
(24) Mestres, J.; Knegtel, R. M. A. Similarity versus docking in 3D virtual screening. *Perspect. Drug Discovery Des.* **2000**, *20*, 191−207.
(25) Vieth, M.; Hirst, J. D.; Dominy, B. N.; Daigler, H.; Brooks, C. L., III. Assessing search strategies for flexible docking. *J. Comput. Chem.* **1998**, *19*, 1623−1631.
(26) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys. J.* **1997**, *72*, 1047−1069.
(27) Monard, G.; Merz, K. M., Jr. Combined Quantum Mechanical/Molecular Mechanical Methodologies Applied to Biomolecular Systems. *Acc. Chem. Res.* **1999**, *32*, 904−911.
(28) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of Docking Performance: Comparative Data on Docking Algorithms. *J. Med. Chem.* **2004**, *47*, 558−565.
(29) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.
(30) Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.- Aided Mol. Des.* **2001**, *15*, 411−428.
(31) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.
(32) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.
(33) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.

(34) Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graphics Modell*. **2003**, *21*, 289−307.
(35) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.
(36) Wang, R.; Lu, Y.; Wang, S. Comparative Evaluation of 11 Scoring Functions for Molecular Docking. *J. Med. Chem.* **2003**, *46*, 2287−2303.
(37) Krovat, E. M.; Langer, T. Impact of Scoring Functions on Enrichment in Docking- Based Virtual Screening: An Application Study on Renin Inhibitors. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1123−1129.
(38) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.
(39) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. Molecular recognition of the inhibitor AG-1343 by HIV-1 Protease: Conformationally flexible docking by evolutionary programming. *Chem. Biol.* **1995**, *2*, 317−324.
(40) Gehlhaar, D. K.; Bouzida, D.; Rejto, P. A. In *Rational Drug Design: Novel Methodology and Practical Applications*; Parrill, L., Reddy, M. R., Eds.; American Chemical Society: Washington, DC, 1999; pp 292−311.
(41) Böhm, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein−ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des*. **1994**, *8*, 243−256.
(42) Böhm, H. J. Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 309−323.
(43) Wang, R.; Gao, Y.; Lai, L. SCORE: A new empirical method for estimating the binding affinity of a protein−ligand complex. *J. Mol. Model.* **1998**, *4*, 379−394.
(44) Wang, R.; Lai, L.; Wang, S. Further development and validation of empirical scoring functions for structure-based binding validation affinity prediction. *J. Comput.-Aided Mol. Des*. **2002**, *16*, 11−26.
(45) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein−ligand interactions: A simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.
(46) Muegge, I. A knowledge-based scoring function for protein−ligand interactions: Probing the reference state. *Perspect. Drug Discovery Des.* **2000**, *20*, 99−114.
(47) Muegge, I. Effect of ligand volume correction on PMF scoring. *J. Comput. Chem.* **2001**, *22*, 418−425.
(48) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.
(49) Ishchenko, A. V.; Shakhnovich, E. I. Small molecule growth 2001 (SMoG2001): An improved knowledge-based scoring function for protein−ligand interactions. *J. Med. Chem.* **2002**, *45*, 2770−2780.
(50) Tame, J. R. H. Scoring functions: a view from the bench. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 99−108.
(51) Kollman, P. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.* **1993**, *93*, 2395−2417.
(52) Schulz-Gasch, T.; Stahl, M. Scoring Functions for Protein−Ligand Interactions: A Critical Perspective. *DTT: Technol.* **2004**, *1*, 231−239.
(53) Klebe, G. Lead Identification in Post-Genomics: Computers as a Complementary Alternative. *DTT: Technol.* **2004**, *1*, 225−230.
(54) Koehler, K. F.; Rao, S. N.; Snyder, J. P. In *Guidebook on Molecular Modeling in Drug Design*; Cohen, N. C., Ed.; Academic Press: San Diego, CA, 1996; pp 253−255.
(55) Pastor, M.; Cruciani, G.; Watson, K. A Strategy for the Incorporation of Water Molecules Present in a Ligand Binding Site into a Three-Dimensional Quantitative Structure−Activity Relationship Analysis. *J. Med. Chem.* **1997**, *40*, 4089−4102.
(56) Silverman, R. A. In *The Organic Chemistry of Drug Design and Drug Action*; Academic Press: San Diego, CA, 1991; pp 62−65.
(57) Poornima, C. S.; Dean, P. M. Hydration in drug design. 1. Multiple hydrogen-bonding features of water molecules in mediating protein−ligand interactions. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 500−512.
(58) Poornima, C. S.; Dean, P. M. Hydration in drug design. 2. Influence of local site surface shape on water binding. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 513−520.
(59) Poornima, C. S.; Dean, P. M. Hydration in drug design. 3. Conserved water molecules at the ligand- binding sites of homologous proteins. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 521−531.
(60) Malamas, M. S.; Sredy, J.; Moxham, C.; Katz, A.; Xu, W.; McDevitt, R.; Adebayo, F. O.; Sawicki, D. R.; Seestaller, L.; Sullivan, D.; Taylor, J. R. Novel Benzofuran and Benzothiophene Biphenyls as Inhibitors of Protein Tyrosine Phosphatase 1B with Antihyperglycemic Properties. *J. Med. Chem.* **2000**, *43*, 1293−1310.
(61) Malamas, M. S.; Sredy, J.; Gunawan, I.; Mihan, B.; Sawicki, D. R.; Seestaller, L.; Sullivan, D.; Flam, B. R. New Azolidinediones as Inhibitors of Protein Tyrosine Phosphatase 1B with Antihyperglycemic Properties. *J. Med. Chem.* **2000**, *43*, 995−1010.
(62) Johnson, T. O.; Ermolieff, J.; Jirousek, M. R. Protein Tyrosine Phosphatase 1b Inhibitors for Diabetes. *Nat. Rev. Drug Discovery* **2002**, *1*, 696−709.
(63) *CERIUS2 LigandFit User Manual*; Accelrys Inc.: San Diego, CA, 2000; pp 3−48.
(64) Kontoyianni, M.; Sokol, G. S.; Mcclellan, L. M. Evaluation of Library Ranking Efficacy in Virtual Screening. *J. Comput. Chem.* **2004**, *26*, 11−22.
(65) Krammer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C. M.; Waldman, M. LigScore: a novel scoring function for predicting binding affinities. *J. Mol. Graphics Modell.* **2005**, *23*, 395−407.
(66) Murthy, V. S.; Kulkarni, V. M. 3D-QSAR CoMFA and CoMSIA on Protein Tyrosine Phosphatase 1B Inhibitors. *Bioorg. Med. Chem*. **2002**, *10*, 2267−2282.
(67) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.
(68) Terp, G. E.; Johansen, B. N.; Christensen, I. T.; Jorgensen, F. S. A new concept for multidimensional selection of ligand conformations (MultiSelect) and multidimensional scoring (MultiScore) of protein−ligand binding affinities. *J. Med. Chem.* **2001**, *44*, 2333−2343.
(69) Paul, N.; Rognan, D. ConsDock: A new program for the consensus analysis of protein−ligand interactions. *Proteins: Struct., Funct., Genet.* **2002**, *47*, 521−533.
(70) Beeley, N. R. A.; Sage, C. GPCRs: an update on structural approaches to drug discovery. *Targets* **2003**, *2*, 19−25.
(71) Waszkowycz, B. In *Advances in Drug Discovery Techniques*; Harvey, A. L., Ed.; John Wiley & Sons: Chchester, U.K., 1998; pp 150−153.
(72) Drew, M. G. B.; Lumley, N. R.; Price, N. R.; Watkins, R. W. In *Proceedings of the 12th European Symposium on Quantitative Structure−Activity Relationships*: *Molecular modeling and Prediction of Bioactivity*; Gundertofte, K., Jørgensen F. S., Eds.; Kluwer Academic/Olenum Publishers: New York, 1998; pp 453−454.
(73) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *Quant. Struct.−Act. Relat. Comb. Sci.* **2003**, *22*, 69−77.
(74) Gasteiger J.; Marsili, M. A new model for calculating atomic charges in molecules. *Tetrahedron Lett.* **1978**, *34*, 3181−3184.
(75) *CERIUS2 OFF*; Accelrys Inc.: San Diego, CA, 1997; pp 5−109.
(76) *CERIUS2 4.8.1 QSAR*; Accelrys Inc.: San Diego, CA, 2003; pp 161−171.
(77) *CERIUS2 4.8.1 QSAR*; Accelrys Inc.: San Diego, CA, 2003; pp 210−235.
(78) Sippl, W. Receptor-based 3D QSAR analysis of estrogen receptor ligands-merging the accuracy of receptor-based alignments with the computational efficiency of ligand-based methods. *J. Comput.-Aided Mol. Des*. **2000**, *14*, 559−572.
(79) Clark, M.; Cramer, R. D. The probability of chance correlation using partial least squares (PLS). *Quant. Struct.−Act. Relat.* **1993**, *12*, 137−145.

JM058047O