An Updated Steroid Benchmark Set and Its Application in the Discovery of Novel Nanomolar Ligands of Sex Hormone-Binding Globulin

Artem Cherkasov,*[,†] Fuqiang Ban,[†] Osvaldo Santos-Filho,[†] Nels Thorsteinson,[†] Magid Fallahi,[‡] and Geoffrey L. Hammond[‡]

Prostate Centre at the Vancouver General Hospital, Faculty of Medicine, University of British Columbia, Vancouver, British Columbia, and Department of Obstetrics and Gynecology, University of British Columbia, Child & Family Research Institute, Vancouver, British Columbia

Received September 13, 2007

A benchmark data set of steroids with known affinity for sex hormone-binding globulin (SHBG) has been widely used to validate popular molecular field-based QSAR techniques. We have expanded the data set by adding a number of nonsteroidal SHBG ligands identified both from the literature and in our previous experimental studies. This updated molecular set has been used herein to develop 4D QSAR models based on "inductive" descriptors and to gain insight into the molecular basis of protein–ligand interactions. Molecular alignment was generated by means of docking active compounds into the active site of the SHBG. Surprisingly, the alignment of the benchmark steroids contradicted the classical ligand-based alignment utilized in previous CoMFA and CoMSIA models yet afforded models with higher statistical significance and predictive power. The resulting QSAR models combined with CoMFA and CoMSiA models as well as structure-based virtual screening allowed discovering several low-micromolar to nanomolar nonsteroidal inhibitors for human SHBG.

Introduction

The blood of vertebrates contains two high-affinity steroid-binding proteins, known as sex hormone-binding globulin (SHBG)[a] and corticosteroid-binding globulin (CBG), whose steroid-binding characteristics have been studied extensively.[1] A group of compounds known to bind these proteins form a popular "steroid benchmark set" utilized in many *in silico* modeling studies[2–11] including popular CoMFA[12] and CoMSiA[13] 3D QSAR methods.

In a series of previous reports, we have investigated the SHBG system and identified various nonsteroidal ligands using several innovative *in silico* screening methods.[17–19] Furthermore, we tested the suggested lead compounds experimentally with tritium-labeled 5α-dihydrotestosterone in a competitive ligand-binding assay. For the purpose of the current study, we have combined the available data on known SHBG ligands (both steroidal and nonsteroidal) and formed an expanded set of 84 compounds (shown in Table 1). This updated benchmark set has been used to develop various QSAR solutions enabling the discovery of nonsteroidal SHBG binders.

Results

We have considered the 84 SHBG ligands (including 21 steroids present in the original benchmark set) as the training set and docked all molecules into the SHBG active site using the Glide program with the default settings of the Extra Precision mode,[20] as in our previous SHBG studies.[17–19] For this purpose, the structure of the protein with cocrystallized ligand 5α-androstane-3β,17α-diol corresponding to the 1LHN entry of the Protein Databank was preoptimized with the MMFF force field.[21] The ligand was then removed, and the protein structure was used in the self-docking analysis, which demonstrated that the crystallographic pose of the bound steroid could be accurately reproduced (Figure 1).

Importantly, the Extra Precision Glide docking protocol reproduced the optimal orientation of androgens and estrogens in the SHBG steroid-binding site, in accordance with recent crystallographic and mutation experiments: specifically, C18 estrogens and C19 androgens have been shown to reside within the SHBG steroid-binding site predominantly in opposite orientations.[14,15] Thus, while a critical ligand-anchoring residue Ser42 in human SHBG[14] coordinates the 17β-hydroxyl of estrogens, the same residue forms a hydrogen bond with functional groups at the C3 position of androgens. This is illustrated in Figure 2, which shows the positions of 5α-dihydrotestosterone (green) and estradiol (yellow) identified in the structures of human SHBG cocrystallized with these steroids (Protein Databank[16] entries 1KDM and 1LHU, respectively).

Notably, these crystallography-derived orientations of estrogens and androgens within the human SHBG steroid-binding site (confirmed by our docking experiments) differ from the field-similarity based alignment of SHBG ligands (Figure 1B) used in the original CoMFA[12] and CoMSiA[13] studies, as well as in all subsequent QSAR reports involving the steroid benchmark set.[2–11]

Out of 84 docked compounds, nine estrogen derivatives (i.e., C18-steroids containing aromatic ring "A"), i.e., estradiol, estriol, estrone, 2-iodoestradiol, 2-methoxyestradiol, equilenin, equilin, 17-deoxyestrone, and estradiol 3-benzoate, all favored a binding pose allowing the coordination of functional groups at C17 with the Ser42 side chain. In addition, one C19 steroid, etiocholanolone, was also docked in such orientation that may be attributed to some structural features of the compound (such as an unusual bending angle of a scaffold) or an artifact of the docking experiment. Otherwise, all other non-estrogen derivatives (compounds without aromatic "A" steroidal ring) demonstrated an opposite docking orientation corresponding to Ser42 anchoring functional groups at C3 of the steroid scaffold.

* To whom correspondence should be addressed. Tel: 604.875.4111 × 69628. Fax: 604.875.5654. E-mail: artc@interchange.ubc.ca.

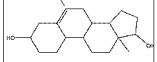
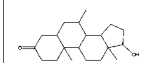
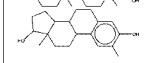
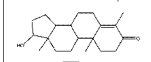
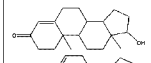
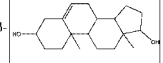
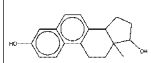
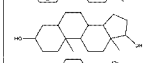
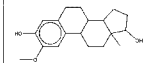
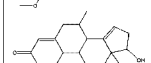
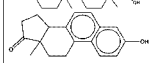
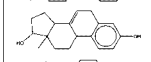
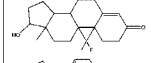
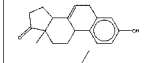
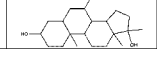
[†] Prostate Centre at the Vancouver General Hospital.

[‡] Department of Obstetrics and Gynecology.

[a] Abbreviations: QSAR, quantitative structure–activity relationships; CoMFA, comparative molecular field analysis; CoMSiA, comparative molecular similarity index analysis; PLS, partial least-squares, SHBG, sex hormone-binding globulin; CBG, corticosteroid-binding globulin; GFA, genetic function approximation.

10.1021/jm7011485 CCC: $40.75 © 2008 American Chemical Society
Published on Web 03/11/2008

Table 1. Known SHBG Binders Utilized for QSAR Modeling (Training Set) and Novel Nonsteroidal Ligands Identified in the Current Study Are Presented along with the Corresponding Protein–Ligand Dissociation Parameters pK_d and Predicted Target Affinities Produced by QSAR and Virtual Screening Approaches

#	Name	Structure	pK_d	I	II	III	IV	V	VI	VII	VIII
	TRAINING SET										
1	5α-Dihydrotestosterone[#]		**9.74**	8.97	8.67	9.71	9.29	9.57	8.79	-13.23	-14.8
2	17-ethinyl-Dihydrotestosterone		**9.74**	9.24	9.37	7.94	9.38	9.03	8.70	-12.8	-12.52
3	1α-methyl-Dihydrotestosterone		**9.60**	8.87	8.76	8.95	9.62	9.09	8.99	-13.75	-15.04
4	5-Androstene-19-nor-3b,17β-diol		**9.54**	8.68	8.84	8.67	8.69	8.88	8.66	-13.38	-13.78
5	7α-methyl-Dihydrotestosterone		**9.38**	8.97	8.76	8.94	9.65	9.07	9.00	-13.69	-15.11
6	2-Iodo Estradiol		**9.32**	8.68	7.46	6.58	9.62	7.24	9.33	-10.56	-11.14
7	4-methyl-Testosterone		**9.31**	9.23	8.94	8.86	9.36	9.14	9.54	-14.14	-15.23
8	Testosterone[#]		**9.20**	8.60	8.65	9.23	9.23	9.23	8.97	-13.89	-14.94
9	5α-Androstene-3β,17β-diol[#]		**9.17**	8.76	8.47	9.24	9.34	9.30	9.04	-13.28	-13.4
10	Dihydroequilenin-17β		**9.12**	8.65	8.14	10.40	9.05	9.05	8.42	-13.72	-13.36
11	5α-Androstane-3α,17β-diol[#]		**9.11**	8.85	8.76	9.10	9.64	9.24	9.25	-14.07	-13.36
12	2-Methoxy Estradiol		**9.08**	8.49	7.70	9.73	9.26	8.98	9.02	-12.45	-12.06
13	7α-methyl-14-dehydro-19-Nortestosterone		**9.07**	8.22	8.26	8.59	8.40	8.90	8.99	-12.98	-14.09
14	6α-fluoro-Dihydrotestosterone		**9.05**	8.84	8.33	8.43	8.82	9.19	8.98	-12.23	-13.92
15	7α-17-dimethyl-Dihydrotestosterone		**9.05**	7.89	7.34	8.79	9.10	8.86	9.40	-14.25	-14.86
16	7α-methyl-5α-Androstane-3β,17β-diol		**9.00**	7.52	7.45	8.48	8.77	9.13	8.95	-12.69	-13.89
17	4-Androstene-3β,17β-diol		**9.00**	8.85	8.86	9.39	9.02	9.01	9.05	-13.84	-15.19
18	17-ethinyl-delta, 5-Androstane		**8.91**	8.18	8.57	8.63	8.53	8.73	8.68	-13.97	-14.24
19	d-Norgestrel		**8.91**	8.44	7.35	7.76	7.89	8.75	7.92	-12.72	-13.39
20	Estradiol[#]		**8.83**	7.14	7.48	8.85	8.35	8.83	7.70	-13.74	-14.29
21	17-methyl-Dihydrotestosterone		**8.81**	7.58	8.21	8.23	9.03	9.01	9.13	-13.56	-14.67
22	17-ethinyl-11-methylene-18-methyl-19-nor-Dihydrotestosterone		**8.80**	8.66	8.33	8.40	8.69	9.23	8.82	-12.61	-12.96
23	Ethisterone		**8.78**	7.83	8.12	8.13	8.60	8.79	8.42	-13.99	-14.01
24	7α,17-dimethyl-Testosterone		**8.76**	8.37	8.57	8.00	8.85	8.96	8.88	-13.81	-12.43
25	6-dehydro-Estradiol		**8.76**	8.45	8.43	10.03	9.34	9.16	8.47	-13.51	-13.9
26	7a-methyl-Testosterone		**8.71**	8.75	8.74	8.79	8.40	9.05	9.06	-13.47	-14.34
27	Equilenin		**8.62**	8.48	8.43	10.27	8.27	9.06	8.92	-12.46	-10.99
28	7-dehydro-Estradiol		**8.62**	8.68	9.06	7.45	8.19	7.56	8.52	-13.51	-13.9
29	17-methyl-1-Dihydrotestosterone		**8.57**	7.62	7.95	8.03	8.84	9.16	8.80	-14.03	-12.95
30	9a-fluoro-Testosterone		**8.51**	9.59	8.11	8.99	8.77	8.65	8.36	-13.71	-14.22
31	Equilin		**8.51**	8.75	9.15	7.57	8.51	7.75	8.15	-11.57	-11.2
32	7α,17-dimethyl-5a-Androstane-3β,17β-diol		**8.46**	8.28	9.06	7.47	8.04	9.18	8.56	-13.22	-12.03

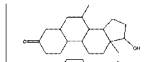
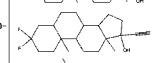
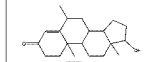
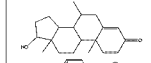
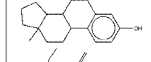
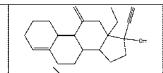
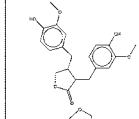
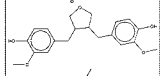
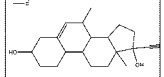
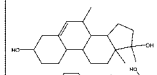
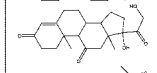
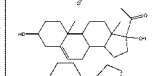
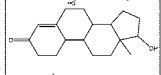
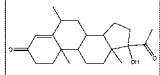
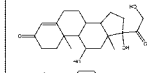
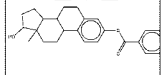
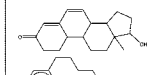
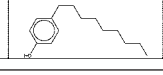
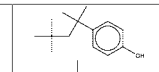
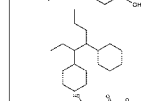
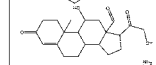
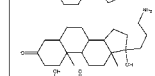
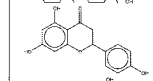
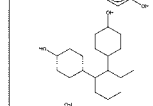
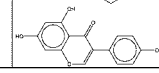
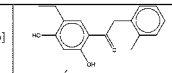
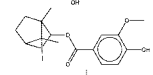
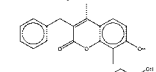
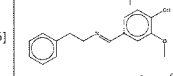
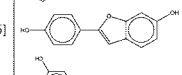
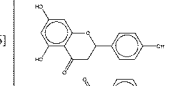
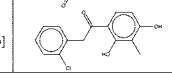
#	Name	Structure	pK_d	I	II	III	IV	V	VI	VII	VIII
	TRAINING SET										
33	7α-methyl-19-nor-Dihydrotestosterone		**8.46**	8.23	8.04	9.26	7.86	9.54	7.91	-12.75	-12.01
34	17-methyl-Testosterone		**8.43**	8.32	8.47	7.97	8.42	8.92	8.79	-13.85	-12.29
35	17-ethinyl-3,3-difluoro-5a-Androstan-17β-ol		**8.36**	7.43	6.69	8.49	8.53	9.16	9.10	-12.18	-11.9
36	6α-methyl-Testosterone		**8.36**	7.24	8.31	8.11	8.47	9.09	8.40	-11.74	-12.27
37	19-Nor-Dihydrotestosterone		**8.36**	8.12	8.66	9.08	7.70	9.42	8.31	-11.8	-13.11
38	7α-methyl-1-dehydro Testosterone		**8.36**	8.81	8.36	8.80	7.61	9.43	7.83	-13.32	-13.03
39	17-methyl-delta-5-Androstane		**8.36**	9.06	8.31	8.63	8.57	9.18	8.39	-13.68	-11.91
40	17-Deoxoestrone		**8.30**	7.42	7.77	8.61	8.67	8.65	8.33	-12.4	-13.65
41	3α-hydroxy-5α-H-17-ethinil-11-methylene-18-methyl-19-nor-Dihydrotestosterone		**8.27**	8.07	7.93	7.86	8.38	8.39	8.40	-12.38	-13.09
42	9α-fluoro-11-oxo-17-methyl-Testosterone		**8.23**	7.88	7.83	7.23	8.37	8.04	7.60	-13.1	-12
43	17-ethinyl-11-methylene-18-methyl-19-nortestosterone		**8.23**	8.34	8.41	7.60	8.71	8.71	8.25	-12.31	-13.59
44	Danazol		**8.20**	8.51	8.97	7.68	8.14	8.57	8.12	-12.85	-8.94
45	Estrone[#]		**8.18**	7.71	8.38	8.16	8.26	8.14	7.46	-12.8	-11.43
46	17-ethinyl-18-methylestra-4,9,13-trien-17β-ol-3-one		**8.11**	7.85	8.39	8.54	7.87	8.88	7.91	-13.34	-13.62
47	Norethindrone		**7.97**	7.91	7.93	8.50	7.92	8.86	7.39	-13.02	-13.83
48	16α-hydroxy Testosterone		**7.92**	7.83	7.59	7.49	7.80	6.03	8.14	-12.44	-13.83
49	3b-hydroxy-17-ethinyl-11-methylene-18-methyl-4-Estren-17β-ol		**7.88**	7.73	8.66	6.87	7.66	7.83	7.43	-8.86	-11.42
50	Dehydroepiandrosterone[#]		**7.84**	8.52	7.53	7.77	7.94	7.69	8.05	-12.53	-1.33
51	3α-hydroxy-17-ethinyl-11-methylene-18-methyl-4-Estren-17β-ol		**7.67**	7.86	7.67	8.90	7.25	9.45	8.26	-11.26	-12.87
52	1α-aminohexyl-17β-hydroxy-5α-Androstan-3-one		**7.53**	8.23	7.66	7.46	7.56	7.44	7.97	-11.53	-12.05
53	Androstenedione[#]		**7.46**	8.01	7.63	8.17	6.67	8.70	6.81	-11.95	-10.7
54	Deoxycortisol[#]		**7.44**[$]	7.51	7.72	7.22	6.77	7.14	7.40	-12.86	-13.29
55	Deoxycorticosterone[#]		**7.38**	6.47	6.92	7.40	7.39	7.29	7.65	-12.19	-12.14
56	Pregnenolone[#]		**7.15**	7.90	7.53	7.03	6.79	6.91	6.80	-11.38	-11.5
57	Androsterone[#]		**7.15**	7.33	6.39	7.18	6.96	7.20	7.64	-13.98	-12.45
58	17-hydroxy-Progesterone[#]		**7.00**	7.37	6.77	6.86	6.65	6.87	6.98	-11.66	-13.29
59	Progesterone[#]		**6.94**	8.59	7.60	6.99	7.09	7.20	7.49	-10.72	-12.59
60	17β-hydroxy-2,2,17-trimethylestra-4,9,11-trien-4-one		**6.86**	5.94	5.86	7.84	6.15	7.64	6.22	-11.13	-9.57
61	17-ethinylestradiol		**6.81**	7.58	7.34	7.87	7.35	8.70	7.07	-13.15	-13.24
62	Estriol[#]		**6.63**	7.45	8.02	6.62	7.16	6.61	7.43	-11.92	-12.57

Table 1. Continued

#	Name	Structure	pK_d	I	II	III	IV	V	VI	VII	VIII
	TRAINING SET										
63	17-ethinyl-11-methylene-18-methyl-4-Estren-17β-ol		**6.60**	6.47	8.00	6.41	6.65	7.16	6.65	-12.78	-13.22
64	(-)-Matairesinol		**6.51**	7.95	8.66	7.52	6.67	8.53	7.15	-10.37	-9.88
65	3,4-divanillyl Tetrahydrofuran		**6.51**	7.82	7.15	7.49	6.72	8.45	6.78	-10.99	-10.36
66	7α-methyl-17-ethinyl-delta5E		**6.44**	6.54	6.63	7.53	6.35	6.76	5.93	-12.15	-13.16
67	7α,17-dimethyl-delta5E		**6.44**	6.17	6.37	6.55	6.41	6.57	6.23	-12.44	-12.08
68	Cortisone[e]		**6.43**	8.10	6.67	6.43	6.37	6.45	6.20	-10.03	-10.07
69	17-hydroxy-5-Pregnen-3b-ol-20-one[g]		**6.36**	6.27	7.92	9.90	7.91	9.37	8.06	-11.63	-9.71
70	Corticosterone[f]		**6.34**	7.47	6.55	6.35	6.73	6.42	7.05	-10.37	-9.05
71	19-Nor-Testosterone		**6.30**	6.04	6.11	6.18	6.34	6.22	5.98	-12.21	-13.28
72	17-hydroxy-6α-methyl Progesterone		**6.20**	8.42	7.97	6.14	5.85	6.10	6.19	-9.68	-11.45
73	Cortisol[g]		**6.20**	5.75	6.55	8.43	6.40	8.74	5.89	-9.91	-11.18
74	Etiocholanolone[g]		**6.15**	6.96	7.04	7.07	5.99	7.74	6.03	-13.85	-12.12
75	Estradiol-3-benzoate		**5.94**	5.70	6.00	8.95	7.18	8.89	7.91	-12.56	-14.37
76	6-dehydro-19-Nortestosterone		**5.94**	5.84	6.13	8.36	5.79	8.39	5.45	-13.57	-3.73
77	4-Nonylphenol		**5.92**	5.39	6.94	6.82	5.91	8.52	5.80	-9.97	-8.52
	TRAINING SET										
78	4-tert-Octylphenol		**5.67**	5.86	6.59	8.18	5.70	9.19	5.98	-8.95	-9.97
79	3,4-dicyclohexyl-hexane (meso)		**5.61**	7.62	7.95	8.58	5.91	7.74	6.00	-9.45	-10.27
80	Aldosterone[h]		**5.32**	5.35	5.14	5.34	5.58	5.34	5.12	-10.39	-8.33
81	17α-aminopropyl-17βb-hydroxy-5α-Androstan-3-one		**4.96**	5.66	5.50	7.22	4.54	6.82	5.05	-12.96	-11.98
82	Naringenin		**4.55**	4.71	4.28	7.70	4.73	7.45	4.63	-10.57	-10.69
83	3,4,-di [(4-hydroxyl)cyclohexyl] hexane (meso)		**4.54**	6.04	7.62	7.28	4.76	8.96	4.99	-11.51	-11.26
84	Genistein		**4.40**	4.72	4.98	8.40	4.50	7.36	4.16	-10.78	-11.04
	SHBG LIGANDS IDENTIFIED IN THE CURRENT STUDY										
85	ZINC00389056 [36]		**6.97**	8.01	5.18	8.52	6.09	7.63	6.97	-12.84	-11.59
86	ZINC00073647 [36]		**6.39**	6.95	5.42	9.24	10.24	8.90	8.19	-12.14	-12.76
87	ZINC00407192 [36]		**5.66**	5.80	6.14	8.08	7.76	8.38	7.67	-12.59	-12.59
88	ZINC02819939 [36]		**6.23**	5.80	5.56	9.16	10.17	9.06	7.34	-11.23	-11.59
89	ZINC00334865 [36]		**6.08**	5.76	6.60	8.37	4.56	8.24	4.40	-11.53	-12.14
90	ZINC00001785 [36]		**6.02**	5.88	6.15	8.25	4.38	7.83	3.83	-11.2	-10.08
91	ZINC00457465 [36]		**5.66**	8.97	3.56	8.21	7.24	8.47	7.73	-12.32	-12.41

** These four entries could not be scored by in silico approaches, as they failed to dock into the 1LHN active site. # Twenty-one steroids forming the original "steroid benchmark set". $ We have used the corrected pK = 7.44 value for deoxycortisol instead of pK = 7.20 from ref 18. The columns labeled with Roman numbers contain predictions by various QSAR models: (I) predictions by QuaSAR-Evolution model created on the basis of the updated dataset; (II) GFA-QSAR model trained on the updated dataset (84 entries); (III) CoMFA model trained on the original dataset (21 entries); (IV) CoMFA model trained on the updated dataset (84 entries); (V) CoMSiA model trained on the original dataset (21 entries); (VI) CoMSiA model trained on the updated dataset (84 entries); (VII) results of 1LHN XP-docking with Glide; (VIII) results 1KDM XP-docking with Glide. Identification codes for compounds **85–91** correspond to internal IDs of the ZINC molecular database.[36]

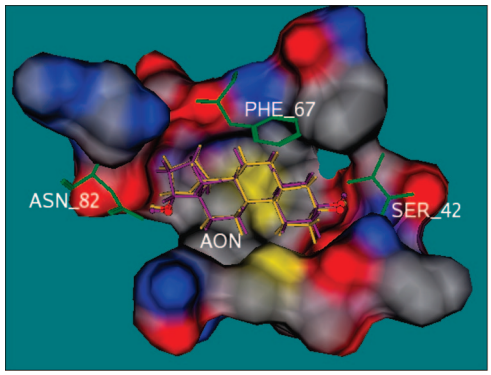


Figure 1. Superposition of a native ligand from 1LHN protein structure (colored in maroon) with the docking pose of 5α-androstane-3β,17α-diol (AON) established by extra precision Glide protocol (in gold).

In addition to the correct identification of different binding modes for C18 estrogens and C19 androgens within the SHBG active site, the virtual screening protocol we have utilized produced docking scores that generally corresponded to the experimental SHBG binding constants. The resulting linear dependences with $r^2 = 0.34$ for the entire set of 84 molecules, and $r^2 = 0.48$ for the 21 benchmark steroids alone, are shown in Figure 3 (it should be noted that such modest correlation coefficients are common for docking scoring functions; for instance see ref 22).

Thus, we conclude that the extra precision docking of known SHBG ligands resulted in binding poses that generally reflect the preferred orientations of specific steroid classes in the crystal structures. Using the resulting docking poses of the canonical benchmark set of 21 steroids, as well as docking poses of the updated set of 84 SHBG ligands, we then created CoMFA and CoMSiA models along with several additional QSAR solutions for *in silico* lead discovery.

CoMFA Models. As described above, we adopted the docking poses of SHBG ligands as the basis for their structural alignment. Using this alignment (which takes into account opposite direction of SHBG binding for C18 and C19 steroids), we assembled two training sets: corresponding to the original

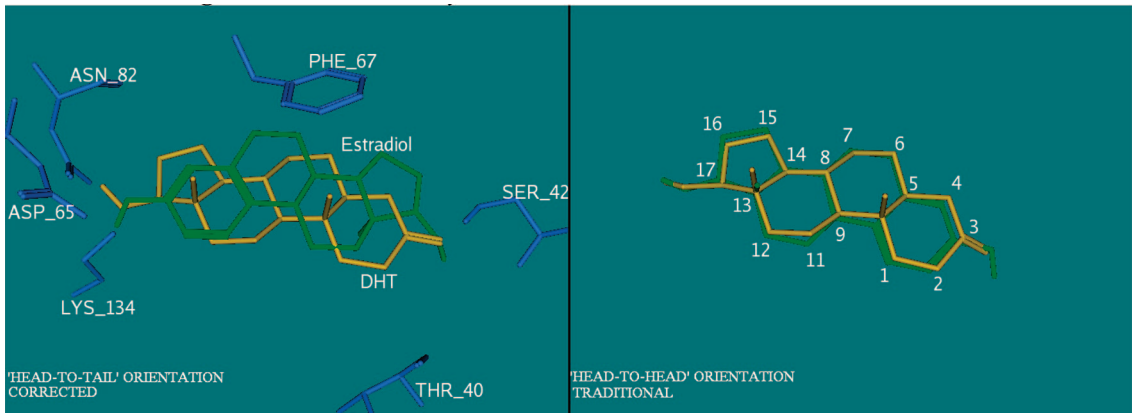


Figure 2. Optimal (left panel) and traditional (right panel) orientations of 5α-dihydrotestosterone (DHT) shown in gold and estradiol shown in green. The correct superposition of the compounds within the human SHBG steroid-binding site was derived from the 1KDM and 1LHU crystal structures. The traditional alignment was obtained by SYBIL.

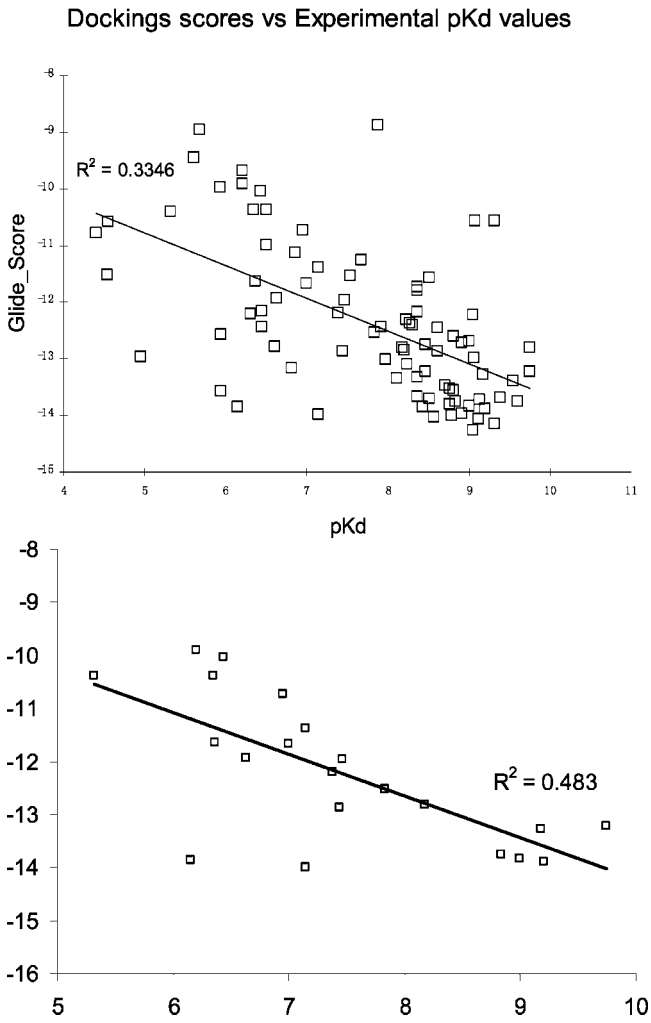


Figure 3. Top panel: the linear dependence between GlideScore values estimated by extra precision docking of 84 trainng set compounds and the corresponding pK_d experimental values (three outliers have been removed). Bottom panel: the same dependence limited to 21 benchmerk steroids.

benchmark set of 21 steroids and an updated set including all 84 SHBG binders.

Utilization of the 1LHN-docking poses of 21 benchmark steroids resulted in a CoMFA model very similar in its statistical characteristics to that in the original study[12] (training statistics are also shown in Table 2). This striking similarity in the

Table 2. Training and Testing Statistics for Computational Models Created and Their Combinations Investigated in the Current Study

models	r^2	q^2_{LOO}	training set **EF**
QSAR_GFA_7.0Å	0.56	0.44	3.5
QuaSAR_Evolution_7.0Å	0.66	0.58	3.5
CoMFA_corrected_21	0.99	0.45	3.0
CoMFA_traditional_21	0.98	0.53	2.0
CoMFA_84	0.99	0.41	5.5
CoMSiA_corrected_21	0.99	0.51	2.0
CoMSiA_traditional_21	0.98	0.53	1.0
CoMSiA_84[a]	0.91	0.49	5.0
1kdm_XP_GlideScore		2.5	
1lhn_XP_GlideScore		2.0	
2Dock_CoMFA84_CoMSiA84_GFA_QuaSAR		5.0	
2Dock_CoMFA84_GFA		5.0	
2Dock_CoMSiA84_GFA		4.0	
2Dock_CoMFA84_CoMSiA84_GFA		5.0	

[a] Three outliers have been removed when training the model. "Corrected" notion reflects alignment of SHBG ligands based on the docking poses. "traditional" notion corresponds to steroid scaffold alignments used in the original CoMFA and CoMSiA studies and based on maximal molecular field similarity.

statistical parameters of CoMFA models irrespective of drastic differences between underlying alignment underscores an apparent (and perhaps unexpected) insensitivity of CoMFA to molecular alignment's nuances. Obviously, the two models, i.e., the "classical"[12] and the one developed here using a different (based of docking orientations) compound alignment, are associated with completely different steric and electrostatic fields suggesting quite different avenues for structural modifications that should putatively lead to more active compounds. This ambiguity of CoMFA model interpretation of statistically indistinguishable alternative models should be kept in mind as a potential source of misleading hypotheses concerning novel compound design.

The CoMFA model created on the basis of the expanded set of 84 SHBG binders produced very similar training r^2 and q^2 values, but allowed 1.8-fold better recovery of the most active compounds from the training set. The corresponding enrichment factor (EF) values calculated with "top 15% hit-list" criteria applied to the predicted training set values are also given in Table 2 (more details on the calculation of enrichment factors EF can be found in the Materials and Methods).

CoMSiA Analysis of the Data Sets. We also utilized the training sets of 21 and 84 superimposed molecules to create CoMSiA models and conduct comparative analysis of their accuracy and enrichment performance. Using the corresponding

sets of aligned structures, we computed the standard CoMSiA fields (steric, electrostatic, hydrophobic, hydrogen bond donor, and hydrogen bond acceptor) and created PLS[23]-based solutions.

On one hand, the results from Table 2 indicate the CoMSiA solutions derived from the original benchmark and expanded sets of SHBG ligands have similar training accuracy with r^2 in 0.91–0.99 range and q^2 neighboring 0.5. On the other hand, as in the case of CoMFA models, the CoMSiA solution created on the basis of the expanded set of SHBG ligands allowed much better enrichment of the training set (EF = 5.0) when compared to the model customized for 21 highly similar original benchmark structures (EF = 2.0). Notably, the use of crystallography-complying and traditional, field similarity-based alignments of 21 benchmark steroids, as in case of CoMFA, did not result in substantially different QSAR models (both solutions are featured in Table 2).

Application of the LFER Principle to Protein–Ligand Interactions Using 4D "Inductive" Descriptors. Previously, we have developed descriptors (called "inductive") that were successfully adopted for QSAR modeling of SHBG ligands.[19–21] These "inductive" 3D-QSAR solutions have been derived from the LFER (linear free energy relationship)-based equations for inductive and steric substituents parameters (see ref 24 for more details)

$$R_{s_{G\to j}} = \alpha \sum_{i\subset G, i\neq j}^{n} \frac{R_i^2}{r_{i-j}^2} \quad (1)$$

$$\sigma^*_{G\to j} = \beta \sum_{i\subset G, i\neq j}^{n} \frac{(\chi_i^0 - \chi_j^0)R_i^2}{r_{i-j}^2} \quad (2)$$

where R_S is the steric influence of a group of n atoms constituting a group G onto a single atom j (reaction center), σ^* is the inductive effect of G onto reaction center j. R corresponds to the covalent atomic radii of an ith atom of a group G, r is the distance between atoms i and j, and χ^0 is atomic electronegativity. Parameters α and β in eqs 1 and 2 normalize them to the format of Taft's original electronic and steric substituent constants.[24,25]

Considering the initial success of "inductive" descriptors in QSAR,[17–19,26–28] we adopted the LFER methodology to describe protein–ligand interactions and updated the scope of eqs 1 and 2 to the effects translated by N-atomic ligand L onto a given receptor atom j

$$R_{s_{L\to p}} = \sum_{i\subset L}^{N} \frac{R_i^2}{r_{i-p}^2} \quad (3)$$

$$\sigma^*_{L\to p} = \sum_{i\subset L}^{N} \frac{(\chi_i^0 - \chi_p^0)R_i^2}{r_{i-p}^2} \quad (4)$$

where parameters R_{sL-p} and $\sigma L\rightarrow p^*$ describe the overall inductive and steric interactions occurring between the entire bound ligand and a receptor's atom considered as a reaction center.

Since the LFER principle is not *a priori* limited to ligand-based considerations, it is reasonable to consider how inductive and steric effects influence it in the context of intermolecular interactions. The exact nature of inductive effects (2) is still debated, but direct electrostatic interactions (applicable for both intra- and intermolecular levels of approximations) are often viewed as the main mechanism of its transduction.[25] Similarly, the model of frontal steric effects[25] underlying eq 1 is not dependent on atomic connectivity or grouping and can be easily applied for quantification of protein–ligand mutual screening.[25]

Thus, we have applied eqs 3 and 4 to the 84 compound training set placed inside the 1LHN active site and calculated normalized R_S and σ^* parameters for their optimal target binding orientations. We have considered all protein atoms (except nonpolar hydrogens) in 7.0 Å ligand proximity (see Materials and Methods for details).

In our opinion, these molecular parameters $R_{sL\to p}$ and $\sigma L\rightarrow p^*$ can be regarded as receptor-dependent 3D-QSAR descriptors because they are derived from three-dimensional structures of compounds and rely on their positioning within the target protein taking into account pairing of protein and ligand atoms. We expected that such 4D "inductive" descriptors would possess good predictive ability and illustrate the advantage of using correct steroidal alignment in modeling SHBG binding.

GFA-Based QSAR Models. Considering all structures from the training set (84 entries), we computed the full spectrum of R_S and σ^* values (one of each for every considered atom of 7.0 Å ligand surrounding). To relate such a large number of descriptors to dependent variables pK_d we employed the Genetic Algorithm approximation, which has been applied for QSAR analyses relying on a heuristic search.[29]

In our current study, we adopted the Genetic Function Approximation (GFA) developed by Rogers and Hopfinger,[30] which is based on the G/SPLINES Genetic Algorithm implementation.[31,32] Given a large number of QSAR descriptors to sample, this approach creates a "population" of QSAR models and applies the "fitness function" to iteratively evolve them to an optimal solution (i.e., to find the most appropriate set of descriptors). The GFA approach uses Friedman's "lack-of-fit" (LOF) fitness criteria

$$\mathrm{LOF} = \frac{\mathrm{LSE}}{\left(1 - \frac{c + dp}{n}\right)^2}$$

where LSE is the least squared error; c is the number of descriptors employed by the model; d is the user-defined smoothing factor, p is the total number of available descriptors, and n is the number of the training set molecules.

We have applied the GFA approach implemented within the WOLF package with the following default settings: the initial population of QSAR models has been limited to 5000; the total number of crossovers was set to 200000, and up to 50% of models were allowed to mutate in every generation (i.e., to randomly sample descriptor values). The resulting linear QSAR solutions have been constructed using the PLS approach[23] and have been further validated by the leave-one-out (LOO) procedure.

The parameters of the final optimal QSAR solution based on six "inductive" descriptors are presented in Table 2, while the predicted activity parameters are listed in Table 1. As these results indicate, the Genetic Algorithm provided modestly accurate but, nonetheless, reasonable and simple models predicting 84 SHBG binding constants with correlation coefficients $r^2 = 0.56$ and $q^2 = 0.45$. The developed GFA-QSAR model could efficiently rank the most active ligands and despite modest training statistics allowed a 3.5-fold hit enrichment.

QuaSAR-Evolution Models. In addition to the GFA method, we used the Genetic Algorithm-based approach implemented by the QuaSAR-Evolution module of the MOE program.[34] This tool enables automated QSAR modeling "on the fly" and is available through the "SVL exchange".[35]

We applied the QuaSAR-Evolution tool with its default settings: (a) the initial population of 100 models; (b) four additional descriptors added to each generation of QSAR

models; (c) multiple linear regression (MLR) mode; (d) the total number of crossovers set to 50000; (e) allowed 50% mutation; and (f) the autotermination factor of 1000 (meaning that the calculation was stopped when the "fitness function" value does not change during 1000 crossovers). The resulting QSAR solution demonstrated better training accuracy compared to the previous GFA models with $r^2 =$ and q^2 estimated as 0.66 and 0.58, respectively (also allowing 3.50-fold enrichment of the training set). It should also be mentioned that the developed linear QSAR model could provide some insight into factors determining SHBG dissociation constants:

$$\begin{aligned} \mathrm{p}K_\mathrm{d} = 4.63 - 0.98\mathrm{Rs\_ASP65} - 1.18\mathrm{Sigma\_ASP65} + \\ 0.36\mathrm{Sigma\_GLY58} + 1.22\mathrm{Sigma\_HIS136} - \\ 0.41\mathrm{Sigma\_LEU69} + \\ 0.66\mathrm{Sigma\_PHE44} - 0.54\mathrm{Sigma\_THR60} - \\ 0.65\mathrm{Sigma\_VAL127};\ N = 84 \mathrm{r2} = 0.66 q2 = 0.58 \mathrm{SE} = 0.79 \end{aligned} \quad (5)$$

The above LFER equation illustrates that one of the most significant contributions to pK_d comes from inductive interaction between a ligand and the Asp65 side-chain (critical anchoring residue located at the "gating" mobile loop of the SHBG active site believed to control ligand uptake and release[15]). Steric interactions with Asp65 that are described by the Rs_ASP65 descriptor also play an important role in ligand binding, such that (5) illustrates its minimization helps to increase pK_d.

Similarly, according to (5), the electron-withdrawing effect exhibited toward Gly58 should also increase pK_d, and it is likely that the backbone oxygen of that residue influences the polar stabilization and hydrogen bonding of some SHBG ligands. The possible role of inductive interactions between a bound ligand and Thr60 is also reflected by (5).

Of note, the involvement of residues His136, Leu69, Phe44, and Val129 featured in (5) with respect to protein–ligand interactions is less obvious, and some known ligand binders such as Ser42 are not reflected by (5). These inconsistencies can perhaps be explained by the limited variability of inductive and steric effects of some residues or by the approximate nature of Genetic Algorithm solutions.

Overall, the above results substantiate the adequate accuracy of the developed QSAR models and their ability to account for specific protein–ligand interactions.

Consensus Scoring by the Developed QSAR Models. To further expand the utility of the developed QSAR models, we have implemented the consensus scoring approach. Thus, eight different predicted parameters of potential activity (two Glide-, two CoMFA-, two CoMSiA-, one GFA-QSAR, and one QuaSAR-Evolution values) have been produced for every entry in the training set. Based on these sorted values, each molecule would then receive a binary 1.0 vote for every "top15% appearance" (thus, the maximum possible vote was set to 8.0). The final cumulative vote was then used to rank the training set entries.

The resulting 5.0-fold enrichment of the top 15% binders in the "hit list" clearly demonstrated that the consensus scoring strategy produces the most balanced predictions and that a synergetic approach can capitalize on the strengths of individual approaches (such as the positive predictive power of ligand-based QSAR techniques and the negative predictive power of docking) and compensate for their weaknesses. Furthermore, we have evaluated several other combined strategies (also featured in Table 2) and discovered that all of them resulted in consistent 4.0–5.0-fold enrichments of the training set.

It should be noted that the use of complementary predicting tools and the implementation of scoring/voting protocols has recently become one of the most important topics in the field of computer-aided drug design.[38]

Selection of Potential SHBG Binders. Overall, the results of QSAR modeling of the training set demonstrated good accuracy of the developed solutions, their useful synergy, and ability to enrich for the most active target binders. These observations encouraged us to apply our scoring systems to electronic collections of commercially available chemicals for the identification of novel nonsteroidal SHBG binders. In this study, we used the ZINC 5.0 molecular database[36] that included 3.3 million entries. From these, we derived 2066886 nonredundant molecules satisfying drug-likeness criteria (see the Materials and Methods for details).

As described in the previous section, all created fields-based and 4D-QSAR solutions were based on high-precision docking possess. Therefore, in order to apply pretrained CoMFA, CoMSiA, and 4D-QSAR models, we docked all 2066886 structures into the 1LHN ligand-binding site. This protein structure was used because, as previously noted, it allowed us to produce the correct docking poses and binding characteristics of the training set compounds. Furthermore, to account for possible induced changes in the SHBG active site, we also docked all 2066886 molecules into a 1KDM protein structure (corresponding to SHBG cocrystallized with 5α-dihydrotestosterone, i.e., compound with the highest binding affinity). All molecular structures that produced GlideScore values <-7.0 were selected, and thus, two redundant hit-lists have been generated, with one of them corresponding to 143,421 best 1KDM-docked ligands and 213,191 to top 1LHN-predicted binders. Next, we implemented a scoring system that assigned a 1.0 vote to the top 5% of both 1KDM (7,171) and 1LHN (10,659) hits, while all other docked ligands were given a vote value of 0. Based on the resulting cumulative vote, we selected 3759 structures for future assessment.

All of the selected docking poses were examined visually; several broken and inconsistently docked structures were removed, and all steroidal derivatives and compounds containing a carboxylic group were eliminated, in order to reduce the total number of selected structures to 1,419. All of these ligands were then redocked into the 1KDM and 1LHN active sites using the XP_Glide.[20]

The resulting "extra precision" docking poses were scored by CoMFA_21, CoMFA_84, CoMSiA_21, CoMSiA_84and QSAR_GFA, QuaSAR-Evolution solutions, where "_21" and "_84" symbols mark models created on the original and updated sets of SHBG ligands respectively.

After sorting all eight sets of predicted activities, we computed the cumulative votes for 1,419 molecular structures, where an entry would receive a vote for every "top 15%" appearance. Based on these cumulative parameters (with the maximal possible value of 8.0), we selected a list of 111 hits, all of which had been voted on more than 3 times.

After the final visual inspection, we formed a list of 87 compounds out of which 41 chemical substances could be readily purchased in sufficient purity and quantity for biochemical verification as SHBG binders, as described below.

Experimental Testing. All 41 compounds selected from the 3.3 million ZINC entries[36] by applying drug-likeness criteria followed by the combination of docking, CoMFA, CoMSiA, and 4D QSAR filters were further screened for their ability to interact with the SHBG steroid-binding site *in vitro*. The screening assay involved a modification of an established competitive steroid ligand-binding assay that employs tritium

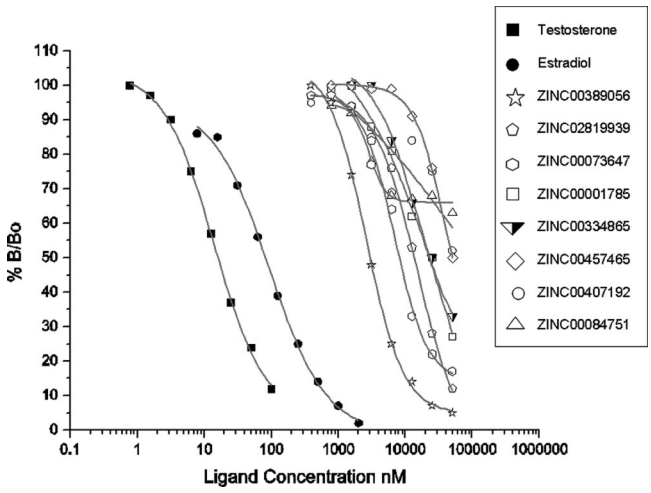


Figure 4. Displacement curves for test compounds used in the in vitro competition assay to determine the relative binding affinities of human SHBG ligands. The percentage of tritium-labeled 5α-dihydrotestosterone bound to SHBG in the presence of increasing concentrations of competitor ligands.

labeled 5α-dihydrotestosterone as the radio-labeled ligand (see the Materials and Methods for details).

The initial experimental screen of test compounds was conducted at a single high concentration (approximately 100 μM), and 25 out of 41 compounds demonstrated some SHBG-binding competition with the tritium-labeled 5α-dihydrotestosterone. The seven substances that displaced more than 50% of bound 5α-dihydrotestosterone from the protein at the 100 μM concentration were further analyzed in a concentration-dependent manner (more data can be found in the Supporting Information).

The competitive displacement curves generated using these nonsteroidal SHBG ligands (entries **85**–**91** in Table 1) are presented in Figure 4, and the corresponding SHBG dissociation constants calculated from the plot are included in Table 1.

As can be seen, the five most active SHBG ligands exhibited nanomolar dissociation constants: 106.9 nM for compound **85**, 408.8 nM for compound **86**, 591.3 nM for compound **88**, 833.6 nM for compound **89**, and 964 nM for compound **90**. The parallelism of the corresponding competitive displacement curves in Figure 4 indicates that these compounds are completely soluble at high concentrations and behave in essentially the same way as a steroid ligand with respect to their kinetics of binding. Taken together with their high affinity toward the target, these novel SHBG binders represent potential therapeutic prototypes.

It is also worth mentioning, that this hit rate appears very good (25 out of 41 tested chemicals showed some activity, with 7 of them being nanomolar-to low micromolar binders), especially considering that we could only test available (as opposed to custom-made) compounds, and taking into account the financial constraints of academia-based drug discovery research.

Nonsteroidal SHBG Binders. Analysis of the docking poses of the eight most active ligands (for seven of them p$K_d$values could be measured) provided additional and important insight into the mechanism of SHBG binding. As Figure 5 illustrates, all eight ligands likely form a hydrogen bond with Asp65 side chain (supported by the secondary H-interaction with Asn82), with two of them, compounds **89** and **90**, also showing strong H-binding toward Ser42 (supported by additional interaction with Val105 backbone oxygen). These observations illustrate the importance of Asp65 and Ser42 anchoring residues previously outlined in numerous SHBG-related publications.[14,15] It should be noted, however, that the presence of two anchoring H-bonds did not make the compounds better binders. In fact, three other substances **85**, **86**, and **88** all formed only one hydrogen bond but demonstrate higher affinity toward the target (likely caused by more favorable hydrophobic interactions).

The importance of hydrophobic forces is well recognized for SHBG binding,[14] and it is therefore no surprise that all eight ligands have sizable aliphatic and aromatic cores that could participate in close-range interactions within hydrophobic pockets. One such pocket is located in close proximity to the Ser42 residue and is formed by the Leu171, Met138, and Val105 side chains of human SHBG. The later residue also forms a hydrophobic patch together with the Phe67 side chain providing additional stabilization for bound ligands. Importantly, Phe67 is also likely involved in π-stacking with aromatic rings of **88**, **89**, and **90** and perhaps with the C═O group of **87** (see Figure 5 for more details). It is also possible that the SHBG affinity for compounds **85**, **86**, and **88**, which involves strong hydrophobic interactions with key residues within its steroid-binding pocket, could be further increased by introducing additional H-bond enabling groups into their Ser42-oriented ends. Such structural modifications could represent an attractive strategy for lead optimization. It is also possible that an extra hydrogen-bond acceptor to the Asp65-oriented end of a ligand that would engage the Asn82 side chain could enhance binding, as would appear to occur in the cases of compounds **88** and **90**.

Another "atypical" interaction within the active site has been found for ligand **91**, which forms an additional H-bond with the Gly58 backbone oxygen. Such coordination has never been previously observed for SHBG ligands, but the possible relevance of this residue was hinted at by the LFER equation (5).

Conclusions

Using available information on known ligands of human sex hormone binding globulin, we have developed several structure–activity models based on conventional (CoMFA and CoMSiA) and newly developed QSAR approaches. While building such QSAR solutions we used molecular alignments that contradict conventional way of superimposing steroidal SHBG ligands, but are in line with direct crystallographic evidence of the preferred steroid-binding poses. We have demonstrated that molecular-field based techniques such as CoMFA and CoMSiA are not very sensitive to ligand alignment, as they result in almost indistinguishable QSAR models derived from the traditional and "crystallographic" alignments of steroidal scaffolds.

To compensate for that drawback, we developed novel ligand-induced active site descriptors (called "inductive" 4D QSAR parameters) which provided additional insights into factors influencing ligand-protein interactions and which have been successfully used in combination with other "*in silico*" drug discovery tools. Thus, the developed range of "*in silico*" solutions have been applied in a consensus manner to more than 2 million structures from the ZINC database[36] and allowed identification of 41 potential SHBG binders. When evaluated experimentally, 25 out of 41 selected candidates demonstrated detectible binding to human SHBG in plasma. Notably, five such novel nonsteroidal SHBG inhibitors demonstrated nanomolar dissociation constants, with the best binder exhibiting K_d = 109 nM and representing the most active nonsteroidal SHBG ligand known to date.

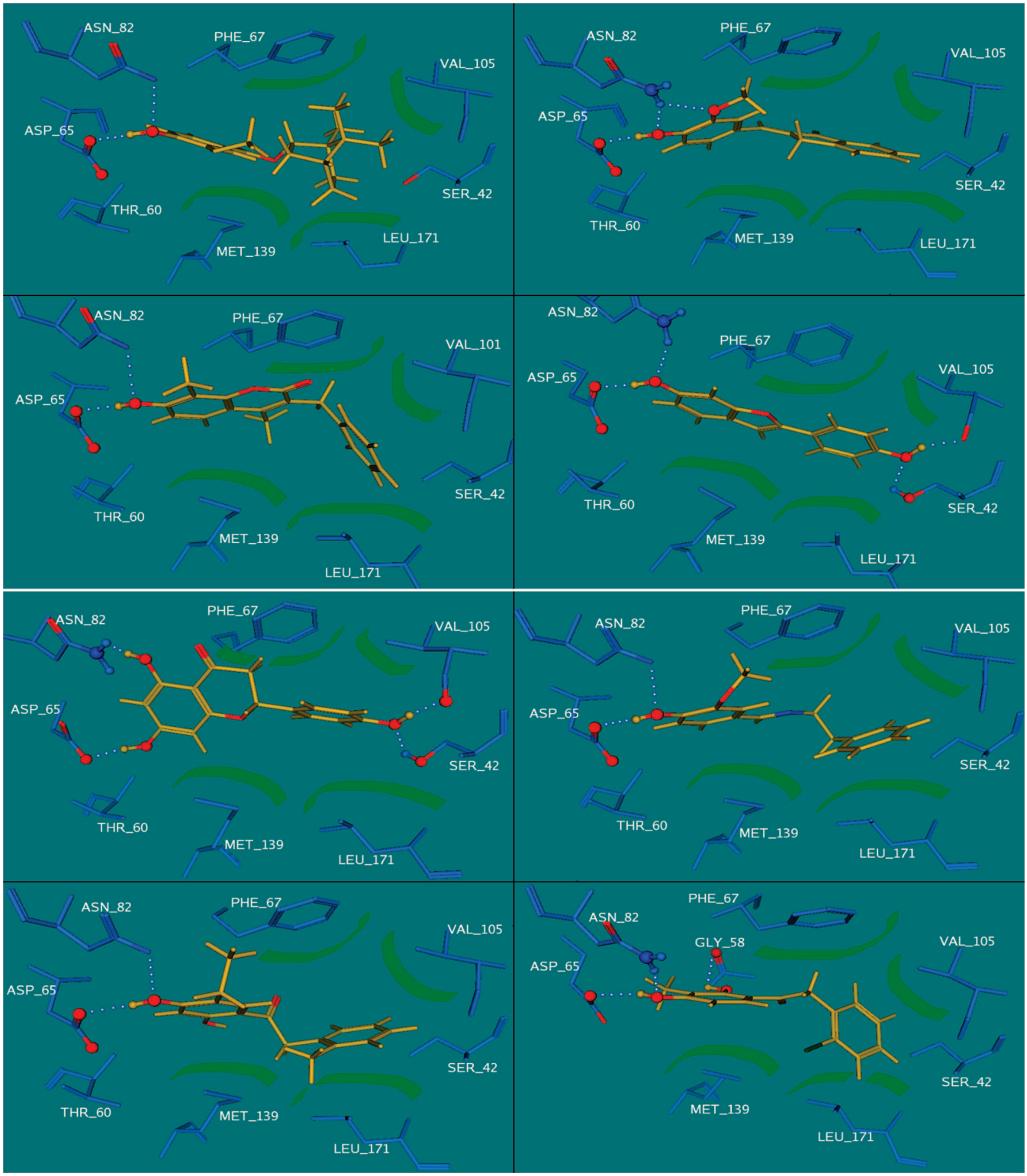


Figure 5. Docked poses of the most active nonsteroidal ligands within the SHBG binding pocket (blue). Only those residues that are most relevant to ligand binding are shown. Hydrogen bonds are represented as white dots; hydrophobic interactions featured by thick green lines. The following eight compounds are shown (ordered from left to right and top to bottom in the figure): **86**, ZINC00084751 (not studied due to solubility issues), **87**, **89**, **90**, **88**, **85**, and **91**.

Since SHBG represents a prospective drug target the identified nonsteroidal lead compounds can be characterized as potential therapeutic agents laying a foundation for future lead optimization studies.

Materials and Methods

Database Preparation. The initially considered set of 3.3 million compounds from the ZINC 5.0 database[36] was reduced to 2,066,886 entries by applying the drug-likeness criteria: molecular weight between 300 and 800 Da; the presence of 1−10 hydrogen bond acceptors; 1−5 hydrogen bond donors; less than 10 rotatable bonds, and overall hydrophobicity below log $P = 5.00$.

The resulting set of 2,066,886 drug-like structures has been washed; i.e. all inorganic components have been removed, and all ionizable groups have been coordinated with pH = 7.0 conditions.

All molecular structures have been optimized using PM3 semiempirical method implemented within the MOE package.[34]

Docking. The Maestro suite[39] was used to prepare 1LHN and 1KDM protein structures for docking. All water and ion–molecules were removed from the corresponding PDB files, and hydrogen atoms were added and adjusted where necessary. Steroid-binding sites were defined as 10 Å surrounding the cocrystallized ligands in the 1KDM and 1LHN.

Subsequent docking was conducted using the Glide 4.0 parallel suite[20] with default settings. All computations were carried out on 10 dual-core LINUX/Centos4.3 IBM stations equipped with Intel Pentium D CPU 3.00 GHz processors and 2 GB RAM of memory. The overall docking time constituted 22 days.

QSAR Descriptors Calculation and Model Building. The extra precision docking poses of 84 training set compounds, 64 chemicals investigated in our previous SHBG studies, and 1,419 selected molecules were used to compute 4D "inductive" QSAR descriptors according to eqs 3 and 4. For every docked molecule placed in the 1LHN active site (defined as 7A surrounding of its native ligand), we computed the direct 3D distances from all atoms of a ligand to the active site's polar hydrogens and heavy atoms. The computed distances have were used in eqs 3 and 4 to compute the cumulative parameters of inductive σ^* and steric R_S influence of a ligand to every considered atom within the active site. All calculations were implemented with customized SVL scripts of the MOE program.[34]

The defined1LHN active site contained 289 heavy atoms and polar hydrogens, and therefore, for every ligand we computed $2 \times 289 = 578$ values of descriptors. These parameters were then used to create predictive QSAR solutions based on the Genetic Algorithm approximation.

The QuaSAR-Evolution models were built using "autoqsar.svl" script obtained from the SVL exchanged site.[35] The default setting was used.

The GFA models were created using WOLF 6.2 software (with default settings) kindly provided by Professor Hopfinger.[40] Both of these programs automatically handle the descriptors' cross-correlation problems and possess built-in capabilities for LOO cross-validation.

The actual values of normalized 4D QSAR descriptors can be obtained upon request.

Molecular Alignment. For generating the "traditional" set of superimposed steroidal structures we used the SYBYL[41] Fit_Atoms functionality, which is based on the BMFIT method.[42] The 1LHN docking pose of its ligand was used as the reference, while other molecules were translated and rotated in a way to fit the weighted centroid of atoms of the "A" ring of steroidal scaffolds. In this way, all steroids have been superimposed in 'head-to-head' orientations, and all nonsteroidal structures were also taken in their docked configurations.

In the "corrected" data sets all steroidal and nonsteroidal ligands were used in their respective 1LHN docking poses.

CoMFA Modeling. The SYBYL package[41] was used to construct all CoMFA models using the partial least-squares fitting, with the cross-validation carried out by the built-in LOO procedure.

Both the traditional and expanded data sets of SHBG binders (containing 21 and 84 entries, respectively) were used independently to compute steric and electrostatic CoMFA fields. The steric ones were calculated on 2 Å grids, by evaluating "6–12" Van der Walls interaction with default CoMFA probes. We used distance-dependent dielectric parameters to compute the Coulombic interactions approximating electrostatic CoMFA fields and set the fields truncation parameter to 30.0 kcal/mol.

For the traditional 21 steroids of the benchmark set we also recreated CoMFA models based on traditional similarity-based molecular alignment used in,[12] with the resulting statistics reproducing original values reported in ref 12.

CoMSiA Modeling. For the studied data sets, we computed 5 CoMSiA properties that included steric-, electrostatic-, hydrophobic-, hydrogen bond donor-, and hydrogen bond acceptor-fields (computed with default settings). The fields were derived according to similarity indexes (computed with 0.3 attenuation factor) of molecules brought into a common alignment. In the CoMSiAstudy, we utilized the same alignment schemes as in CoMFAmodeling.

All calculations were carried out with default settings; each CoMSiAproperty of a given atom was scaled to 74.1% for its 1 Å proximity, to 30.1% for >2 Å surrounding, and to 6.7% for the area within 3 Å.

The final CoMSiA models were constructed using the partial least-squares (PLS) algorithm[23] and cross-validated by the LOO procedure implemented by the SYBIL package.[41]

Enrichment Calculations. All predicted SHBG affinity parameters (including Glide docking scores, CoMFA, CoMSiA, and 4D-QSAR outputs) have been processed into the parameters of percent yield (%Y), percent accurate (%A), enrichment factor (E), and Goodness of Hit list (GH) parameters custom for *in silico* screening studies:

$$\%\mathrm{Y} = H_a/H_t$$

$$\%\mathrm{A} = H_a/A$$

$$\mathrm{EF} = (H_a/H_t)/(A/D)$$

$$\mathrm{GH} = \left(\frac{H_a(3A + H_t)}{4H_tA}\right)\left(1 - \frac{H_t - H_a}{D - A}\right)$$

where H_t is the total number of compounds in the hit list (in our case, top 15% portion of the sorted predictions), H_a is the number of known actives in the hit list (true positives), A is the active compounds in the database, and D is the number of compounds in the database. We have only reported the EF values; other parameters can be obtained from authors upon request.

The corresponding calculations have been carried out using in-house SVL scripts.

SHBG Ligand-Binding Assay. An established competitive ligand-binding assay was used to determine the relative binding affinities of the studied compounds to human SHBG, compared to testosterone and estradiol standards.[43] In brief, the assay involved mixing 100 μL aliquots of diluted (1:200) human pregnancy serum containing approximately 1 nM SHBG, which was pretreated with dextran-coated charcoal (DCC) to remove endogenous steroid ligand, with 100 μL of tritium-labeled 5α-dihydrotestosterone at 10 nM as labeled ligand. For the screening assay, triplicate aliquots (100 μL) of a fixed amount (100 μM) of test compound were added to this mixture and incubated overnight at room temperature. After further 10 min incubation at 0 °C, 500 μL of a DCC slurry was added at 0 °C and incubated for 10 min prior to centrifugation to separate SHBG-bound from free 5α-dihydrotestosterone. Compounds that displaced more than 35% of the tritium-labeled 5α-dihydrotestosterone from the SHBG in this assay were then diluted serially, and triplicate aliquots (100 μL) of known concentrations of test compounds were used to generate complete competition curves by incubation with the SHBG/5α-dihydrotestosterone mixture, and separation of SHBG-bound from free steroid, as in the screening assay. The amounts of 5α-dihydrotestosterone bound to SHBG at each concentration of competitor ligand were determined by scintillation spectrophotometry and plotted in relation to the amount of 5α-dihydrotestosterone bound to SHBG at zero concentration of competitor. From the resulting competition curves, IC_{50} concentrations could be calculated if displacement of more than 50% of tritium labeled 5α-dihydrotestosterone from SHBG was achieved.

The dissociation constants (K_d) have been calculated from the relative binding affinity parameters using the following equation: $1/\{K_a(\text{dihydrotestosterone})/[(1 + R)/\text{RBA} - \text{R}]\}$, where $K_a(\text{dihydrotestosterone}) = 0.98 \times 10^9\ \text{M}^{-1}$ is the association constant of the 5α-dihydrotestosterone and R (0.05) is the ratio of bound-to-free 5α-dihydrotestosterone at 50% displacement in the assay.

Acknowledgment. G.L.H. is a Canada Research Chair in Reproductive Health and is supported by operating grants from the Canadian Institutes of Health Research, F.B. is a Canadian Institutes of Health Research Postdoctoral Scholar. N.T. ac-

knowledges the support of the CIHR/MSFHR Strategic Training Program in Bioinformatics (http://bioinformatics.bcgsc.ca). O.S.F.'s work has been supported by the Genome Canada-funded PREPARE project (www.prepare.med.ubc.ca). A.C. acknowledges financial support from the Prostate Centre at the VGH. We thank Professor Anthony Hopfinger for providing the WOLF 6.0 molecular modeling program.

Supporting Information Available: Table containing information on compounds experimentally tested for binding to human sex hormone binding globulin. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Westphal, U. *Steroid−protein interactions*; Monographs on endocrinology, Vol. 4; Springer: New York, 1971; 567 pp.
(2) Tuppurainen, K.; Viisas, M.; Perakyla, M.; et al. Ligand intramolecular motions in ligand-protein interaction: ALPHA, a novel dynamic descriptor and a QSAR study with extended steroid benchmark dataset. *J. Comput. Aided Drug Design* **2004**, *18*, 175–187.
(3) Asikainen, A. H.; Ruuskanen, J.; Tuppurainen, K. A. Performance of (consensus) kNN QSAR for predicting estrogenic activity in a large diverse set of organic compounds. *SAR QSAR Environ. Res.* **2004**, *15*, 19–32.
(4) Korhonen, S. P.; Tuppurainen, K.; Laatikainen, R.; Perakyla, M. FLUFF-BALL, a template-based grid-independent superposition and QSAR technique: validation using a benchmark steroid data set. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1780–1793.
(5) Liu, S. S.; Yin, C. S.; Wang, L. S. Combined MEDV-GA-MLR method for QSAR of three panels of steroids, dipeptides, and COX-2 inhibitors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 749–756.
(6) Tuppurainen, K.; Viisas, M.; Laatikainen, R.; Perakyla, M. Evaluation of a novel electronic eigenvalue (EEVA) molecular descriptor for QSAR/QSPR studies: validation using a benchmark steroid data set. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 607–613.
(7) Liu, S. S.; Yin, C. S.; Li, Z. L.; Cai, S. X. QSAR study of steroid benchmark and dipeptides based on MEDV-13. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 321–329.
(8) Polanski, J.; Walczak, B. The comparative molecular surface analysis (COMSA): a novel tool for molecular design. *Comput. Chem.* **2000**, *24*, 615–25.
(9) Turner, D. B.; Willett, P.; Ferguson; Heritage, T. W. Evaluation of a novel molecular vibration-based descriptor (EVA) for QSAR studies: 2. Model validation using a benchmark steroid dataset. *J. Comput. Aided Mol. Des.* **1999**, *13*, 271–296.
(10) Robinson, D. D.; Winn, P. J.; Lyne, P. D.; Richards, W. G. Self-organizing molecular field analysis: a tool for structure-activity studies. *J. Med. Chem.* **1999**, *42*, 573–583.
(11) Jain, A. N.; Koile, K.; Chapman, D. Compass: predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. *J. Med. Chem.* **1994**, *37*, 2315–2327.
(12) Cramer, R. D., III; Patterson, D. E.; Bunce, J. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
(13) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict their Biological Activity. *J. Med. Chem.* **1994**, *37*, 4130–4146.
(14) Grishkovskaya, G. V.; Avvakumov, G. V.; Hammond, G. L.; Catalano, M. G.; Muller, Y. A. Steroid ligands bind human sex hormone binding globulin in specific orientations and produce distinct changes in protein conformation. *J. Biol. Chem.* **2002**, *207*, 32086–32093.
(15) Hammond, G. L.; Avvakumov, G. V.; Muller, Y. A. Structure/function analysis of human sex hormone binding globulin: effects of zinc and steroid-binding specificity. *J. Steroid Biochem. Mol. Biol.* **2003**, *85*, 195–200.
(16) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucl. Acid Res.* **2000**, *28*, 235–242.
(17) Cherkasov, A.; Shi, Z.; Fallahi, M.; Hammond, G. L. Successful in Silico Discovery of Novel Non-Steroidal Ligands for Human Sex Hormone Binding Globulin (SHBG). *J. Med. Chem.* **2005**, *48*, 3203–3213.
(18) Cherkasov, A.; Li, Y.; Fallahi, M.; Hammond, G. L. 'Progressive docking': A Hybrid QSAR/Docking Approach for Accelerating in silico High Throughput Screening. *J. Med. Chem.* **2006**, *49*, 7466–7478.
(19) Cherkasov, A.; Shi, Z.; Li, Y.; Jones, S. J. M.; Fallahi, M.; Hammond, G. L. Inductive Charges on Atoms in Proteins: Comparative Docking with the Extended Steroid Benchmark Set and Discovery of a Novel SHBG Ligand. *J. Chem. Inf. Model.* **2005**, *45*, 1842–1853.
(20) Glide; Version 4.0, Schrödinger Inc., San Diego, CA, 2006.
(21) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–519.
(22) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.
(23) Stahle, L.; Wold, S. Multivariate Data Analysis and Experimental Design in Biomedical Research. *Progr. Med. Chem* **1988**, *25*, 292–334.
(24) Cherkasov, A. 'Inductive' Descriptors. 10 Successful Years in QSAR. *Curt. Comput.-Aided Drug Design* **2005**, *1*, 21–42.
(25) Cherkasov, A.; Galkin, V. I.; Cherkasov, R. A. The problem of the quantitative evaluation of the inductive effect: correlation analysis. *Russ. Chem. Rev.* **1996**, *65*, 641–656.
(26) Karakoc, A.; Cherkasov, A.; Sahinalp, S. C. Distance Based Algorithms for Small Biomolecule Classification and Structural Similarity Search. *Bioinformatics* **2006**, *22*, e243–251.
(27) Karakoc, A.; Sahinalp, S. C.; Cherkasov, A. Comparative QSAR- and Fragments Distribution Analysis of Drugs, Drug-likes, Metabolic Substances and Antimicrobial Compounds. *J. Chem. Inf. Model.* **2006**, *46*, 2167–2182.
(28) Cherkasov, A. Can 'Bacterial-Metabolite-Likeness' Model Improve Odds of in silico Antibiotic Discovery? *J. Chem. Inf. Model.* **2006**, *46*, 1214–1222.
(29) Holland J. H. *Adaptation in Natural and Artificial Systems*; Ann Arbor, MI, 1975.
(30) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.
(31) Rogers, D. G/SPLINES: A Hybrid of Friedman's Multivariate Adaptive Regression Splines (MARS) Algorithm with Holland's Genetic Algorithm. Proc. Fourth Int. Conf. Genet. Algorithms, San Diego, July 1991.
(32) Rogers, D. Data Analysis using G/SPLINES. *Advances in Neural Processing Systems 4*; Kaufmann: San Mateo, CA, 1992.
(33) Friedman, J. *Multivariate Adaptive Regression Splines*; Technical Report No. 102, Laboratory for Computational Statistics, Department of Statistics; Stanford University: Stanford, CA, Nov 1988 (revised Aug 1990).
(34) *MOE: Molecular Operational Environment; Version 2004.03*; Chemical Computation Group, Inc.: Montreal, Canada, 2004.
(35) SVL exchange: http://svl.chemcomp.com/viewcat.php
(36) Irwin, J. J.; Shoichet, B. K. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(37) Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov. Today* **2006**, *11*, 580–594.
(38) Feher, M. Consensus scoring for protein-ligand interactions. *Drug Discov. Today* **2006**, *11*, 421–428.
(39) Maestero; Schrödinger Inc., San Diego, CA, 2004.
(40) WOLF package, version 6.2, Chem21 Group, Inc., 2007.
(41) *SYBYL, version 7.2*, Tripos, Inc.: St. Louis, MO, 2006.
(42) Nyburg, S. C. Some Uses of a Best Molecular Fit Routine. *Acta Crystallogr.* **1974**, *B30*, 251–253.
(43) Hammond, G. L.; Lahteenmaki, P. L. A versatile method for the determination of serum cortisol binding globulin and sex hormone binding globulin binding capacities. *Clin. Chim. Acta* **1983**, *132*, 101–110.

JM7011485