

## Progressive Docking: A Hybrid QSAR/Docking Approach for Accelerating In Silico High Throughput Screening

Artem Cherkasov,<sup>\*,†</sup> Fuqiang Ban,<sup>†</sup> Yvonne Li,<sup>‡</sup> Magid Fallahi,<sup>§</sup> and Geoffrey L. Hammond<sup>§</sup>

Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, 675 West 10th Avenue, Vancouver, British Columbia, V5Z 1L3, Department of Obstetrics and Gynecology, Child and Family Research Institute, University of British Columbia, and Division of Infectious Diseases, Faculty of Medicine, University of British Columbia, Vancouver, British Columbia V5Z 3J5

Received August 8, 2006

A combination of protein–ligand docking and ligand-based QSAR approaches has been elaborated, aiming to speed-up the process of virtual screening. In particular, this approach utilizes docking scores generated for already processed compounds to build predictive QSAR models that, in turn, assess hypothetical target binding affinities for yet undocked entries. The “progressive docking” has been tested on drug-like substances from the NCI database that have been docked into several unrelated targets, including human sex hormone binding globulin (SHBG), carbonic anhydrase, corticosteroid-binding globulin, SARS 3C-like protease, and HIV1 reverse transcriptase. We demonstrate that progressive docking can reduce the amount of computations 1.2- to 2.6-fold (when compared to traditional docking), while maintaining 80–99% hit recovery rates. This progressive-docking procedure, therefore, substantially accelerates high throughput screening, especially when using high accuracy (slower) docking approaches and large-sized datasets, and has allowed us to identify several novel potent nonsteroidal SHBG ligands.

### Introduction

This work continues our efforts on the development and application of “inductive” QSAR descriptors for in silico modeling studies, in particular, for the discovery and optimization of drug leads.

In a series of previous studies, we reported the development of inductive 3D-sensitive QSAR descriptors that are related to atomic electronegativity, covalent radii, and intramolecular distances and can be computed by equations for steric and inductive constants,<sup>1–4</sup> inductive electronegativity,<sup>5,6</sup> inductive partial charge,<sup>7,8</sup> and inductive analogues of chemical hardness and softness<sup>6,8</sup> (see eqs 1–6 in the Materials and Methods section). To date, 50 global inductive descriptors (calculated for a whole molecule) have been developed and they have already demonstrated high effectiveness in predicting physical-chemical properties, reactivity, and the biological activity of compounds. The detailed description of inductive QSAR descriptors and their applications can be found in a recent review.<sup>9</sup>

In our recent studies, we utilized global inductive descriptors in combination with conventional in silico drug design tools for the discovery of novel nonsteroidal ligands for human sex hormone-binding globulin (SHBG).<sup>8,10</sup>

**SHBG as a Drug Target.** Plasma SHBG is the liver-expressed protein that binds biologically active androgens and estrogens and plays a pivotal role in regulating the metabolic clearance of these sex steroids and their access to target tissues. Abnormal levels of SHBG, resulting in alterations in unbounded sex steroids, have been implicated in numerous human diseases including endometrial cancer,<sup>11</sup> ovarian dysfunction,<sup>12</sup> infertility,<sup>13</sup> osteoporosis,<sup>14,15</sup> diabetes,<sup>16</sup> and cardiovascular diseases<sup>17</sup> among others. Thus, the competitive formation of SHBG-

endogenous steroid complexes through the use of novel ligands represents a possible way of liberating endogenous steroids to enhance their biological activities, and this could have potential therapeutic ramifications in the context of diseases associated with steroid insufficiency. Therefore, the discovery of potent nonsteroidal SHBG inhibitors represents an attractive drug design task that can lead to useful alternatives for potentially harmful hormone-replacement therapies.

Aside from being an attractive drug target, human SHBG represents an important model system for conventional in silico chemical studies. Thus, association constants of SHBG with the series of steroidal derivatives form a well-established “steroid benchmark set” that has been investigated in a variety of molecular modeling studies (for the latest examples, see refs 18–27). The structure of the SHBG protein can also serve as a useful and challenging test system. Thus, nine crystal structures of the N-terminal domain of this protein have been solved to date and deposited in the Protein Databank.<sup>28</sup> These structures corresponding to native SHBG complexes with different steroidal ligands provide detailed insight into the topology of the steroid-binding sites and the molecular basis of interactions between steroid ligands and SHBG. In addition, experimental binding affinities between native SHBG or SHBG variants and a range of chemicals are also available.<sup>8,10,29–32</sup>

**Use of Inductive Parameters in Discovery of SHBG Ligands.** In our published studies,<sup>8,10</sup> we tested several in-house molecular modeling solutions utilizing global inductive descriptors in combination with conventional drug design tools and discovered a number of nonsteroidal antagonists that bind to human SHBG with affinity constants of up to  $1.2 \times 10^6 \text{ M}^{-1}$  and which can efficiently displace bound sex hormones.

**Pharmacophore-Based Lead Discovery.** Thus, in our recent work<sup>10</sup> we developed several pharmacophore models for SHBG binders and used them to screen an electronic collection of ~24 000 natural compounds. As a result, we identified 105 natural derivatives that met the pharmacophore requirements. Thus, we faced a well-known and long-standing challenge of ranking the pharmacophore-derived “hits”. To prioritize the

\* To whom correspondence should be addressed. Tel.: 604.875.4588. Fax: 604.875.4013. E-mail: artc@interchange.ubc.ca.

<sup>†</sup> Division of Infectious Diseases, University of British Columbia.

<sup>‡</sup> British Columbia Cancer Agency.

<sup>§</sup> Department of Obstetrics and Gynecology, University of British Columbia.

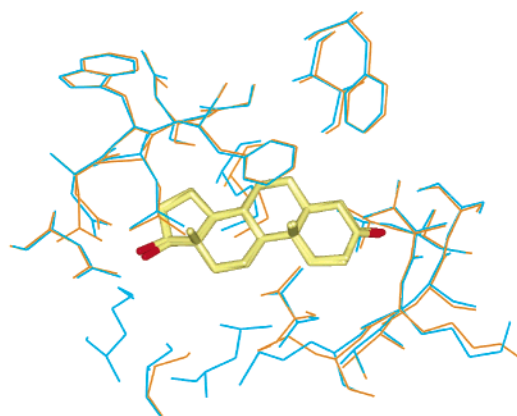
identified 105 nonsteroidal chemicals for further experimental testing, we used inductive descriptors in combination with the method of artificial neural networks (ANN).<sup>33</sup> Namely, we assembled a training set of 78 compounds known to interact with SHBG and compiled a “negative control” set of 165 chemicals with unknown affinities to SHBG. Furthermore, we have trained the ANN model, enabling us to relate 28 inductive descriptors, calculated for those 243 compounds, to their Boolean (1|0) protein binding criteria. The resulting ANN-based binary QSAR model has then been used to prioritize pharmacophore-identified natural substances for their potential ability to bind SHBG. As a result, 22 top-ranked nonsteroidal substances have been tested empirically, and eight of them (corresponding to four different molecular scaffolds) demonstrated sufficient steroid replacement activity.

**Docking with Inductive Protein Charges.** In another recent work,<sup>10</sup> we reported a novel iterative approach that allows rapid and conformation-sensitive computation of inductive atomic charges in proteins. The use of the inductive charge values in a comparative docking study involving human SHBG, and an extended steroid benchmark set, demonstrated their superior performance compared to that of several conventional protein charging systems (including CHARMM, AMBER, MMFF, OPLS, and PEOE among others), and this allowed additional potential drug hits to be discovered for the SHBG target. Thus, inductive reactivity indices can assist in various aspects of *in silico* drug research, because they effectively cover a broad range of bound atoms and molecules whose properties vary in relation to their size, polarizability, electronegativity, compactness, mutual inductive and steric influence, distribution of electronic density, and so on.<sup>9</sup> We now report the development of a novel QSAR/docking protocol that utilizes QSAR solutions based on inductive and conventional global QSAR descriptors to speed up the procedure of virtual high throughput screening.

## Results and Discussion

As indicated above, inductive descriptors cover a broad range of molecular properties and can be used effectively to create various binary inductive QSAR classifiers, such as the previously reported models of antibiotic-likeness,<sup>9,34</sup> drug-likeness,<sup>9,35</sup> and bacterial-metabolite-likeness,<sup>35</sup> which we have created using inductive descriptors. Combinations of global inductive parameters with other QSAR descriptors have also been utilized successfully to distinguish metabolic substances isolated from human, bacterial, plant, and fungal cells and have helped identify the extent of overlap between drugs, inactive chemicals, and metabolites in the descriptor space.<sup>36,37</sup> In the current study, we have employed QSAR solutions based on a combination of global inductive and other QSAR parameters to create quantitative models that enable assessments of *hypothetical* docking scores for defined protein–ligand systems.

It is a commonly held view that one of the main purposes of virtual high-throughput screening is to filter out nonbinders, while positive docking predictions (hits) typically undergo additional *in silico* evaluation. Thus, QSAR models capable of producing hypothetical docking scores could help identify the most probable nonbinders in docking databases and, therefore, reduce the amount of computations required. Such an approach could complement various predocking filters, such as the rejection of docking candidates based on their size, mass, volume, number of atoms,<sup>38</sup> pharmacophore constraints,<sup>39</sup> and/or drug-likeness criteria<sup>40</sup> among others. Effective reduction of docking computations becomes increasingly important as the content of conventional docking databases expands and can



**Figure 1.** The structures of the active sites of 1D2S and 1KDM SHBG structures, superimposed. 1D2S and 1KDM residues within 4.5 Å of the ligand are shown as orange and blue wires, respectively. Both proteins are in complex with DHT, and the ligands are shown in stick mode. Two of the active site residues in 1KDM were not resolvable in 1D2S.

routinely encompass millions of chemical structures. Naturally, ligand-based QSAR has been suggested as a plausible alternative<sup>41</sup> or supplementary<sup>42,43</sup> approach to structure-based lead discovery

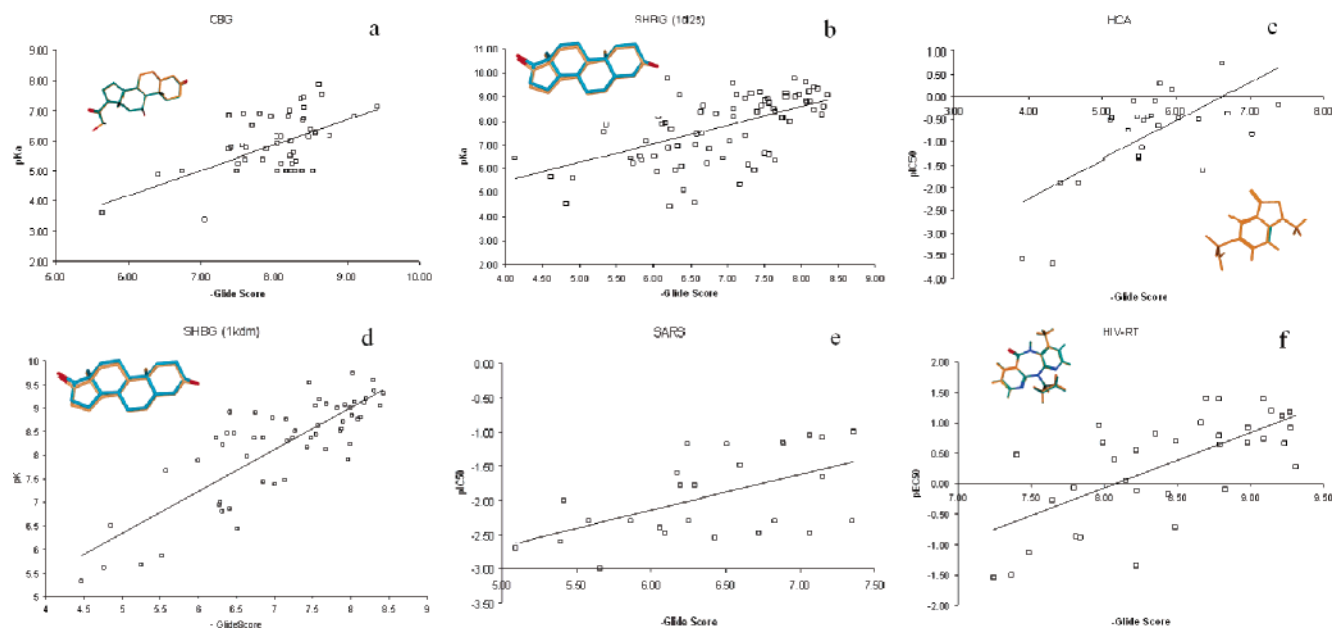
In this study, we conducted *in silico* screening of the National Cancer Institute (NCI) drug-like dataset to identify potential binders for human SHBG using two crystal structures of the protein corresponding to 1KDM and 1D2S entries in the Protein Databank. These two PDB entities were selected for the study because they capture important variations in the SHBG active site that may occur upon ligand binding (see ref 8 for relevant discussion). Figure 1 features the simplified superimposed structures of the active sites of 1D2S and 1KDM, defined as the 4.5 Å environment of the bound ligand—dihydrotestosterone. The two proteins are respectively marked in orange (1D2S) and blue (1KDM) wires, illustrating the flexibility of a loop segment Pro130–Arg135 that can “gate” the active site entry and affect coordination of the functional groups at C3 and C17 of the bound androgen or estrogen, respectively.<sup>29</sup>

For the purpose of protein–ligand docking, we utilized the Glide program,<sup>44</sup> as in our previous studies,<sup>8,10</sup> which is generally viewed as one of the most accurate docking packages.<sup>45,46</sup> The Glide samples conformational space of a ligand during docking using an incremental construction method that makes it a very useful approach. However, the high accuracy of Glide significantly reduces its speed when compared to many other docking protocols, and on average it requires up to 4 min to dock one molecule.<sup>47</sup>

To validate the applicability of the docking protocol, we used Glide software to dock numerous compounds with known SHBG binding affinities into the 1KDM and 1D2S active sites, and this involved using default program parameters. The resulting GlideScores for the majority of compounds studied (the structures can be found in the Supporting Information) reproduced the corresponding experimental association constants with good accuracy (Figure 2b,d).

In addition, the ligand position was also reproduced accurately by docking (Figure 2b,d), which confirmed the overall adequacy of the adopted Glide protocol.

The hypothesis we investigated further is that QSAR models trained on a limited set of docking scores can sufficiently approximate potential target-binding affinities for all compounds in the docking database. If that is true, the obvious nonbinders can be removed from the docking database, and the amount of



**Figure 2.** Dependences between the negative GlideScore values produced by the known binders of the studied protein targets vs the corresponding experimental affinity/activity values. (a) CBG, pK<sub>a</sub> values vs the docking scores; (b) SHBG 1D2S PDB structure, pK<sub>a</sub> values vs the docking scores; (c) HCA, pIC<sub>50</sub> values vs the docking scores; (d) SHBG 1KDM PDB structure, pK<sub>a</sub> values vs the docking scores; (e) SARS protein, pIC<sub>50</sub> values versus the docking scores; (f) HIV reverse transcriptase protein, pEC<sub>50</sub> values vs the docking scores.

required docking computations should, therefore, be significantly decreased. The underlying assumption upon which this is based is the following: when variations in binding orientations of potential ligands is not significant (as it may be expected in the case of a small steroid-binding pocket) and the nature of the target atomic interior remains relatively constant, the differences in binding affinities of compounds can be roughly related to their own structures. The assumption of invariability of the docking orientations for a given target is common for large-scale docking studies<sup>42,43,48</sup> and, if valid, the target affinity values (experimental,  $K_a$  or IC<sub>50</sub>, or theoretical, the docking scores) can be approximated by ligand-based QSAR solutions.

Consider a typical docking experiment that produces a vast amount of scoring values that are usually utilized in a very limited way, such that the docking scores are usually used only to identify potential binders upon the completion of the docking. In this context, we propose that the generated docking scores can be used “on the fly” to gain insight into the factors that determine successful docking and to create intermediate ligand-based QSAR models that allow us to reduce the amount of remaining, queried docking jobs. It is hard to expect a priori, that such QSAR solutions will have very high predictive power, but nonetheless, we anticipated that they would be sufficient for identifying the most probable nonbinders in the docking set.

**Compound Database.** For the purpose of this study, we selected a database of compounds offered by the NCI.<sup>49</sup> The original set of 223 536 entries has been reduced to 89 941 compounds by applying the expanded Lipinski’s drug-likeness criteria: molecular weight between 300 and 800 Da; the presence of 1–10 hydrogen-bond acceptors and 1–5 hydrogen-bond donors; less than 10 rotatable bonds; and overall hydrophobicity below  $\log P = 5.0$ . For all 89 941 nonsteroidal structures from the NCI set satisfying the above criteria, we calculated 28 inductive parameters for further QSAR modeling (outlined in Table 1).

Thus, 90 184 substances, including NCI drug-like compounds and compounds from the expanded steroids benchmark set,<sup>10</sup> have all been docked into the 1KDM active site using the Glide 2.7 protocol. The resulting GlideScore values have then been

used to simulate the progressive-docking procedure. In particular, we investigated how the SHBG ligands could have been recovered (compared to conventional docking) if we had incorporated intermediate QSAR solutions into the docking pipeline. The idea was to determine if a significant number of nonbinders could be rejected without docking them into the protein’s active site, while preserving the true binders (compounds with the GlideScore < -8.5).

**Database Clustering.** To maximize the range of GlideScore values to be used for QSAR modeling within a drawn sample, the original docking database consisting of 90 272 entries was clustered using QSAR descriptors and the value-based clustering procedure implemented in MOE<sup>50</sup> (with a Tanimoto coefficient of 0.85). This approach enabled us to implement a sampling algorithm that selects the most diverse set of chemical structures from the database. In particular, we developed the SVL-script that draws upon a defined number of compounds (such as 1000, 5000, 10 000, and 20 000) from the clustered database in a way that the sample is maximally represented within all database clusters. In this way, the QSAR solutions created for the sampled GlideScore values could be expected to cover the entire descriptor space of the docking database and could be applied to untested database entries.

**Simulated Progressive Docking.** Figure 3 illustrates the progressive-docking procedure we have developed. The process begins with the previously mentioned computation of inductive descriptors for all database entries and subsequent descriptor-based database clustering (steps 2–3). Other predocking filters, such as drug-likeness criteria, can also be incorporated into the process (marked as step 1 on the chart). It is perhaps worth mentioning that this procedure is not limited by the nature of compounds in the docking database and can be applied to substantially focused libraries enriched by various predocking filters.

At the following step 4, the initial set of compounds, covering the chemical space of the docking database, gets selected. At step 5, they undergo docking with the target using Glide software to produce the initial set of GlideScore values.

During the following step 6, all significant docking scores

**Table 1.** Statistical Parameters of QSAR Models Relating GlideScore Values to Inductive Descriptors<sup>a</sup>

	step 1	step 2	step 3	step 4	step 5	step 6	total
comps docked	10 083	20 328	30 216	37 251	45 805	51 366	90 184
comps docked efficiently (GScore < -4)	688	1403	2105	3101	3257	3676	5653
TP (GScore < -8.5)	27	58	89	111	137	155	281
<b>expected</b>							
TP (GScore < -8.5)	31	69	120	158	223	270	
<b>observed</b>							
false negatives	0	2	4	10	11	11	
Comps rejected from docking	16 020	26 787	33 423	37 428	38 818		
RMSE	0.77	0.79	0.80	0.80	0.79		
Q <sup>2</sup>	0.61	0.56	0.53	0.63	0.49		
intercept	6.97	-5.80	-9.26	-7.77	-11.16		
average_EO_neg	-0.65	-0.46	-0.37	-0.43	-0.34		
average_EO_pos	-7.45	-3.91	-2.90	-4.46	-1.77		
average_neg_charge	1.33	0.77	0.48	0.54	0.77		
average_pos_charge	-0.77	-6.16	-7.72	-3.04	-7.92		
EO_equalized	0.96	0.63	1.05	2.67	0.56		
global_hardness	18.65	34.04	33.28	23.32	34.21		
hardness_of_most_pos	-0.35	-0.10	-0.15	0.04	0.27		
largest_neg_softness	0.01	0.02	0.01	0.03	0.05		
largest_rs_i_mol	4.58	6.97	7.36	5.42	8.14		
largest_rs_mol_i	0.36	1.05	0.89	0.58	0.97		
most_neg_charge	0.71	-0.26	-0.24	-0.46	-0.24		
most_neg_rs_mol_i	-1.54	-1.29	-1.27	-1.34	-1.16		
most_neg_sigma_i_mol	0.58	-0.06	-0.23	-0.26	-0.59		
most_neg_sigma_mol_i	-0.38	0.04	0.05	0.18	0.10		
most_pos_charge	-0.16	-0.28	-0.34	-0.38	-0.42		
most_pos_rs_i_mol	0.17	0.31	0.33	0.31	0.29		
most_pos_sigma_i_mol	-0.01	0.11	0.12	0.21	0.15		
most_pos_sigma_mol_i	2.46	1.14	0.90	0.69	0.43		
smallest_pos_softness	-0.85	-0.63	-0.37	-0.14	-0.46		
softness_of_most_neg	0.05	0.10	0.12	0.11	0.12		
sum_neg_hardness	-0.23	0.22	0.18	0.06	0.29		
sum_pos_hardness	0.14	0.32	0.30	0.30	0.29		
total_charge	-0.15	0.01	0.00	-0.14	0.00		
total_neg_softness	0.00	-0.01	0.00	0.00	-0.01		
total_pos_softness	0.01	0.00	0.00	0.00	0.00		
total_sigma_mol_i	-0.42	-0.31	-0.28	-0.33	-0.25		

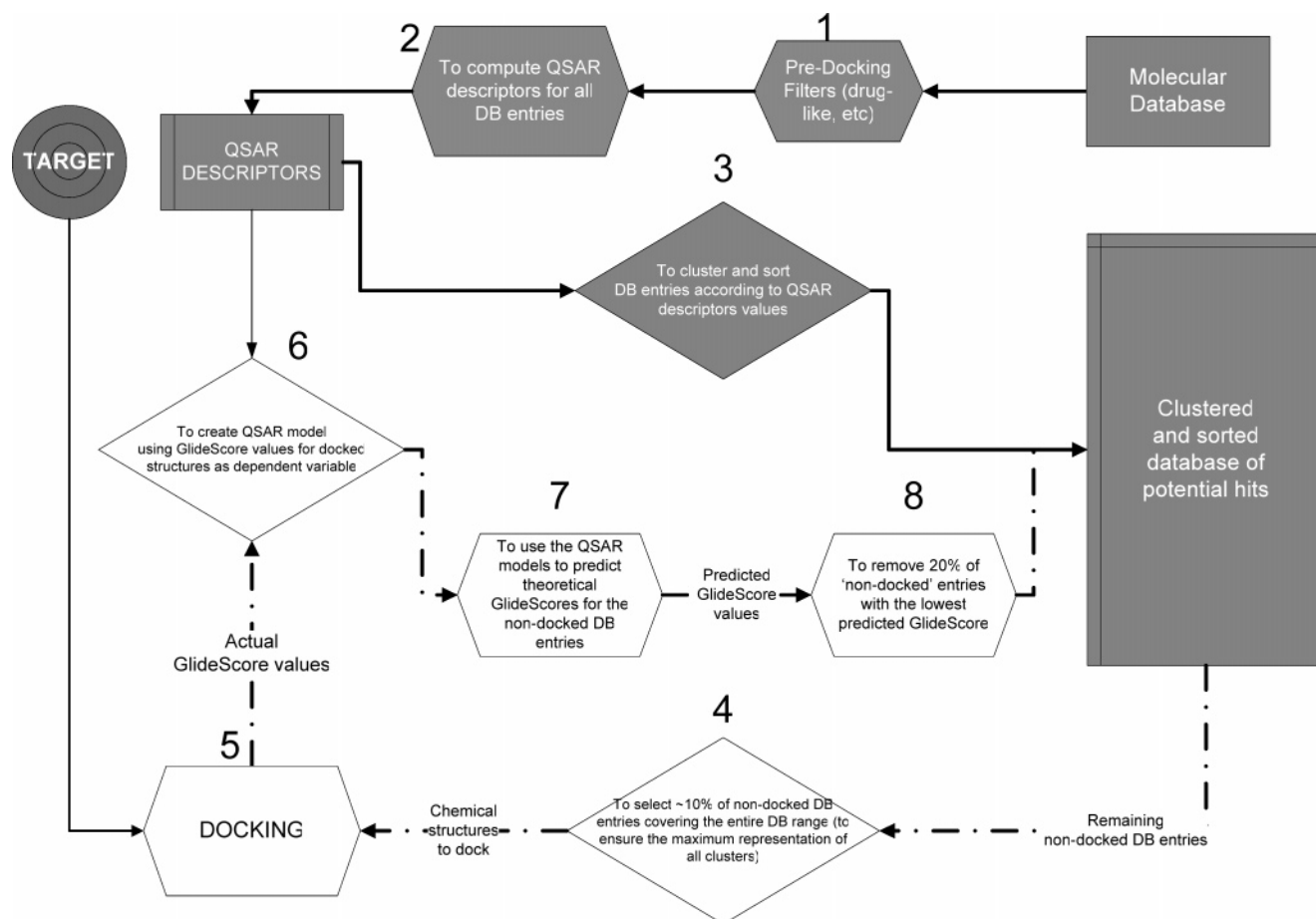
<sup>a</sup>The reported solution corresponds to the progressive docking with the 1KDM structure, QSAR sampling for every 10 000 docked substances, and with 20% rejection rate.

are utilized to build a linear QSAR model approximating these values by QSAR descriptors computed for the corresponding molecules. As Figure 3 illustrates, the resulting QSAR solution is then used at step 7 to estimate hypothetical GlideScores for the remaining undocked chemical structures in the queried database. Based on these predicted values, unprocessed database entries that produce the worst docking scores are removed from further docking (step 8). Steps 4–8 of the progressive-docking procedure can then be repeated until all entries of the docking database are processed. At each cycle, a fraction of the remaining undocked entries are removed. We investigated two possibilities, that is, when 10% or 20% of entries with the most positive predicted docking scores are removed from consideration. Figure 4 illustrates the QSAR-predicted GlideScore for ~80 000 compounds from the NCI drug-like database plotted against the corresponding empirically established values. The QSAR model created for the initially docked 10 000 structures has been used for the prediction; the light-blue color codes are for those 20% of the 80 000 untested entries that are rejected from further analysis. As the graph illustrates, none of ~16 000 substances rejected at this stage of progressive docking produced significant (< -8.5) GlideScore values. The chart also indicates that there is a general correspondence between the predicted and the actual docking scores for the substances studied.

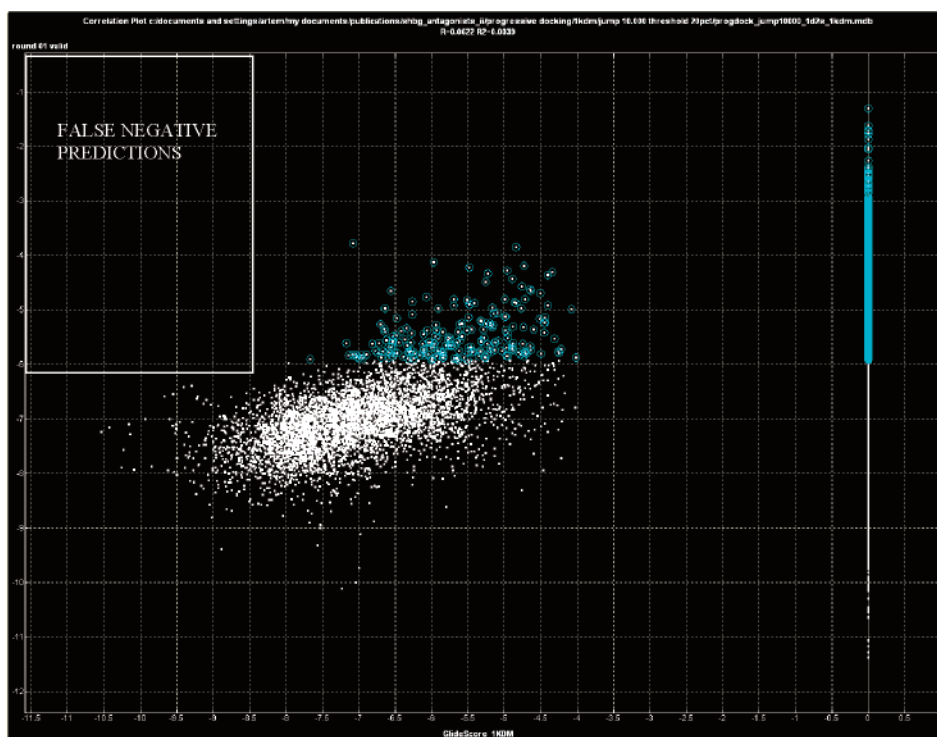
The initial steps 1–3 of the progressive-docking protocol described above are presented in bold solid lines in Figure 3. All subsequent steps (steps 4–8 of the procedure) that occur in a recursive manner are shown in dashed contours. We imple-

mented this type of recursive protocol to establish a gradual enrichment of QSAR solutions by empirical data (docking scores). This allows a more conservative rejection of undocked entries and avoids false negative predictions. On the other hand, more aggressive approaches could be implemented: one can use a limited set of *X* entries (covering the descriptors space of the database) to create a QSAR model that can be approximate GlideScore values for the remaining 90 - *X* database entries. Although such an approximation might be much faster, it could only be used if the quality of the initial QSAR model is very high. Otherwise, such a procedure will likely result in a number of false negative predictions and reject many potential hits. It is also apparent that a reasonable balance between data processing speed and an ability to recover true positives by the progressive-docking procedure will depend on the choice of rejection criteria and the sampling set size.

As mentioned above, we considered a 10% and 20% rejection criteria and sample sizes varying from 1000, 5000, 10 000, and 20 000 substances and evaluated the simulated recovery of potential SHBG ligands (GlideScore < -8.5) from the original set of 90 184 molecules. Figure 5 illustrates the results of simulated progressive docking with the 1D2S structure and the NCI drug-like dataset using various settings. The horizontal axis corresponds to the number of the database entries processed by the Glide program, and the vertical axis reflects the number of the true positives recovered. It is important to note that Figure 5 does not illustrate the true positive content of the already processed and sorted docking database, as in conventional decoy



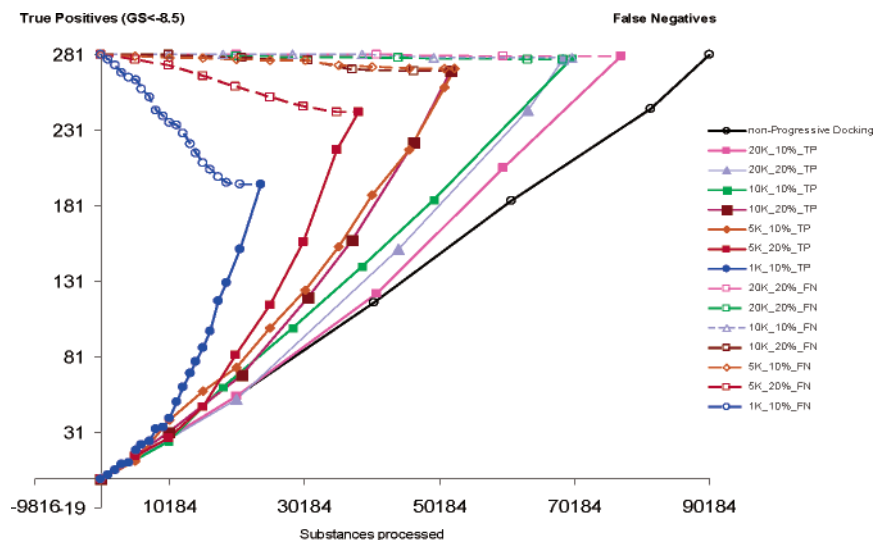
**Figure 3.** General scheme for progressive docking. The recursive cycle is illustrated by the dashed lines and noncolored objects, and the preparation cycle is illustrated by the solid lines and grayed objects. Other scheme notations: rectangles, storage; diamonds, decisions; and hexagons, processes.



**Figure 4.** Experimental vs QSAR-predicted GlideScore values for the entries of NCI drug-like database (first round of progressive docking). The 20% of nondocked entries to be excluded from the databases are color-coded in light blue.

docking studies; instead, it features the *actual* recovery of potential binders (compounds with the Glide Score <  $-8.5$ ) during the routine.

Thus, it is understandable that a similar nonprogressive-docking protocol, which processed a randomized docking database in sequential order, resulted in a linear dependence



**Figure 5.** Recovery of 1KDM effective binders (GlideScore < -8.5) from the NCI set of drug-like compounds using various progressive-docking settings (sample size and rejection percentage).

**Table 2.** Characteristics of the Progressive-Docking Performance Using Various Sample-Size and Rejection Cutoff Settings

progressive-docking settings (sampling size; rejection threshold)	docked substances	TP	FN	% yield	% hit rate	enrichment
20 K; 10%	77 012	280	10	36	99.64	1.17
20 K; 20%	68 447	278	3	41	98.93	1.30
10 K; 10%	69 797	279	2	40	99.29	1.28
10 K; 20%	51 883	270	11	52	96.09	1.67
5 K; 10%	52 353	272	9	52	96.80	1.67
5 K; 20%	38 163	243	38	64	86.48	2.04
1 K; 10%	23 742	195	86	82	69.40	2.64

between the number of processed compounds and identified SHBG hits, as indicated in black in Figure 5. It is also apparent that the smaller size of the sampling set and the larger rate of rejection of nondocked entries promotes faster processing of molecular database and results in greater numbers of false negatives (compounds that could be docked sufficiently but were underestimated by the QSAR models).

The data in Figure 5 also illustrate that the sampling set of 10 000 entries (roughly 10% of the docking database) and a 20% rejection criteria results in the most balanced docking outcomes: these settings allowed us to reject 43% of compounds from the docking set (38 818 out of 90 184 entries) and to recover 270 of the successfully docked substances, while nonprogressive docking identified 281 substances. The statistical parameters of linear QSAR solutions and the number of true positive SHBG hits estimated at every step of progressive docking with these settings are summarized in Table 1.

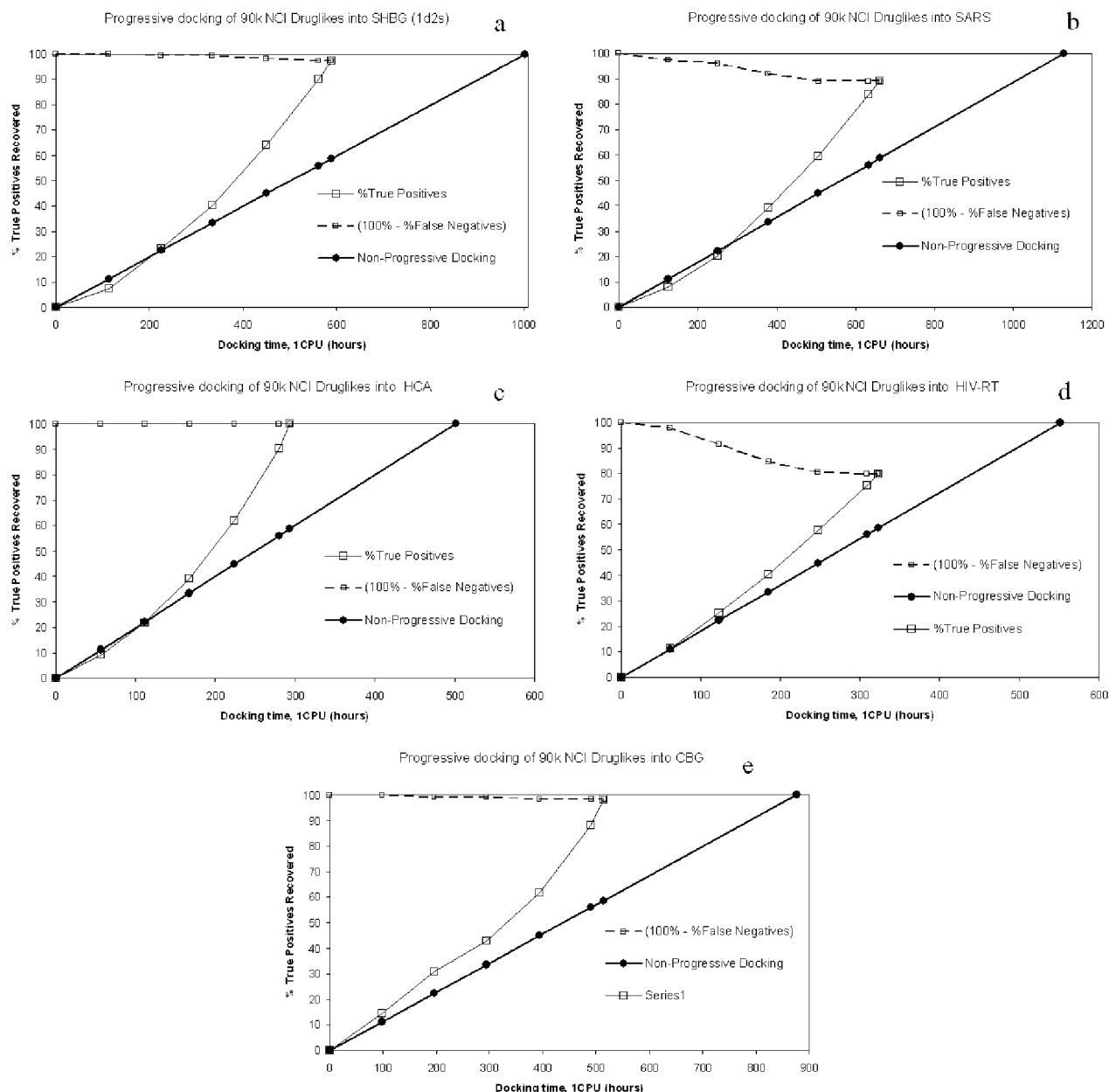
The quality of QSAR solutions featured in Table 1 is reasonable at all stages. The results also indicate that the progressive-docking protocol allowed processing of less than 60% of the original docking dataset (51 366 out of 90 272 entries), and this resulted in a 70% enrichment of the docking database without loss of significant useful information, that is, only 11 false negatives corresponding to a 97% hit rate. As Figure 5 illustrates, a less-conservative setting, such as a 5000 sample size and a 20% rejection criteria, can speed up the docking by 2-fold and still ensures a good enrichment of potential hits (86% hit rate). Figure 5 also illustrates the generation of false negative predictions by the progressive-docking protocol. As this shows, all progressive-docking experiments allowed an exponential recovery of true positive

predictions, as opposed to the linear hit recovery by traditional docking, and that the number of false negative predictions generated was not very high in most cases (only the protocol utilizing a 1000 sampling set resulted in a higher false positive rate).

The performance criteria of the progressive-docking procedure (hit rate, enrichment factor, true positive predictions, etc) are summarized in Table 2.

It is important to emphasize that the established progressive-docking *hit rate* and *enrichment factor* parameters (Table 2) possess different meaning when compared to conventional docking studies operating on decoy sets of compounds. Typically, to assess the performance of docking or other database searching procedures of a protein, authors mix a set of already known ligands with a limited number of presumably inactive compounds (often structurally similar to the target binders) and assess how the known binders are being recovered (for recent studies see refs 38–40). While this approach is adequate, it possesses certain drawbacks, namely, the decoy sets tend to be biased toward certain types of compounds (as known active substances are usually structurally similar), while inactivity of the negative control substances is only assumed. In contrast, we investigated the ability of the progressive-docking procedure to recover *potential binders* from a set of compounds solely by docking score. The fact that we used a large dataset of 90 000 entries possessing a significant level of noise also made the simulation more relevant to in silico high throughput studies.

As Table 2 illustrates, progressive docking could achieve a 1.2–2.0-fold enrichment of *potential binders*, while maintaining high (87–99%) hit rates. Thus, the use of a 10 000 sampling size and 20% rejection criteria resulted in a very balanced 96%



**Figure 6.** Recovery of true positives (effective binders) and production of false negatives (rejected effective binders) from the NCI set of drug-like compounds, using a progressive-docking procedure that is based on the QSAR sampling of every 10 000 docked structures and rejecting 20% of the remaining nondocked molecules. The studied targets: (a) SHBG 1D2S PDB structure, (b) SARS protein, (c) HCA, (d) HIV reverse transcriptase protein, and (e) CBG.

hit rate and 67% enrichment, as it allowed filtering out 38 301 of 90 184 molecules. These particular settings have been used in the following progressive-docking experiments that also involved the drug-like NCI dataset, which has been docked into several nonrelated and diverse targets. By conducting this study, we expected to further validate the developed hybrid QSAR/docking approach.

#### Validation of Progressive Docking on Additional Targets.

The above developed procedure and settings have been applied to several additional targets that included human carbonic anhydrase (HCA), SARS CoV protease (3CL), HIV viral reverse transcriptase (HIV-RT), and human corticosteroid-binding globulin (CBG). The choice of these proteins aimed to explore applicability of progressive docking to very diverse targets with different binding affinity ranges. Thus, the HCA protein has a relatively small and mostly polar cavity that includes a zinc

atom, but it can also accommodate ligands with long aliphatic chain fragments,<sup>51</sup> while CBG is similar to the previously described SHBG protein in so much as its ligand-binding affinity can be significantly influenced by minor substitutions. However, CBG ligands have more rotatable bonds, when compared to the relatively flat testosterone and estrogen derivatives, and it exhibits very different affinity patterns when compared to SHBG.<sup>52</sup> In one of the studied viral systems, 3CL, its ligands bind to the protein via covalent interactions, while respecting strict steric complementarity,<sup>53</sup> whereas the HIV-RT possess a very large active site formed by the interacting subunits of the homodimer and represents a known challenge for docking studies.<sup>54</sup>

The nature of known ligands for the selected targets is also very different (thus HCA binds small polar chemicals containing few rings, while 3CL and HIV-RT tend to interact with flexible

**Table 3.** Characteristics of the Progressive-Docking Performance on Five Studied Targets Using 20% Rejection Criteria and 10 000 Docking Sampling

target	docked molecules	rejected molecules	time spent (hours)	time saved (hours)	TP	FN	hit rate (%)
SHBG (1d2s)	52 893	37 379	588	415	147	4	97.35
CBG	52 931	37 341	515	363	131	2	98.50
HCA	52 858	37 414	294	208	958	1	99.90
HIV-RT	52 914	37 358	323	228	3959	1002	79.80
SARS	52 930	37 342	662	467	66	8	89.19

poly-ring systems, and CBG strongly prefers corticosteroids when compared to many other steroid classes). In addition, their ligand-binding affinities vary significantly from low nanomolar for CBG to high micromolar levels in the case of 3CT.

The developed progressive-docking procedure described in the previous sections has been applied to all four additional targets in the following way. First, for each of the proteins, we identified a set of known ligands with established binding affinities or corresponding *in vitro* inhibiting potentials.<sup>51–54</sup> The identified structures were then added to the original docking database of NCI drug-like molecules and steroids, expanding the dataset to 90 272 entries.

All targets were preprocessed for further docking (more details are in the Materials and Methods section), and all self-ligands were docked into the proteins using the Glide 2.7 program. The resulting dependences between the docking scores and the corresponding binding affinity/activity values for 3CL, HCA, HIV-RT, and CBG proteins were plotted into panels b–e of Figure 2, which also contain the superimposed structures of corresponding natives ligands identified from crystal structures and by the docking. As Figure 2 illustrates for all proteins (except CBG, where a homology model was used), the Glide 2.7 program reliably reproduced the crystal ligand orientations and affinity trends for known binders.

At the next step, we applied the progressive-docking routine, illustrated in Figure 3, to all four proteins, using the sampling step of 10 000 molecules and a 20% rejection criteria, as considered optimal in the previous sections. We have also expanded the range of QSAR descriptors and utilized 58 noncorrelated parameters that included 26 inductive descriptors for progressive docking with 1KDM, as well as 32 additional parameters representing conventional QSAR descriptors implemented with in the MOE package,<sup>50</sup> as described in Materials and Methods. As in the case of docking into the 1KDM structure, we constructed the curves of the true-positive (GlideScore < -8.5) recovery presented in Figure 6.

To illustrate the actual time-saving capabilities of the progressive-docking procedure, we transformed the horizontal axes of the recovery curves into actual docking times. Thus, according to our estimates, Glide 2.7 requires 40 s, 45 s, 22 s, 22 s, and 35 s to dock one ligand into the SHBG, 3CL, HIV-RT, HCA, and CBG active sites, respectively (CPU specifications: P4, 2 GHz, 512 MB RAM). Thus, Figure 6 illustrates that progressive docking saved 415 h on the SHBG target (out of 1003 single CPU hours required for docking of 90 272 compounds), 363 h on CBG (out of 878 h required), 206 h on HCA (out of 502 h required), 228 h on HIV-RT (out of 551 h required), and 467 h for 3CL (out of 1120 h required). Thus, hundreds of CPU processing hours can be saved using intermediate QSAR solutions in relation to generating docking scores. It is worth mentioning, of course, that the total time required for computing the QSAR descriptors, database clustering, the automated creation of QSAR solutions, and the filtering out predicted nonbinders from the docking database did not exceed several minutes.

Considering the fact that the Glide running time may reach

4–5 min per ligand depending on the setting,<sup>47</sup> the overall time-saving provided by the hybrid QSAR/docking procedure we have developed could be even more significant. From our own experience, in those cases when the docking site is considered flexible, the Glide may require ~10 min per molecule and, therefore, the progressive docking may save thousands of hours of docking even for a modest-sized database of ~100 000 molecules.

Following the completion of progressive docking with HIV-RT, 3CL, HCA, and CBG, we docked all the rejected compounds (on average 37 000 per target, see Table 3) to estimate the number of potential binders rejected by the QSAR modeling, that is, to establish the rates of false negative predictions.

The estimated numbers of molecules rejected by the QSAR, but which could nonetheless be docked into the corresponding targets with GlideScore < -8.5 are presented in Table 3. The false negative predictions accumulated by the progressive docking on 3CL, HCA, CBG, and HIV-RT are included in the respective panels of Figure 6.

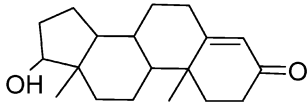
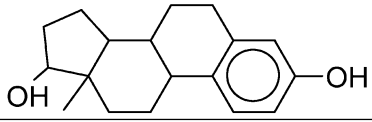
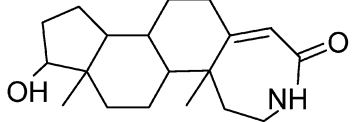
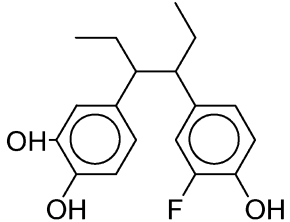
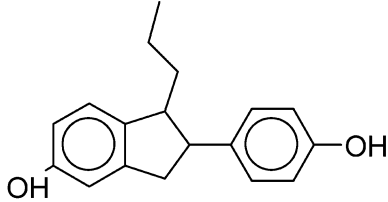
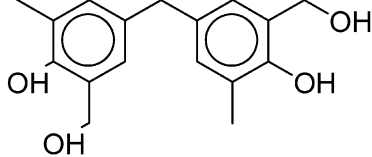
As the figure and Table 3 illustrate, the hybrid QSAR/docking procedure managed to preserve of 80–99% of all true hits. Thus, out of 37 341 molecular structures removed from the docking database of the CBG target, only two compounds could, in fact, be successfully docked into the protein. The progressive docking with the HCA resulted in losing only 1 out of 958 potential binders (compounds that could be docked with GlideScore < -8.5). The application of progressive docking to 3CL and HIV-RT proteins produced more significant numbers of false negative predictions, and these could be determined by less than optimal performance of the Glide on these targets. Nonetheless, even for the 3CL and HIV-RT systems, the progressive-docking hit rates remained at reasonably high 90 and 80% levels, respectively.

Thus, the performance of the developed hybrid procedure on some diverse target structures demonstrated that it saves up to 40% of the Glide processing time (transforming into hundreds of CPU hours), while still recovering 80–100% of potential docking hits. This, in our opinion, makes this new approach an attractive accessory for more accurate but slower HTS protocols involving flexible docking and/or *in silico* experiments involving large molecular datasets.

**Progressive Docking on the 1D2S Structure.** Finally, the developed progressive-docking procedure has been applied to find those chemicals among the 90 272 substances studied that would fit the active site of the 1D2S SHBG structure. This structure was selected to complement a previous study involving the 1KDM structure, because 1D2S features the loop segment Pro130–Arg135' that is believed to be important for steroid binding.<sup>29</sup> All the progressive-docking steps described in the previous section have been applied: at each step, 10 000 compounds that maximally covered all QSAR clusters of the NCI drug-like database were drawn and docked into 1D2S. The resulting GlideScore values were then used to create the QSAR model with inductive descriptors, which allow hypothetical docking scores to be established for the remaining untested substances and to, thereby, remove 20% of the remaining entries



**Table 4.** Experimental Relative Binding Affinities (as Compared to Testosterone), IC<sub>50</sub> Values, and Association Constant K<sub>a</sub> Values Established for Most Active Leads Identified from the NCI Drug-Like Dataset

compound	formula	RBA (testosterone)	*IC <sub>50</sub> ( $\mu$ M)	K <sub>a</sub> M <sup>-1</sup>
testosterone		1	0.014	1.1 x 10 <sup>9</sup>
estradiol		3.5	0.05	6 x 10 <sup>8</sup>
NSC 627265		66	0.95	1.54 x 10 <sup>7</sup>
NSC 32082		714	10.35	1.42 x 10 <sup>6</sup>
NSC 34319		768	11.14	1.32 x 10 <sup>6</sup>
NSC 48435		3326	48.23	3.04 x 10 <sup>5</sup>

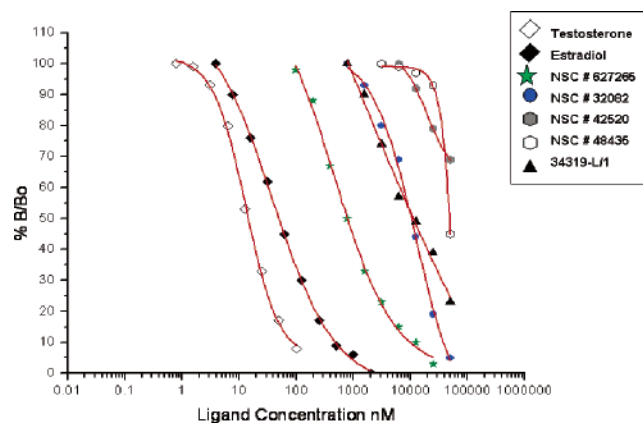
from further consideration. Five consecutive iteration steps 4–8 were taken during the progressive docking with 1D2S, and a total of 52 983 substances (out of 90 272) have been docked into the protein's active site (see Table 3). As in previously studies, the recovery of 1D2S hits (compounds docked with the GlideScore < -8.5) is illustrated in Figure 6a, and this also features the linear trend for conventional docking recovery of potential hits.

As a result, 147 potential 1D2S binders were identified, and the most promising hits identified for the 1KDM and 1D2S crystal structures of the SHBG protein were selected for experimental evaluation. To accomplish this, we utilized the less stringent cutoff values of GlideScore = -8.5 to maximize the number of potential SHBG ligands identified.

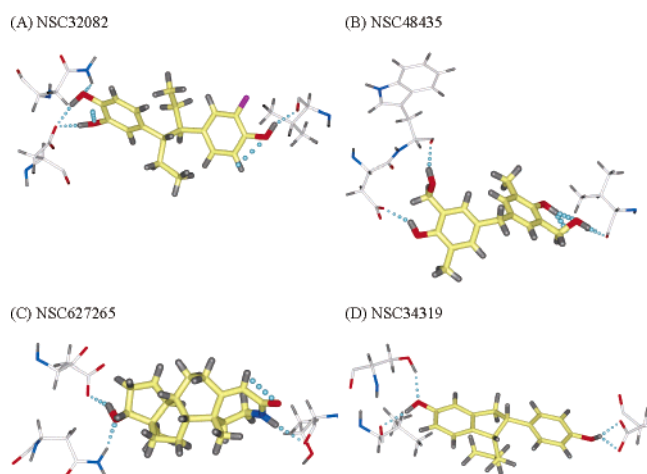
**Experimental Testing of the Docking Hits.** The docking results for 1KDM and 1D2S differed, and this may be attributed to the complex nature of SHBG ligand binding and/or the possibility that bound ligands induce different changes in the active sites of these protein structures. We chose 16 top-ranked nonsteroidal substances from the 1KDM and 1D2S docking experiments for in vitro testing. All 16 compounds (NSC: 74429, 367779, 376464, 627265, 2801-Z/2, 36398-U/1, 59266-A/2, 309136-Z/1, 350993-W/1, 32082-F/1, 34319, 42520-Y/1, 48435-F/2, 105825-L/2, 118073-X/1, 167385-X/1) were tested

for their ability to interact with the SHBG ligand-binding site by using a competitive binding assay that employs tritium-labeled dihydrotestosterone (<sup>3</sup>H]DHT; see Materials and Methods for details). The initial screening of compounds was conducted at a single high concentration (approximately 200  $\mu$ M), and the results demonstrate that three nonsteroidal compounds (NSC: 32082, 34319, 48435) and one steroid-like substance (NSC627265) displace up to 35–95% of the [<sup>3</sup>H]-DHT from the SHBG steroid-binding site (see Table 4). These compounds were then selected for a more detailed analysis of their ability to compete [<sup>3</sup>H]DHT from the human SHBG steroid-binding site, relative to known concentrations of the physiologically most important androgen (testosterone) and estrogen (17 $\beta$ -estradiol). The resulting competitive displacement curves generated using these test compounds (see Figure 7) illustrate that their potencies as SHBG ligands are in line with their rank potencies in the preliminary screening assay.

The structures of the top 4 binders (NSC: 627265, 32082, 34319, and 48435) are presented in Table 4 along with the corresponding IC<sub>50</sub> values calculated from the displacement curve shown in Figure 7. These IC<sub>50</sub> values (NSC627265 = 950 nM, NSC32082 = 10.35  $\mu$ M, NSC34319 = 11.14  $\mu$ M, NSC48435 = 48.23  $\mu$ M) were compared with those of testosterone (15.6 nM) and estradiol (61.6 nM), and relative binding



**Figure 7.** Displacement curves for test compounds used in the in vitro competition assay to determine the relative binding affinities of human SHBG ligands. The amount of [ $^3\text{H}$ ] testosterone bound to SHBG in the presence of increasing concentrations of competitor ligands (B).



**Figure 8.** Poses of the four most active hit compounds docked in the active sites of their respective target protein structures. Each compound is shown as thicker, yellow sticks, and the protein residues each compound is thought to interact with are shown as thinner, white sticks. Potential hydrogen bonds formed in the interaction are shown as blue circles. (A) NSC32082, (B) NSC48435, (C) NSC627265 are docked in the active site of 1D2S, and (D) NSC34319 is docked in the active site of 1KDM.

affinity (RBA) values were calculated using testosterone as the reference (Table 4).

Thus, we have identified four compounds that bind SHBG with binding affinities in the low micromolar range. The binding affinity of the best binder, NSC627265, is only about 1 order of magnitude weaker than that of estradiol. However, this is not surprising because this compound closely resembles a steroid, even though it does not contain a typical steroid ring structure, which is why it was not filtered out from the docking dataset. The next best binders, NSC32082 and NSC34319, are more than 2 orders of magnitude weaker in binding affinity than estradiol, with estimated  $K_a$  values of around  $1.4 \times 10^6 \text{ M}^{-1}$ . These two compounds are stronger SHBG binders than other nonsteroidal ligands identified in previous studies, where the highest  $K_a$  was  $1.2 \times 10^6 \text{ M}^{-1}$ .<sup>8</sup> It should also be mentioned that NSC32082 and NSC34319 bind SHBG almost as effectively as a natural lignan derivative<sup>55</sup> that is the best nonsteroidal SHBG binder reported to date with a  $K_a$  of  $3.2 (\pm 1.7) \times 10^6 \text{ M}^{-1}$ .<sup>55</sup> Moreover, NSC32082 and NSC34319 are synthetic chemicals that are easier to produce and are amenable to further structural modifications.

The four SHBG hits we have identified are predicted to reside within the SHBG steroid-binding pocket in a similar manner as the biological sex steroid ligands. This is illustrated in Figure 8, which features the hydrogen bonds that likely exist between the compounds and residues within the 1KDM active site. Notably, each of the compounds are predicted to form at least one hydrogen bond on either end anchoring them to SHBG, just as steroids form hydrogen bonds with SHBG in the PDB structures.

To summarize, the hybrid procedure we have developed accomplishes one of the main purposes of docking, that is, the rejection of predicted nonbinders. The procedure allows a reduction in the amount of docking computations by 1.2–2 times while still allowing 86–99% recovery of all *potential* binders from the database. Its application has revealed four new SHBG ligands that are among the most active nonsteroidal hits for the target identified to date.

## Conclusions

The Glide program<sup>44</sup> we used in the current study approximates a systematic search of positional, orientational, and conformational space of the docked ligand. This methodology has enabled Glide to perform favorably in several recent comparative docking studies,<sup>56</sup> including our own.<sup>10</sup> However, the Glide protocol and other accurate docking programs are characterized by relatively low data processing speed, typically requiring a few minutes of computation per ligand. Thus, the speed of reliable docking becomes a critical challenge for in silico high throughput investigations because conventional molecular datasets routinely include several million molecular structures. Numerous computational approaches aim to reduce the amount of computational screening required, including predocking filters, database clustering, and parallelization protocols. The incorporation of QSAR solutions, approximating hypothetical docking scores, may represent yet another strategy for reducing computational time. Such QSAR strategies can be used in addition to other conventional predocking filters, such as those based on molecular size, volume, or drug-like criteria.

For each of the six studied targets, the developed procedure saved hundreds of hours in terms of the docking time required to process  $\sim 90\,000$  potential ligands, while maintaining 80–100% hit recovery rates. Thus, the developed progressive-docking procedure is an effective approach for accelerating high throughput screening, especially when applied to accurate but slower docking approaches and for large-scale studies. The applicability of advanced, possibly nonlinear QSAR techniques, such as random forest, support vector machines, bayesian networks, and artificial neural networks should also be investigated. For the sake of the current study, we only used linear data fitting with the PLS approach as we implemented rather conservative rejection criteria for predicted nonbinders. If one desires more aggressive progressive docking, special attention should be paid to the high quality of the intermediate QSAR solutions.

It is possible that the progressive-docking approach may not be suitable for docking systems allowing multiple and diverse binding poses of potential ligands, but this remains to be investigated.

## Materials and Methods

**Inductive Descriptors.** In summary, the inductive QSAR variables can be computed by the following equations

$$R_{S_{j \rightarrow N-1}} = \alpha \sum_{i \neq j} \frac{R_j^2}{r_{j-i}^2} = \alpha R_j^2 \sum_{i \neq j} \frac{1}{r_{j-i}^2} \quad R_{S_{G \rightarrow j}} = \alpha \sum_{i \in G, i \neq j} \frac{R_i^2}{r_{i-j}^2} \quad (1)$$

$$\sigma_{j \rightarrow N-1}^* = \beta \sum_{i \neq j} \frac{(\chi_i^0 - \chi_j^0) R_j^2}{r_{j-i}^2} \quad \sigma_{G \rightarrow j}^* = \beta \sum_{i \in G, i \neq j} \frac{(\chi_i^0 - \chi_j^0) R_i^2}{r_{i-j}^2} \quad (2)$$

$$\chi_{N-1 \rightarrow}^0 = \left( \sum_{i \neq j} \frac{N-1 \chi_i^0 (R_i^2 + R_j^2)}{r_{i-j}^2} \right) / \left( \sum_{i \neq j} \frac{N-1 R_i^2 + R_j^2}{r_{i-j}^2} \right)$$

$$\chi_{N-1 \rightarrow j}^0 = \left( \sum_{i \neq j} \frac{N-1 \chi_i^0 (R_i^2 + R_j^2)}{r_{i-j}^2} \right) / \left( \sum_{i \neq j} \frac{N-1 R_i^2 + R_j^2}{r_{i-j}^2} \right) \quad (3)$$

$$\Delta N_j = Q_j + \gamma \sum_{i \neq j} \frac{N-1 (\chi_j - \chi_i) (R_j^2 + R_i^2)}{r_{j-i}^2} \quad Q_j = \text{formal charge of } j \quad (4)$$

$$\eta_j = 1 / \left( 2 \sum_{j \neq i} \frac{N-1 R_j^2 + R_i^2}{r_{j-i}^2} \right) \quad \eta_{\text{MOL}} = \frac{1}{s_{\text{MOL}}} = 1 / \left( 2 \sum_{j \neq i} \frac{N-1 R_j^2 + R_i^2}{r_{j-i}^2} \right) \quad (5)$$

$$s_i = 2 \sum_{j \neq i} \frac{N-1 R_j^2 + R_i^2}{r_{j-i}^2} \quad s_{\text{MOL}} = \sum_{j \neq i} \sum_{j \neq i} \frac{N-1 R_j^2 + R_i^2}{r_{j-i}^2} \quad (6)$$

where  $R$  is the covalent atomic radii,  $r$  is the interatomic distance,  $\chi$  is the inductive electronegativity,  $R_s$  is the steric constant,  $\sigma^*$  is the inductive constants,  $\Delta N$  is the inductive partial charge, and  $\eta$  and  $s$  are the inductive analogues of chemical hardness and softness.

It should be noted that the variables indexed with  $j$  subscript describe the influence of a single atom onto a group of atoms  $G$  (typically the rest of  $N$ -atomic molecule), while  $G$  indices designate group (molecular) quantities. The linear character of eqs 1–6 makes inductive descriptors readily computable and suitable for sizable databases and positions them as appropriate parameters for large-scale QSAR models.

**Target Preparation.** The Maestro suite<sup>57</sup> was used to prepare protein structures for docking. The following PDB files were used for targets: 1KWR for HCA, 1VRT for HIV-RT, 1KDM and 1D2S for SHBG, and 1UK4 for 3CL. The structure of human CBG was obtained by homology modeling on  $\alpha 1$ -antichymotrypsin serpin structure, (PDB entry 1QMN) using the MOE programs<sup>50</sup> with default settings.

From all PDB structures, water and ion molecules were removed and hydrogen atoms were added and adjusted where necessary. The ligand-binding sites were defined as 10 Å surrounding the cocrystallized ligands in the crystal structures. No water molecules or ions were retained in the active sites. In the case of the 3CL protein, the covalently bound ligand was removed from the protein during the docking site preparation.

**Molecular Docking.** The consequent docking has been conducted using the Glide 2.7 program, with the default settings and inductive partial charges assigned to protein molecules according to the previously published procedure.<sup>10</sup> The docking database has been separated into 10 equal parts that have been docked on 10 machines in parallel (PC specifications: Intel P4, 2.0 GHz processor, 512 MB RAM, Centos 4.0 OS).

**QSAR Descriptors Calculation and Model Building.** The optimized structures of 90 272 compounds were used for calculating 26 non-cross-correlating inductive QSAR descriptors<sup>1–5</sup> and 32 conventional QSAR parameters calculated by the MOE program (the corresponding description values can be obtained for authors upon request).

Inductive QSAR parameters used for creating the models: *Average\_EO\_Neg*, *Average\_EO\_Pos*, *Average\_Neg\_Charge*, *Average\_Pos\_Charge*, *EO\_Equalized*, *Global\_Hardness*, *Hardness\_of\_Most\_Pos*, *Largest\_Neg\_Softness*, *Largest\_Rs\_i\_mol*, *Largest\_Rs\_mol\_i*, *Most\_Neg\_Charge*, *Most\_Neg\_Rs\_mol\_i*, *Most\_Neg\_Sigma\_i\_mol*, *Most\_Neg\_Sigma\_mol\_i*, *Most\_Pos\_Charge*, *Most\_Pos\_Rs\_i\_mol*, *Most\_Pos\_Sigma\_i\_mol*, *Most\_Pos\_Sigma\_mol\_i*, *Smallest\_Pos\_Softness*, *Softness\_of\_Most\_Neg*, *Sum\_Neg\_Hardness*, *Sum\_Pos\_Hardness*, *Total\_Charge*, *Total\_Neg\_Softness*, *Total\_Pos\_Softness*, *Total\_Sigma\_mol\_i*.

MOE QSAR parameters used for creating the models: *b\_double*, *b\_rotN*, *b\_rotR*, *b\_triple*, *chiral*, *a\_nN*, *a\_nO*, *a\_nS*, *FCharge*, *lip\_don*, *KierFlex*, *a\_base*, *usa\_acc*, *usa\_acid*, *usa\_base*, *usa\_don*, *density*, *logP(o/w)*, *a\_ICM*, *chi1v\_C*, *chiral\_u*, *balabanJ*, *logS*, *ASA*, *ASA+*, *ASA-*, *ASA\_H*, *ASA\_P*, *CASA+*, *CASA-*, *DASA*, *DCASA*

For more details on inductive parameters, see references 1–5, while the conventional QSAR parameters correspond to notations implemented within the MOE program.<sup>50</sup>

The inductive QSAR descriptors were calculated by the custom SVL scripts of the MOE program, which can be downloaded through the SVL exchange.

**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.<sup>58</sup> In brief, the assay involved mixing 100  $\mu\text{L}$  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  $\mu\text{L}$  tritium-labeled DHT ( $[^3\text{H}]$  DHT) at 10 nM as labeled ligand. For the screening assay, triplicate aliquots (100  $\mu\text{L}$ ) of a fixed amount (200  $\mu\text{M}$ ) of test compound were added to this SHBG/ $[^3\text{H}]$  DHT mixture and incubated overnight at room temperature. After a further 10 min incubation at 0  $^\circ\text{C}$ , 500  $\mu\text{L}$  of a DCC slurry was added at 0  $^\circ\text{C}$  and incubated for 10 min prior to centrifugation to separate SHBG-bound from free  $[^3\text{H}]$  DHT. Compounds that displaced more than 35% of the  $[^3\text{H}]$  DHT from the SHBG in this assay were then diluted serially, and triplicate aliquots (100  $\mu\text{L}$ ) of known concentrations of test compounds were used to generate complete competition curves by incubation with the SHBG/ $[^3\text{H}]$  DHT mixture and separation of SHBG-bound from free  $[^3\text{H}]$  DHT, as in the screening assay. The amounts of  $[^3\text{H}]$  DHT bound to SHBG at each concentration of competitor ligand were determined by scintillation spectrophotometry and plotted in relation to the amount of  $[^3\text{H}]$  DHT bound to SHBG at zero concentration of competitor. From the resulting competition curves,  $\text{IC}_{50}$  concentrations could be calculated if displacement of more than 50% of  $[^3\text{H}]$  DHT from SHBG was achieved.

The association constants ( $K_a$ ) have been calculated from the RBA parameters using the following equation:  $K_a(\text{DHT}) / [(1 + R) / \text{RBA} - R]$ , where  $K_a(\text{DHT}) = 0.98 \times 10^9 \text{ M}^{-1}$  is the association constant of the DHT and  $R$  (0.05) is the ratio of bound-to-free tritium-labeled DHT 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, FB is the Canadian Institutes of Health Research Postdoctoral Scholar, and Y.Y.L. acknowledges the support of the CIHR/MSFHR Strategic Training Program in Bioinformatics (<http://bioinformatics.bcg-sc.ca>).

**Supporting Information Available:** The resulting GlideScores for the majority of compounds studied. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## References

- Cherkasov, A. R.; Jonsson, M.; Galkin, V. I. A Novel Approach to the Analysis of Substituent Effects. Quantitative Interpretation of Ionization Potentials and Gas Basicity of Amines. *J. Mol. Graphics Modell.* **1999**, *17*, 28–43.
- Cherkasov, A. R.; Galkin, V. I.; Cherkasov, R. A. A. New Approach to the Theoretical Estimation of Inductive Constants. *J. Phys. Org. Chem.* **1998**, *11*, 437–447.

- (3) Cherkasov, A. R.; Jonsson, M.; Galkin, V. I.; Cherkasov, R. A. Correlation Analysis in the Chemistry of Free Radicals. *Russ. Chem. Rev.* **2001**, *70*, 1–26.
- (4) Cherkasov, A.; Sprou, D.; Chen, R. 3D Correlation Analysis—New Method of Quantification of Substituent Effect. *J. Phys. Chem. A* **2003**, *107*, 9695–9704.
- (5) Cherkasov, A. R.; Galkin, V. I.; Cherkasov, R. A. “Inductive” Electronegativity Scale. *J. Mol. Struct.* **1999**, *489*, 43–46.
- (6) Cherkasov, A. R.; Galkin, V. I.; Cherkasov, R. A. “Inductive” Electronegativity Scale. 2. “Inductive” Analog of Chemical Hardness. *J. Mol. Struct.* **2000**, *497*, 115–123.
- (7) Cherkasov, A. Inductive Electronegativity Scale. Iterative Calculation of Inductive Partial Charges. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2039–2047.
- (8) “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.
- (9) Cherkasov, A. “Inductive” Descriptors. 10 Successful Years in QSAR. *Curr. Comput.-Aided Drug Des.* **2005**, *1*, 21–42.
- (10) Cherkasov, A.; Shi, Z.; Fallahi, M.; Hammond, G. Successful in Silico Discovery of Novel Nonsteroidal Ligands for Human Sex Hormone Binding Globulin (SHBG). *J. Med. Chem.* **2005**, *48*, 3203–3213.
- (11) Nisker, J. A.; Hammond, G. L.; Davidson, B. J.; Frumar, A. M.; Takaki, N. K.; Judd, H. L.; Siiteri, P. K. Serum Sex Hormone-Binding Globulin Capacity and the Percentage of Free Estradiol in Postmenopausal Women with and without Endometrial Carcinoma. A New Biochemical Basis for the Association between Obesity and Endometrial Carcinoma. *Am. J. Obstet. Gynecol.* **1980**, *138*, 637–642.
- (12) Hogeveen, K. N.; Cousin, P.; Pugeat, M.; Dewailey, D.; Soudan, B.; Hammond, G. L. Human Sex Hormone-Binding Globulin Variants Associated with Hyperandrogenism and Ovarian Dysfunction. *J. Clin. Invest.* **2002**, *109*, 973–981.
- (13) Anderson, D. C. Sex-Hormone-Binding Globulin. *Clin. Endocrinol.* **1974**, *3*, 69–96.
- (14) Van Pittelgergh, I.; Goemaere, S.; Zmierzczak, H.; Kaufman, J. M. Perturbed Sex Steroid Status in Men with Idiopathic Osteoporosis and their Sons. *J. Clin. Endocrinol. Metab.* **2004**, *89*, 4949–4953.
- (15) Rapuri, P. B.; Gallagher, J. C.; Haynatzki, G. Endogenous Levels of Serum Estradiol and Sex Hormone Binding Globulin Determine Bone Mineral Density, Bone Remodeling, the Rate of Bone Loss, and Response to Treatment with Estrogen in Elderly Women. *J. Clin. Endocrinol. Metab.* **2004**, *89*, 4954–4962.
- (16) Lindstedt, G.; Lundberg, P. A.; Lapidus, L.; Lundgren, L.; Björntrop, P. Low Sex-Hormone-Binding Globulin Concentration as Independent Risk Factor for Development of NIDDM. 12-Yr Follow-Up of Population Study of Women in Gothenburg, Sweden. *Diabetes* **1991**, *40*, 123–128.
- (17) Haffner, S. M.; Katz, M. S.; Stern, M. P.; Dunn, J. F. Association of Decreased Sex Hormone Binding Globulin and Cardiovascular Risk Factors. *Arteriosclerosis* **1989**, *9*, 136–143.
- (18) 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 Des.* **2004**, *18*, 175–187.
- (19) 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.
- (20) 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.
- (21) 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.
- (22) 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.
- (23) 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.
- (24) Polanski, J.; Walczak, B. The Comparative Molecular Surface Analysis (COMSA): A Novel Tool for Molecular Design. *Comput. Chem.* **2000**, *24*, 615–625.
- (25) Turner, D. B.; Willett, P.; Ferguson, A. M.; 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.
- (26) 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.
- (27) 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.
- (28) 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.
- (29) Grishkovskaya, I.; 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**, *277*, 32086–32093.
- (30) Avvakumov, G. V.; Grishkovskaya, I.; Muller, Y. A.; Hammond, G. L. Crystal Structure of Human Sex Hormone-Binding Globulin in Complex with 2-Methoxyestradiol Reveals the Molecular Basis for High Affinity Interactions with C-2 Derivatives of Estradiol. *J. Biol. Chem.* **2002**, *277*, 45219–45225.
- (31) Grishkovskaya, I.; Avvakumov, G. V.; Hammond, G. L.; Muller, Y. A. Resolution of a Disordered Region at the Entrance of the Human Sex Hormone-Binding Globulin Steroid-Binding Site. *J. Mol. Biol.* **2002**, *318*, 621–626.
- (32) Hammond, G. L.; Avvakumov, G. V.; Muller, Y. A. Structure/Function Analyses of Human Sex Hormone-Binding Globulin: Effects of Zinc on Steroid-Binding Specificity. *J. Steroid Biochem. Mol. Biol.* **2003**, *85*, 195–200.
- (33) *SNNS: Stuttgart Neural Network Simulator*, version 4.0, University of Stuttgart: Stuttgart, Germany, 1995.
- (34) Cherkasov, A. Inductive QSAR Descriptors. Distinguishing Compounds with Antibacterial Activity by Artificial Neural Networks. *Intern. J. Mol. Sci.* **2005**, *6*, 63–86.
- (35) Cherkasov, A. Can Bacterial-Metabolite-Likeness Model Improve Odds of *In silico* Antibiotic Discovery? *J. Chem. Inf. Model.* **2006**, *46*, 1222–1244.
- (36) Karakoc, E.; Cherkasov, A.; Sahinalp, S. C. Distance Based Algorithms for Small Biomolecule Classification and Structural Similarity Search. *Bioinformatics* **2006**, *22*, e243–251.
- (37) Karakoc, E.; 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*, in press.
- (38) Kairys, V.; Fernandes, M. X.; Gilson, M. K. Screening Drug-Like Compounds by Docking to Homologous Models: A Systematic Study. *J. Chem. Inf. Model.* **2006**, *46*, 365–379.
- (39) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual Screening Using Protein-Ligand Docking: Avoiding Artificial Enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793–780.
- (40) Wang, J.; Kang, X.; Kuntz, I. D.; Kollman, P. A. Hierarchical Database Screenings for HIV-1 Reverse Transcriptase Using a Pharmacophore Model, Rigid Docking, Solvation Docking, and MM-PB/SA. *J. Med. Chem.* **2005**, *48*, 2432–2444.
- (41) Medina-Franco, J. L.; Golbraikh, A.; Oloff, S.; Castillo, R.; Tropsha, A. Quantitative Structure-Activity Relationship Analysis of Pyridinone HIV-1 Reverse Transcriptase Inhibitors Using the k Nearest Neighbor Method and QSAR-Based Database Mining. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 229–242.
- (42) Sciabola, S.; Carosati, E.; Baroni, M.; Mannhold, R. Comparison of Ligand-Based and Structure-Based 3D-QSAR Approaches: A Case Study on (aryl)-Bridged 2-Aminobenzonitriles Inhibiting HIV-1 Reverse Transcriptase. *J. Med. Chem.* **2005**, *48*, 3756–3767.
- (43) Baber, J. C.; Shirley, W. A.; Gao, Y.; Feher, M. The Use of Consensus Scoring in Ligand-Based Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 277–288.
- (44) *Glide*, version 2.7, Schrödinger, Inc.: San Diego, CA, 2004.
- (45) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, *47*, 1750–1759.
- (46) 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 of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.
- (47) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative Evaluation of Eight Docking Tools for Docking and Virtual Screening Accuracy. *Proteins* **2004**, *57*, 225–242.
- (48) Fernandes, M. X.; Kairys, V.; Gilson, M. K. Comparing Ligand Interactions with Multiple Receptors via Serial Docking. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1961–1970.
- (49) National Cancer Institute (NCI) database: [http://dtp.nci.nih.gov/docs/dtp\\_search.html](http://dtp.nci.nih.gov/docs/dtp_search.html).

- (50) MOE: *Molecular Operational Environment*, version 2003.10; Chemical Computation Group, Inc.: Montreal, Canada, 2004.
- (51) Gruneberg, S.; Stubbs, M. T.; Klebe, G. Successful Virtual Screening for Novel Inhibitors of Human Carbonic Anhydrase: Strategy and Experimental Confirmation. *J. Med. Chem.* **2002**, *45*, 3588–3602.
- (52) Westphal, U. Steroid–Protein Interaction II. *Monographs in Endocrinology*; Springer-Verlag: Berlin and Heidelberg, Germany, 1986.
- (53) Tsai, K-C.; Chen, S-Y.; Liang, P-H.; Lu, I-L.; Mahindroo, N.; Hsieh, H-P.; Chao, Y-S.; Liu, L.; Liu, D.; Lien, W.; Lin, T-H.; Wu, S-Y. Discovery of a Novel Family of 3CL Protease Inhibitors by Virtual Screening and 3D-QSAR Studies. *J. Med. Chem.* **2006**, *49*, 3485–3495.
- (54) Maruyama, T.; Kozai, S.; Demizu, Y.; Witvrouw, M.; Pannecouque, C.; Balzarini, J.; Snoecks, R.; Anrei, C.; De Clercq, E. Synthesis of Anti-HIV-1 and Anti-HCMV Activity of 1-Substituted 3-(3,5 Dimethylbenzyl)uracil derivatives. *Chem. Pharm. Bull.* **2006**, *54*, 325–333.
- (55) Schottner, M.; Gansser, D.; Siteller, K. M. Lignans Interfering with 5  $\alpha$ -Dihydrotestosterone Binding to Human Sex Hormone-Binding Globulin. *J. Nat. Prod.* **1997**, *61*, 119–121.
- (56) Krovat, E. M.; Steindl, T.; Langer, T. Recent Advances in Docking and Scoring. *Curr. Comput.-Aided Drug Des.* **2005**, *1*, 93–102.
- (57) *Maestero*; Schrödinger, Inc.: San Diego, CA, 2004.
- (58) 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.

JM060961+