Using Molecular Docking, 3D-QSAR, and Cluster Analysis for Screening Structurally Diverse Data Sets of Pharmacological Interest

Osvaldo A. Santos-Filho* and Artem Cherkasov

Jack Bell Research Centre at Vancouver General Hospital, Faculty of Medicine, University of British Columbia, 2660 Oak Street, Vancouver, British Columbia V6H 3Z6, Canada

Received June 9, 2008

In this study, we propose a drug design approach which includes docking, molecular fingerprints based cluster analysis, and ‘induced’ descriptors based receptor-dependent 3D-QSAR. The method was shown to be very useful for screening and modeling structurally diverse data sets of pharmacological interest. Different from other receptor-dependent 3D-QSAR, no ambiguous alignments are required for the construction of the models, and the computational cost is relatively lower. Moreover, ‘induced’ descriptors were shown to be very powerful in “capturing” ligand−receptor intermolecular interactions. The methodology was validated for eight data sets sampled from the literature and from public databases: human sex hormone-binding globulin, human corticosteroid-binding globulin, anthrax lethal factor, HIV-1 reverse transcriptase, neuraminidase A, thrombin, trypsin, and *Pneumocystis carinii* dihydrofolate reductase data sets. The resulting models were interpretable; the constructed QSAR equations have high statistical significance and predictive strength; and the drug design solutions were shown to be useful for guiding ligand modification for the development of new inhibitors for a broad range of molecular targets.

INTRODUCTION

Three-dimensional quantitative structure−activity relationship (3D-QSAR)[1] analysis represents an important tool in the field of computer-aided drug design. This methodology is typically implemented in two distinct schemes: receptor-independent (RI) and -dependent (RD) QSAR. The first scheme neglects the structure of the molecular target, while the second one utilizes it for constructing the models.

With CoMFA,[2] a very popular 3D-QSAR method, each data set molecule is placed into a 3D grid. Then, at each grid point, steric (Lennard-Jones potential) and electrostatic energies are computed for each molecule ligand by a probe atom. As the result, steric and electrostatic molecular fields are computed and used as descriptors for the construction of the QSAR models. Other popular 3D-QSAR methods are CoMSIA[3] and SOMFA.[4] Both of them are derived from CoMFA. In addition to steric and electrostatic fields, CoMSIA computes hydrophobic, hydrogen bonding donor, and hydrogen bond acceptors fields for each ligand, whereas SOMFA uses intrinsic molecular properties (i.e., molecular shape and electrostatic potentials) to create QSAR models, instead of using probe interaction energies. Such grid-based methods rely on frequently ambiguous procedures of molecular alignment and utilize the partial least-squares (PLS)[5] approach to fit the grid-derived descriptors to the dependent variable (the biological activity, in the majority of cases). Despite the fact that CoMFA, CoMSIA, and SOMFA can employ docking poses of the ligands to construct the molecular alignments, they all are characterized as receptor-independent 3D-QSAR approaches, since they do not explicitly take into consideration the structure of the receptor.

Another important grid-based 3D-QSAR method is the 4D-QSAR[6] method, which incorporates conformational and alignment freedom into the development of structure−activity models by performing conformational ensemble sampling on each data set ligand. The descriptors of the method are Cartesian coordinates of each of the atom-types of the ligands at each point of the 3D-grid. Such atom-types are classified as bulky (steric), polar positive, polar negative, hydrogen bonding donor, hydrogen bonding acceptor, and aromatic atoms. The conformational ensemble of the molecules (the fourth dimension of the method) is obtained from molecular dynamics simulations (MDS). Such an approach can be used in both receptor-independent and -dependent schemes. The difference is that in the first approach the structure of the receptor is not used during the conformational sampling of the ligands, whereas in the second one that structure is explicitly used. However, since intermolecular terms are not part of the 4D-QSAR algorithm, ligand−receptor intermolecular interactions can only be indirectly inferred.

Besides 4D-QSAR, there are only a few receptor-dependent QSAR methods reported, including 6D-QSAR,[7] free-energy force field (FEFF) 3D-QSAR,[8] Raptor,[9] and adaptation of fields for molecular comparison (AFMoC).[10] As in the 4D-QSAR method, in 6D-QSAR the fourth dimension refers to the possibility of representing each molecule by an ensemble of conformations, orientations, protonation states, and tautomers; the fifth dimension refers to the possibility of considering an ensemble of different induced-fit models by adapting the mean van der Waals surface, generated about all ligands defining the training set; and the sixth dimension allows for the evaluation of different salvation models. Ligand−receptor interactions are estimated based on the force field terms used during the simulations. The used “receptor” of the method is actually a hypothetical receptor

* Corresponding author e-mail: o_santosfilho@yahoo.com.

10.1021/ci8001952 CCC: $40.75 © 2008 American Chemical Society
Published on Web 09/24/2008

site, characterized by the 3D-surface that surrounds the ligand molecules at van der Waals distance and which is populated with atomistic properties (hydrophobicity, electrostatic potential, and hydrogen-bonding propensity) mapped onto it. In other words, the topology of that ligand-based surface mimics the 3D shape of the binding site.

Overall, FEFF 3D-QSAR analysis[8] includes all contributions to the free energy of the ligand−receptor binding process in order to construct 3D-QSAR models. Moreover, all members of the binding process (ligand−receptor complexes, the unbound receptor, and unbound ligands) are considered during the analysis. The binding thermodynamic properties are inferred from the terms of the used force field during the molecular dynamics simulations at different temperatures, through a warm up and cooling down cycle. Then, PLS and genetic algorithm are used for sampling the most significant terms (thermodynamic descriptors), and the QSAR models are constructed. This method was satisfactorily used in several structure-based design studies. However, since a large number of molecular dynamics simulations needs to be done for each ligand and the receptor, in both bond and nonbonded states at several temperatures, the overall computational cost is relatively high.

Raptor[9] is an innovative receptor-modeling technology which is based on a novel scoring function including hydrophobicity, hydrogen bonding, entropy, and the quantification of an explicitly simulated induced fit. The philosophy underpinning the new concept is a dual-shell representation of the binding-site surrogate, allowing for the simulation of anisotropic effects. In this context, this software is an extension of the 6D-QSAR approach as it is quite powerful in generating surrogates for predominantly hydrophobic binding sites as well as in handling mixed-charge ligand systems.

AFMoC[10] makes use of available modeling data by tailoring DrugScore knowledge-based potentials specifically toward a given protein using inhibitor potency data. AFMoC was found to provide superior performance, based both on cross-validation runs as well as for inhibitors not considered in the training set. In particular, AFMoC's ability to gradually transform between generally applicable unadapted interaction fields to case specifically adapted ones proved to be of major importance.

In this study, we propose a RD 3D-QSAR approach which includes molecular docking simulations, cluster analysis, and 'induced' descriptors. Structurally diverse ligand data sets from the literature and from known databases were used to define its applicability domain and to evaluate its advantages and limitations. When needed, molecular fingerprints based cluster analysis was used for resampling the data sets. Extensive docking calculations were carried out in order to "capture" optimum intermolecular interactions of the ligands with their respective receptor binding sites. PLS and genetic algorithm were used for constructing the QSAR models. In our approach, no ambiguous alignment step is required, since the conformations of the ligands are regarding their respective docked poses. Moreover, the corresponding computational cost is relatively modest, as compared to those found in other RD QSAR methodologies. The constructed QSAR models showed statistical and predictive strength and were used for developing drug design hypotheses.

METHODS

Data Sets and Molecular Targets. In this study, eight data sets were used for constructing receptor-dependent 3D-QSAR models. Six of them were sampled from the Binding Database (BindingDB)[11] (120 anthrax lethal factor ligands, 1318 HIV-1 reverse transcriptase ligands, 170 neuraminidase A ligands, 181 thrombin ligands, and 197 trypsin ligands); two data sets were used in our previous works[12−14] (84 ligands for human sex hormone-binding globulin (SHBG) and 53 ligands for human corticosteroid-binding globulin (CBG)); and 756 *Pneumocystis carinii* dihydrofolate reductase (*pc*DHFR) was assembled by Sutherland and colleagues.[15]

The reasons for selecting the data sets in the current study are as follows: (a) they are regarded as distinct classes of receptors; (b) the number of ligands on each data set is equal to at least 50 molecules; (c) the range of biological (activity) data, in logarithmic scale, is equal to at least 4; and (d) they are associated with different diseases. From the data sets, just the SHBG and CBG ones are not constituted of structurally diverse molecules.

The bioactivity data (K_i, K_d, or IC_{50}) were expressed as pK_i, pK_d, or pIC_{50}, where p is the cologarithm (−log) of the bioactivity, Each data set was split into two subsets: a training set, used for training the QSAR models, and a test set (equal to 15% of each corresponding data set), used for validating the calculated models.

All molecular targets (hereafter called "receptors"), except the CBG one, were obtained from the Protein Data Bank.[16] The CBG receptor[13] was constructed by using homology modeling.

Proteins are stored in BindingDB, based on their importance as drug targets as well as on the availability of suitable data.[11] Moreover, a specific data set of ligands can show bioactivity data for more than one specific protein. In other words, mutant proteins can be found for the same data set. Consequently, a careful analysis was done, in order to sample ligands for unique proteins.

Molecular Docking. The Maestro suite[17] was used to prepare the receptor structures for docking calculations. All water and ion molecules were removed from the corresponding PDB structures, and hydrogen atoms were added. In order to prevent steric hindrance on the system, it was found to be necessary to minimize the hydrogen atoms, keeping the remaining atoms (heavy atoms) fixed during this process.

The docking calculations were conducted using Glide 4.5 parallel suite,[18] with default settings. All computations have been carried out on 10 dual-core LINUX/Centos4.3 IBM workstation equipped with Intel(R) Pentium(R) D CPU 3.00 GHz processors and 2GB RAM of memory.

'Inductive' Descriptors and Its Usage in Evaluating Receptor−Ligands Interactions. In previous works, we have presented a set of molecular descriptors, called 'inductive' descriptors. Such descriptors are derived from linear free energy relationships (LFER)-based equations for inductive and steric substituent parameters. The theory and mathematical formalism of those 'inductive' descriptors was reviewed recently[19] and is summarized below.

Table 1. 3D-QSAR Models

data set	3D-QSAR model	eq no.
anthrax lethal factor	pIC$_{50}$(M) = −0.45 + 6.70 RsHIS686ND1 + 0.91 RsHIS690CE1 − 5.65 RsLEU658CB (6)+ 3.08 RsLEU658CD2 − 1.39 RsTYR659O − 2.69 SigmaGLU687CA + 1.22 SigmaHIS654N + 2.97 SigmaLEU330N + 2.23 SigmaSER655OG + 3.00 SigmaTHR731OG1 + 4.18 SigmaTYR659C − 4.08 SigmaVAL653N (*N* = 62,r^2 = 0.81, q^2 = 0.70)	(6)
HIV-1 reverse transcriptase	pIC$_{50}$(M) = 6.83 +3.85 RsLYS102O + 2.04 RsLYS103O − 8.24 RsTYR181CB (7)+ 3.49 SigmaHIS235N + 4.73 SigmaLYS103N − 3.85 SigmaTYR181OH − 5.01 SigmaTYR318OH (*N* = 33, r^2 = 0.82, q^2 = 0.67)	(7)
neuraminidase A	pIC50(M) = −0.86 - 3.81 RsARG118CZ + 2.91 RsARG224CD + 9.37 RsARG371NH2 (8)+5.56 RsTYR406CD2 − 4.58 SigmaARG152CD − 1.49 SigmaARG152NH2 +5.78 SigmaARG156O − 6.30 SigmaARG371CZ + 3.09 SigmaILE222CA +1.93 SigmaSER246C (*N* = 53, r^2 = 0.83, q^2 = 0.73)	(8)
thrombin	pIC50(M) = 0.91 + 6.15 RsGLU217N + 2.98 RsGLY226CA − 1.19 RsHIS57ND1 (9) +6.32 SigmaASP189C − 3.34 SigmaCYS220N + 1.46 SigmaGLU192CB + 5.80 SigmaGLY196N + 9.31 SigmaGLY219O − 4.67 SigmaHIS57CB +6.14 SigmaLEU99CA − 13.72 SigmaPHE227CB + 8.70 SigmaSER195CB −10.09 SigmaSER195N − 0.42 SigmaSER214CA + 1.04 SigmaTRP60CZ3 − 2.20 SigmaTYR60OH (*N* = 110, r^2 = 0.84, q^2 = 0.76)	(9)
trypsin	pKi(M) = −4.92 − 4.19 Rs _LEU99CD2 + 3.02 RsSER217OG + 14.46 RsTYR228OH (10) +11.81 SigmaCYS191N + 4.58 SigmaGLN192NE2 + 4.11 SigmaGLY216O +7.61 SigmaLEU99CB − 21.95 SigmaVAL213C (*N* = 42, r^2 = 0.88, q^2 = 0.81)	(10)
SHBG	p*K*$_{d}$(M) = 1.12 + 6.03 RsLEU171CD2 − 17.72 RsLYS106CA + 22.4 RsMET107CB (11)− 4.11 RsPHE56CE1 + 9.40 RsTRP66CH2 + 8.54 RsTRP84CZ3 + 6.41 RsVAL112CG1 −17.60 RsVAL127CA − 6.43 SigmaASP59OD1 − 13.11 SigmaSER41O + 2.92 SigmaSER42CA +23.41 SigmaTHR40OG1 −2.22 SigmaVAL105CB − 3.79 SigmaVAL127C (*N* = 72, r^2 = 0.84, q^2 = 0.72)	(11)
CBG	p*K*$_{a}$(M) = −7.54 − 5.61 RsASP358OD2 + 4.21 RsILE255CG1 − 4.83 RsILE255O (12) +6.22 RsLYS359CG + 5.68 RsTRP362CB − 1.58 SigmaARG252CD − 2.59 SigmaGLN224CD −3.26 SigmaTRP362CE3 + 4.32 SigmaTRP362N (*N* = 45,r^2 = 0.86, q^2 = 0.80)	(12)
P. carinii DHFR	pIC$_{50}$(M) = 3.43 − 3.85 RsARG75NH2 − 6.91 RsSER64OG + 10.14 RsTHR61N (13) + 4.84 SigmaGLU32OE1 − 3.10 SigmaILE33CD1 + 5.61 SigmaILE65CG2 −17.87 SigmaLEU72N + 17.10 SigmaLYS37N + 6.29 SigmaPRO66CD −16.57 SigmaTRP27CA − 5.03 SigmaTRP27NE1 + 4.33 SigmaTYR129OH (*N* = 60, r^2 = 0.82, q^2 = 0.72)	(13)

The basic equations of the 'inductive' descriptors are

$$Rs_{G\to j}=\alpha\sum_{i\subset G,i\neq j}^{n}\frac{R_i^2}{r_{i-j}^2} \tag{1}$$

and

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

where *Rs* is the steric influence of a group of *n* atoms constituting a substituent *G* onto a single atom *j* (reaction center), and σ^* is the inductive effect of *G* onto reaction center *j*. *R* corresponds to the covalent atomic radii of an *i-th* atom of group *G*, *r* is the distance between atoms *i* and *j*, and χ^0 is the 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.[19,20]

'Inductive' descriptors have been successfully applied in constructing QSAR models[21−26] and were used in this work for evaluating protein−ligand interactions as well as deriving the corresponding receptor-dependent molecular descriptors. In this context, eqs 1 and 2 can be rearranged in order to represent ligand−receptor interactions as follows

$$Rs_{L\to p}=\sum_{i\subset L}^{N}\frac{R_i^2}{r_{i-p}^2} \tag{3}$$

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

where the parameters $Rs_{L\to p}$ and $\sigma^*_{L\to p}$ describe, respectively, the overall steric and inductive interactions occurring between the entire docked ligand and a receptor's atom considered as a reaction center.

Table 2. Relative Importance (Significance) of the Descriptors in the Optimum QSAR Models

data set	relative significance of the descriptors
anthrax lethal factor	1.00: SigmaTYR659C 0.91: RsLEU658CB 0.89: RsHIS686ND1 0.53: SigmaVAL653N 0.49: RsLEU658CD2 0.46: SigmaGLU687CA 0.35: SigmaSER655OG 0.34: RsTYR659O 0.33: SigmaLEU330N 0.31: SigmaTHR731OG1 0.17: RsHIS690CE1 0.19: SigmaHIS654N
HIV-1 reverse transcriptase	1.00: SigmaTYR318OH and RsTYR181CB 0.81: RsLYS102O 0.76: SigmaLYS103N 0.75: SigmaHIS235N 0.56: SigmaTYR181OH 0.47: RsLYS103
neuraminidase A	1.00: RsARG371NH2 0.82: SigmaARG371CZ 0.61: RsTYR406CD2 0.50: SigmaARG156O 0.45: SigmaARG152CD 0.41: SigmaILE222CA 0.31: RsARG118CZ 0.25: RsARG224CD 0.24: SigmaSER246C 0.18: SigmaARG152NH2
thrombin	1.00: SigmaSER195N 0.95: SigmaSER195CB 0.82: SigmaGLY219O 0.66: SigmaPHE227CB 0.51: RsGLU217N 0.50: SigmaGLY196N 0.43: SigmaASP189C 0.42: SigmaLEU99CA 0.33: SigmaHIS57CB 0.26: SigmaCYS220N 0.25: SigmaTYR60OH 0.20: RsGLY226CA 0.16: RsHIS57ND1 0.13: SigmaGLU192CB 0.09: SigmaTRP60CZ3 0.03: SigmaSER214CA
trypsin	1.00: SigmaVAL213C 0.83: SigmaCYS191N 0.71: RsTYR228OH 0.42: SigmaGLN192NE2 and SigmaLEU99CB 0.38: RsLEU99CD2 0.31: RsSER217OG and SigmaGLY216O
SHBG	1.00: RsMET107CB 0.94: RsVAL127CA 0.93: SigmaTHR40OG1 0.92: RsLYS106CA 0.65: SigmaSER41O 0.56: RsTRP66CH2 0.47: RsTRP84CZ3 0.41: RsVAL112CG1 0.36: SigmaASP59OD1 0.25: RsPHE56CE1 0.24: SigmaVAL127C 0.23: SigmaSER42CA and 0.23 RsLEU171CD2 0.13: SigmaVAL105CB
CBG	1.00: RsTRP362CB 0.76: SigmaTRP362N 0.71: RsILE255O 0.70: RsLYS359CG 0.65: SigmaTRP362CE3 0.62: RsILE255CG1 0.58: SigmaGLN224CD 0.57: RsASP358OD2 0.31: SigmaARG252CD
P. carinii DHFR	1.00 SigmaLEU72N 0.95: SigmaLYS37N 0.54: RsTHR61N 0.48: SigmaTRP27CA 0.45: SigmaPRO66CD0.39: RsSER64OG and SigmaGLU32OE1 0.38: SigmaILE65CG2 0.32: SigmaTYR129OH 0.23: RsARG75NH2 0.22: SigmaILE33CD1 0.18: SigmaTRP27NE1

Fingerprint Cluster Analysis. Molecular fingerprints (also known as structural keys or molecular signatures) represent a set of features derived from the structure of a molecule. It is a function of the topology of the molecule or of its conformation. Several schemes do exist for representing molecular fingerprints, and the fundamental idea is to encapsulate relevant properties useful for 'defining" the structure of the molecules. In this work several schemes were tested, and for the data sets used, the 2-point pharmacophore based fingerprint approach (FP: TGD) was found to be appropriate. With this approach, each atom is given a specific type (*hydrogen-donor*, *hydrogen-acceptor*, *polar*, *anion*, *cation*, and *hydrophobic*). Then, all pairs of atoms are coded as features, and the topology of the molecule (the graph distance of each atom pair) is taken into consideration for the definition of the molecular fingerprint of the system.

The used clustering analysis approach is based on the Jarvis-Patrick algorithm[27] and is summarized below:

1. *Calculation of the molecular fingerprints*.
2. *Calculation of the similarity matrix*. A matrix A is created, where $A[i,j]$ contains the similarity metric between fingerprints i and j.
3. *Derivation of the similarity thresholding*. From the similarity matrix A, a binary matrix B is created such that $B[i,j]$ has the value 1 if $A[i,j] \geq S$, or 0 otherwise. Here, S is the *similarity threshold* used to determine if two fingerprints are similar.
4. *Clustering analysis*. The rows of the matrix B are treated as the fingerprints. Two molecules are "put" in the same cluster if the Tanimoto coefficient of their corresponding rows in B is greater than or equal to T, the *overlap threshold*. It is important to mention that the Tanimoto coefficient is the number of features in common divided by the union of the features and always lies between 0 and 1.

According to that clustering scheme, two molecules are considered to be in the same cluster if the lists of molecules to which they are similar overlap to a reasonable degree. In other words, two molecules belong in the same cluster if they are similar to the same set of molecules.

Optimization of the QSAR Models with Genetic Algorithm. The optimization of the 3D-'inductive' QSAR models was carried out with a genetic algorithm, as implemented on the *QuaSAR-Evolution* script,[28] which is based on Hopfinger's genetic function approximation (GFA).[29] Such a script was loaded on the MOE package.

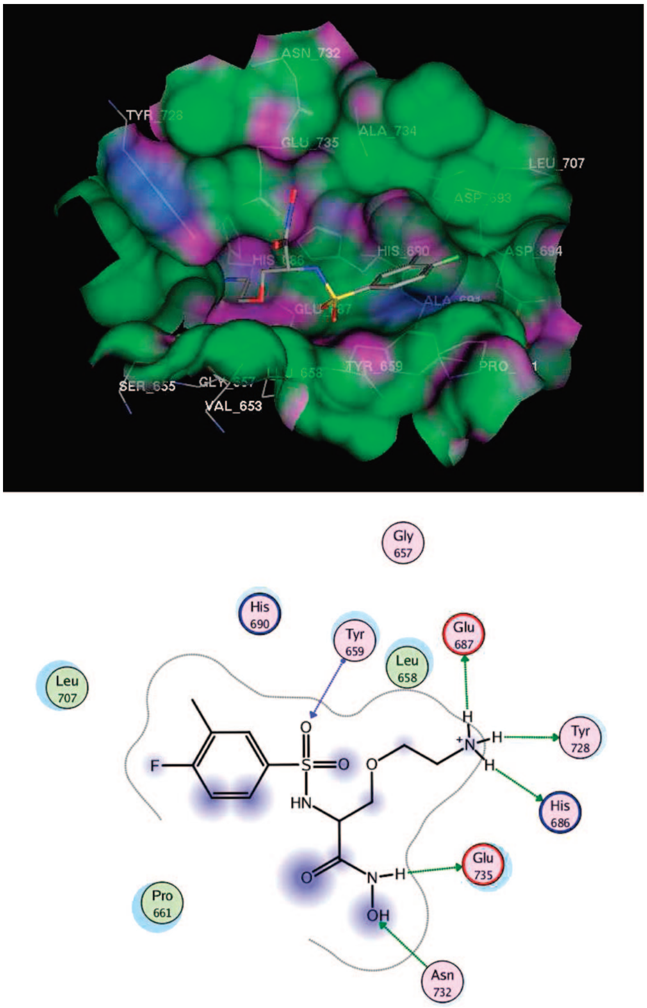

Figure 1. 3D- and 2D-representations of the interaction of the most active (training) ligand of the anthrax data set with its receptor.

As all evolutionary algorithms, GFA[29] uses biology terms such as *inheritance*, *mutation*, *selection*, and *crossover* (recombination) to identify optimal mathematical solutions.[30] The method 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 a ‘fitness function’ to iteratively evolve them to an optimal solution (i.e., to find the most appropriate set of descriptors). Such a function is the Friedman's ‘lack-of-fit’ (*LOF*) function

$$LOF = \frac{LSE}{\left(1 - \frac{c + dp}{n}\right)^2} \tag{5}$$

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.[29]

RESULTS

Data Sets and Molecular Targets. For the SHBG and DHFR data sets, the crystallographic structures of their respective ligand−receptor complexes were available from PDB: 1D2S[33] and 1E26,[34] respectively. For the CBG data set, a homology structure of the receptor (modeled in house) was used.[13]

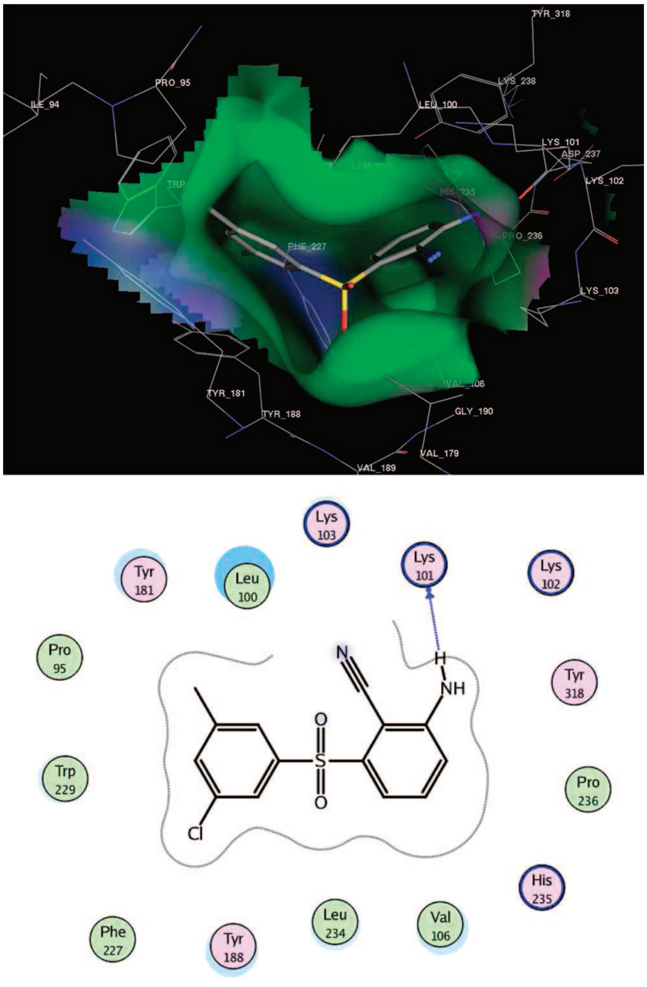

Figure 2. 3D- and 2D-representations of the interaction of the most active (training) ligand of the HIV-1 reverse transcriptase data set with its receptor.

In BindingDB, the entries of each data set are constituted by ligands, amino acid sequences of the receptors, and the corresponding bioactivity data. In that database, bioactivity data are determined for homologue enzymes, instead of for unique ones. In this context, in order to use BindingDB in QSAR analyses, a preliminary analysis on each data set was done to ensure that the bioactivity data of the ligands are regarding the unique receptor. That procedure results in a resampling of the ligands and can be considered as a first screening on that database. The following scheme was used. First, the most active ligands of each data set as well as the respective amino acid sequence of its receptor were identified. Then, a search for PDB entries with both similar amino acid sequences and ligands was carried out. The quantitative criteria used for that search were as follows: (a) a Blast E-value cutoff equal to 1.0 × 10^{-25} and (b) a ligand similarity of at least 90%. Thus, each data set from BindingDB was resampled, and the following PDB structures were selected for the subsequent docking calculations: 1PWP,[35] for the anthrax lethal factor data set; 1C0T,[36] for the HIV-1 reverse transcriptase data set; 2HU0,[37] for the neuraminidase A data set; 1ABJ,[38] for the thrombin data set; and 1F0U,[39] for the trypsin data set.

Molecular Docking and Cluster Analysis. Before running the docking calculations, required preliminary computations were carried out: missing hydrogen atoms were added to the receptor structures; the charges states near their binding

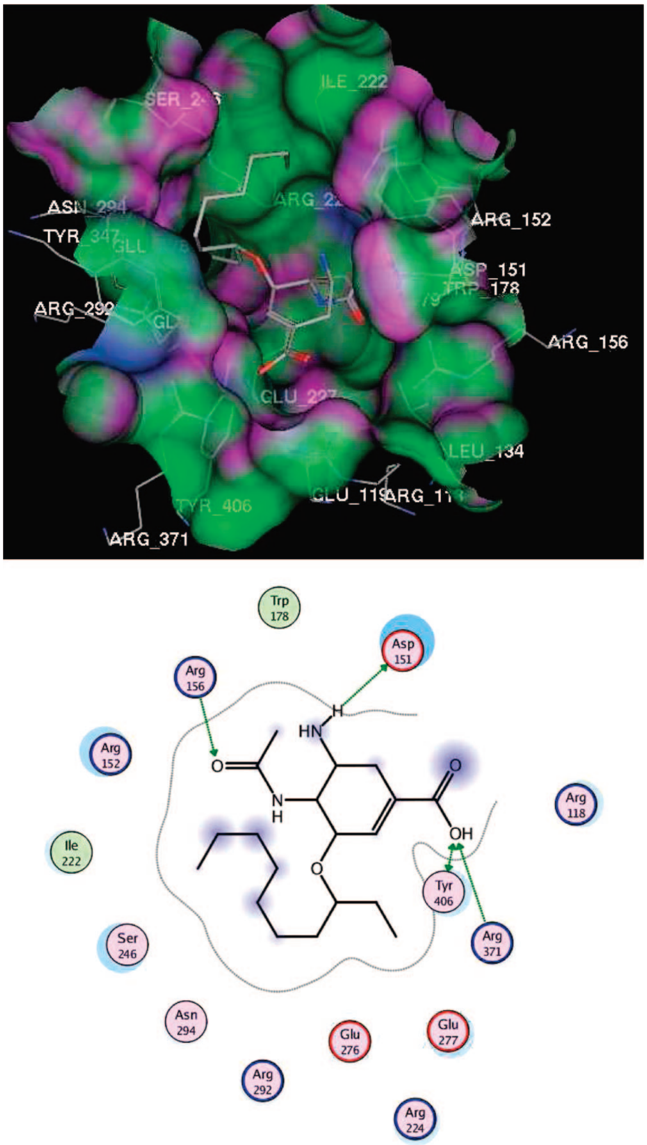

Figure 3. 3D- and 2D-representations of the interaction of the most active (training) ligand of the HIV-1 neuraminidase data set with its receptor.

sites were corrected, hybridizations were fixed, and major steric clashes were removed. Both Maestro 8.5[17] and MacroModel 9.5[40] programs were used. Subsequently, each ligand was docked to their respective receptors with the Glide program.[18] A flexible-ligand based docking scheme with an extra precision scheme was used.

It was noticed that not all ligands docked to their respective receptors. Thus, the docking calculations "filtered" the initial pool of ligands, working as a second data set screening. The original data sets were reduced as follows: (a) 111 anthrax lethal factor ligands; (b) 194 factor XA ligands; (c) 950 HIV-1 reverse transcriptase ligands; (d) 170 neuraminidase A ligands; (e) 157 thrombin ligands; (f) 191 trypsin ligands; (g) 84 SHBG ligands; (h) 53 CBG ligands; and (i) 600 *P. carinii* DHFR ligands.

For the construction of the QSAR models just the best (energy-scored) docking pose of each ligand was selected, and each data set was organized into two subsets (training and test sets). It was noticed that for the case of the larger data sets (those obtained from BindingDB and the Sutherland's DHFR data sets) the respective QSAR models showed very low predictivity quality, with cross-correlation coefficients (q^2) lower than 0.40. Moreover, a large number of outlier ligands were found, which indicates the presence of highly diverse data sets. Thus, a molecular fingerprint cluster analysis was carried out in order to resample those data sets. A large number of ligands were found forming singleton clusters (clusters with one single ligand), which confirms the structural diversity on the data sets. In this context, new QSAR models were constructed by using the largest ligand cluster of each data set. This can be considered the third and final screening procedure. The resulting data sets are constituted by the following: (a) 73 (62 training set and 11 test set) anthrax lethal factor ligands; (b) 39 (33 training set and 6 test set) HIV-1 reverse transcriptase ligands; (c) 62 (53 training set and 9 test set) neuraminidase A ligands; (d) 129 (110 training set and 19 test set) thrombin ligands; (e) 49 (42 training set and 7 test set) trypsin ligands; (f) 84 (72 training set and 12 test set) SHBG ligands; (g) 53 (45 training set and 8 test set) CBG ligands; and (h) 70 (60 training set and 10 test set) *P. carinii* DHFR ligands.

The QSAR Models. In this study, 'inductive' descriptors (eqs 3 and 4) were calculated for different orientations of each ligand (molecular poses generated from docking calculations) into the active site of their respective receptors.

The quality of a RD QSAR is a function of its ability in capturing as much information regarding (ligand−receptor) intermolecular interactions as possible. Moreover, since the involved atoms are those from both ligands and their respective active sites, the remaining atoms can be excluded from the QSAR analysis. Such an approximation not only represents a considerable reduction of computational costs but also eliminates unnecessary "noise" during the analysis of the models.

Each receptor was protonated with (the default MOE[41]) a pH value of 7.4. After that, the 'optimum' size of each receptor was determined by using the MOE/Site Finder methodology.[41,42] By using such an approach several probable binding sites were determined. However, a comparison with the crystallographic structure of each receptor (from PDB) was done in order to identify the correct active site.

The 'inductive' descriptors were calculated for the highest (Glide) scored docking pose of each ligand. Moreover, before constructing the QSAR models, all descriptors were normalized. The best QSAR models, optimized with *QuaSAR-Evolution* (MOE) script, are shown in Table 1. The MOE package calculates the relative importance (significance) of the descriptors in the optimum QSAR models (Table 2).

DISCUSSION

Previous works have shown the importance of including the structure of receptors for constructing 3D-QSAR models for structurally diverse data sets.[43,44] This study showed that not always the usage of the receptor structure is enough for the development of significant models. In this study all attempts of using molecular entries of the largest data sets for constructing the QSAR models failed. Not only a large number of outlier molecules was found but also the statistical significance of the models was very low ($q^2 < 0.40$). Six of the eight data sets used in this study (those assembled on BindingDB and the *P. carinii* DHFR assembled from the literature by Sutherland and colleagues) are in this category.

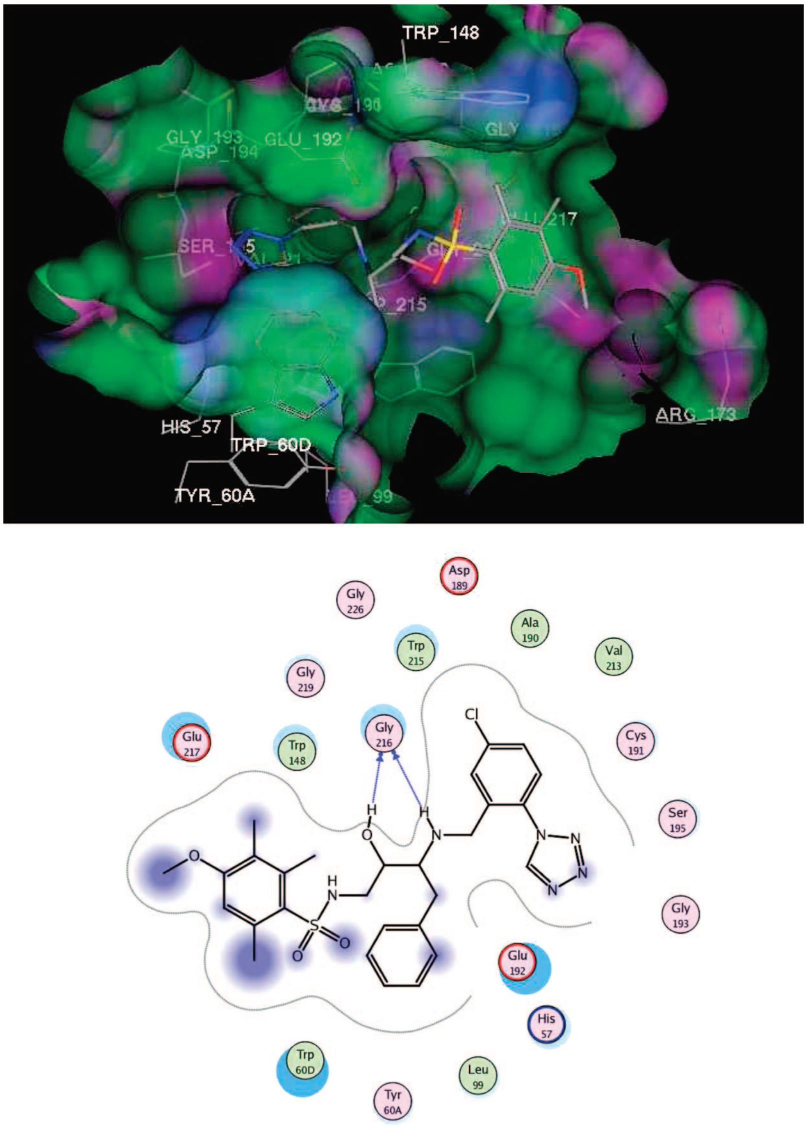

Figure 4. 3D- and 2D-representations of the interaction of the most active (training) ligand of the thrombin data set with its receptor.

The solution for the problem was to carry out cluster analysis on the large data sets before constructing the QSAR models. By doing that, we noticed a remarkable reduction in the number of molecules. The reduction for the anthrax lethal factor, HIV-1 reverse transcriptase, neuraminidase A, thrombin, trypsin, and *P. carinii* DHFR data sets was, respectively, 34%, 96%, 63%, 18%, 74%, and 88%. In this context, cluster analysis worked as an additional and significant, screening strategy.

There is a debate in the QSAR community regarding how to sample test set molecules from a molecular population. The goal is to construct a test set that represents the molecular features of the training set, so that the former could be used as an external validation tool for the respective QSAR model. The adopted criterion in this study was to cluster each data set, once again, followed by the selection of about 15% of ligands to be used as test sets. This way not only ligands ranging from active to inactive but also those with similar fingerprints features were selected. Such a method worked properly, due to the significant correlation found between the bioactive data of the data sets and their respective QSAR predictions (correlation plots are shown in the Supporting Information).

Regarding the construction of the models, trial QSAR models were optimized by using both genetic algorithm (GA) and PLS regression. No outlier ligand was found. The QSAR models are shown in Table 1, and Figures 1−8 show 3D- and 2D-representations of the interaction of the most active (training) ligand of each data set with their respective receptors.

By analyzing the QSAR models (eqs 6−13) and Figures 1−8, it is possible to investigate the intermolecular interactions between the ligands and their respective receptors. That can be useful to the development of drug design hypothesis. During this study, crystallographic structures of the active sites of the receptors were used as the main reference.

3D-QSAR and Docking Models for the Anthrax Lethal Factor Data Set. Anthrax lethal factor is a highly specific protease that cleaves members of the mitogen-activated protein kinase (MAPKK) family near to their amino termini, leading to the inhibition of one or more signaling pathways.[36] Equation 6 shows the preponderance of polar ‘inductive’ descriptors (Sigma) descriptors. Figure 1 shows that polar residues, Tyr-659, Tyr-728, His-686, Asn-732, Glu-687, and Glu-735, are involved in hydrogen bond interactions with the ligand. The negative contribution of the steric ‘inductive’ descriptor (negative coefficient value: −1.39) on the oxygen

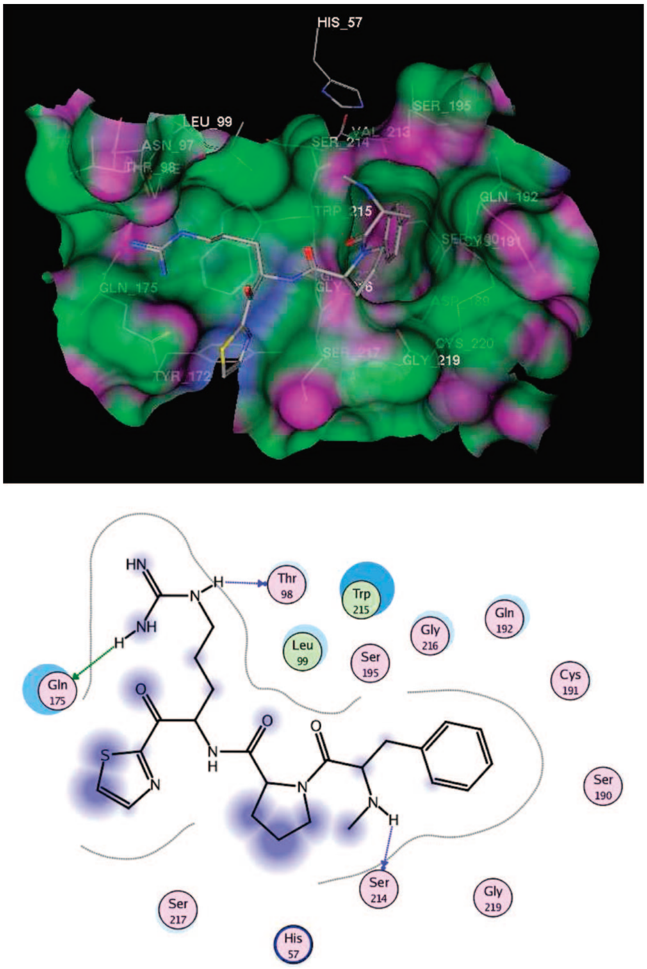

Figure 5. 3D- and 2D-representations of the interaction of the most active (training) ligand of the trypsin data set with its receptor.

of Tyr-659 indicates that the design of lead molecules with bulky and/or nonpolar substituent groups close to that oxygen would "disturb" the formation of hydrogen bond, important for the stabilization of the interaction of the ligand into the active site. Moreover, the highly positive contributing (+4.18) of the polar 'inductive' descriptor of the same residue seems to confirm that hypothesis as well. The relatively high significance of the steric contribution of the nonpolar residue Leu-658 (91% and 49% for RsLEU658CB and RsLEU658CD2, respectively) seems to be contradictory for an acidity active site. However, a closer look at the QSAR model (eq 6) shows that, actually, such a residue is negatively contributing to the bioactive response (a net coefficient value equal to −2.57). Probably, the relative proximity of Leu-658 to the acid Glu-687 is an indication of the design of lead molecules with a nonpolar group in that region which would disturb the formation of a favorable acid interaction with Glu-687. According to Figure 1, His-686 forms a hydrogen bond with an exposed (to the solvent) proton, which seemed to be "acquired" by the ligand from the aqueous environment. Equation 6 shows a steric 'inductive' descriptor on the nitrogen of the His, instead. In order to investigate such an unusual term, an analysis of the interaction of the second most active ligand (pIC_{50} value difference equal to 0.01) was carried out. It was noticed that the second most active ligand is not directly interacting with His-686. Moreover, such a residue is not directly interacting with the ligand found on the crystallographic structure of the receptor as well. Consequently, even not being essential to the interaction of ligands to the anthrax receptor, due to its proximity to the active site, His-686 is able to participate in eventual interactions with the ligands. The steric 'inductive' descriptor associated with His-686 could be contributing to the orientation of the ligands toward the polar groove of the receptor binding pocket. If such an idea is true, then this feature should be considered during the drug design process.

3D-QSAR and Docking Models for the HIV-1 Reverse Transcriptase Data Set. HIV-1 reverse transcriptase[29] is used by the HIV-1 virus to transcribe its single-stranded RNA genome into single-stranded DNA and subsequently construct a complementary strand of DNA, providing a DNA double helix capable of integration into host cell chromosomes. Figure 2 shows the interaction of the most active training set ligand and the active site of that receptor. The active site is predominantly hydrophobic, being constructed primarily from Leu, Val, Trp, and Tyr. Thus, for this specific receptor, the main drug design strategy is to focus on nonpolar substituents. Equation 7 shows that 'inductive' descriptors regarding Lys-102, and Lys-103 are positively contributing to the model bioactivity. The net contribution of Lys-102 is +3.85 and for Lys-103 is +6.77. Moreover, the relative importance of those residues during the optimization of the model is also relatively high: 81% and 76%, respectively. Whereas the descriptors regarding both Tyr-181 and Tyr-318 are negatively contributing to the bioactivity (net contribution of −12.09 and −5.01), with relative importance during the optimization of 100%. The contribution of those four descriptors is in accordance to the hydrophobic nature of the active site. The positive contribution of the 'inductive' descriptor regarding His-235 as well as the shown hydrogen bond of the ligand with Lys-101 seems to be an indication of the possibility of eventual hydrogen bonds, depending on the structure. This is a secondary effect but should not be totally discarded.

3D-QSAR and Docking Models for the Neuraminidase A Data Set. Neuraminidase A[37] is a protein that is necessary for virus proliferation, which makes it an ideal target for the development of anti-influenza drugs. The optimization of the 'induced' descriptors for the neuraminidase A data set resulted in a QSAR model constituted with basic residues (Arg-118, Arg-152, Arg-156, Arg-224, Arg-371, and Tyr-406), polar and small amino acid (Ser-246), and hydrophobic residues (Ile-222 and Tyr-406). The descriptors with higher relative importance are as follows: RsARG371NH2(100%),SigmaARG371CZ(82%),RsTYR406CD2 (61%), and SigmaARG156O (50%). According to eq 8, the net contributions of Arg-371, Tyr-406, and Arg-156 are, respectively, +3.07, +5.56, and +5.78. Three terms are negatively contributing to the bioactivity of the model: RsARG118CZ, SigmaARG152CD, and SigmaARG152NH2. Figure 3 shows that Arg-156, Arg-371, and Tyr-406 form hydrogen bonds with the ligand. Moreover, Asp-151, a residue whose descriptor was not "captured" during the optimization of the model, can also form a hydrogen bond. Consequently, those hydrogen bonds must be considered for the design of potential active lead molecules. Due to the steric nature (Rs) of the 'induced' descriptors found on Arg-371 and Tyr-406, it is necessary to be careful when proposing substituent groups to the ligands capable of interacting with those residues, and a size-charge balance needs to be

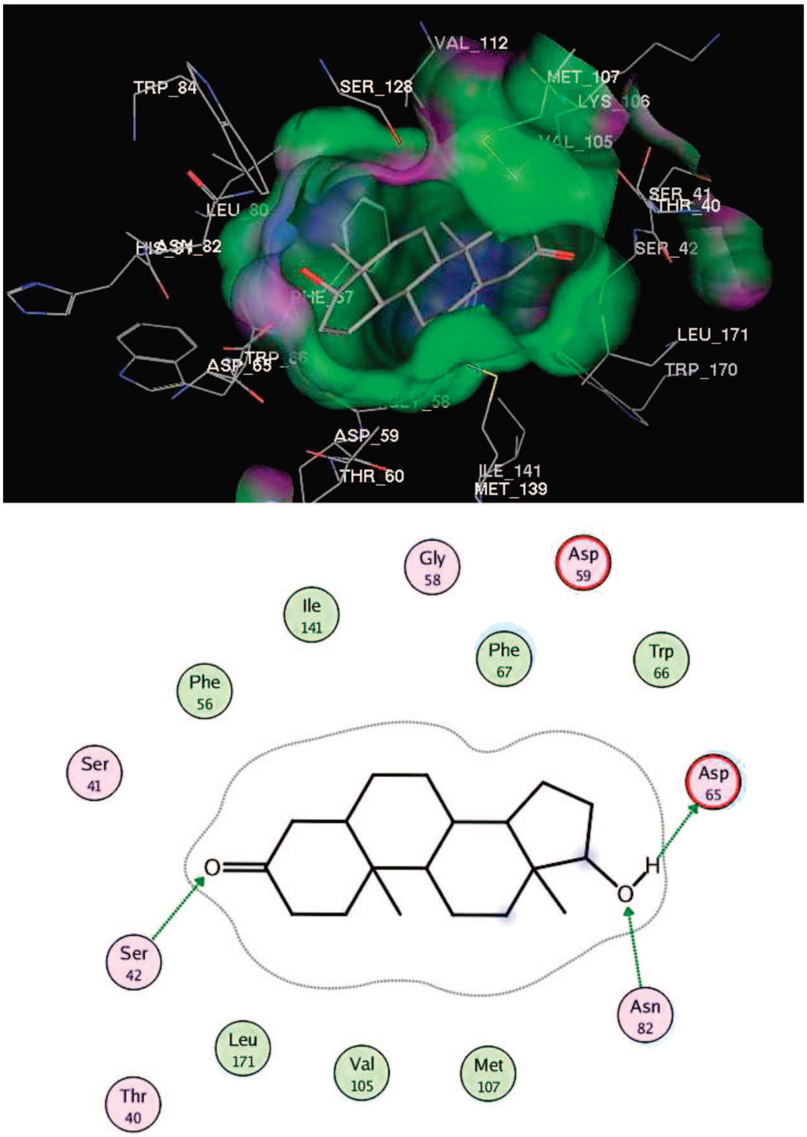

Figure 6. 3D- and 2D-representations of the interaction of the most active (training) ligand of the SHBG data set with its receptor.

considered. In other words, it seems that small groups capable of forming hydrogen bonds with Arg-156, Arg-371, and Tyr-406 would be more appropriate than larger ones.

3D-QSAR and Docking Models for the Thrombin Data Set. Thrombin,[38] a coagulation protein that has many effects in the coagulation cascade, converts fibrinogen into soluble strands of fibrin as well as catalyzing many other coagulation-rated reactions. Its active site is mainly constituted by polar residues, and it is remarkable in the presence of the catalytic triad His-57, Asp-102, and Ser-195. The QSAR model (eq 9) shows descriptors regarding two of those residues (His-57 and Ser-195). Moreover, three of the sixteen ‘induced’ descriptors are related to polar interactions (Sigma). Ser-195 show descriptors with relatively high importance (SigmaSER195N [100%] and SigmaSER195CB [95%]), whereas His-57 has a relatively low contribution (33% for the descriptor SigmaHIS57CB and 16% for RsHIS57ND1). On the model, the net contribution is lightly negative (−1.39) for the Ser-195 descriptors and negative (−5.86) for His-57 descriptors. Probably, those “negative” contributions are regarding the conformations adopted by the ligands (from the docking calculations). As Figure 4 shows, no stabilizing hydrogen bond was formed with those two residues. Another relatively important descriptor is SigmaGly219O (82%). In the crystallographic structure of the receptor, Gly-219 forms a hydrogen bond with the ligand. Such an interaction was not formed during the docking calculations. However, due to the high value of the coefficient of that descriptor on the model (+9.16), the possibility of changes on the ligands, so that it makes them able to form hydrogen bonds to that residue, seems to be relevant. Another residue that was “missed” during the model optimization, but that forms hydrogen bonds in both crystallographic and docked models, is Gly-216 and should also be considered. SigmaPHE227CB, which 66% of relative importance among the descriptors, has a very low coefficient on the model (−13.72). This seems to be reasonable since this is an indication of how unfavorable a polar interaction would be with such a nonpolar residue.

3D-QSAR and Docking Models for the Trypsin Data Set. Trypsin,[39] found in the digestive system, is a serine protease that hydrolyses proteins into smaller peptides or amino acids. The descriptors regarding the catalytic triad His-57, Asp-102, and Ser-195 is not present on the optimized QSAR model. However, according to Figure 5, His-57 and Ser-195 are relatively close to the ligands and should not be disregarded in full for the design of potential new lead

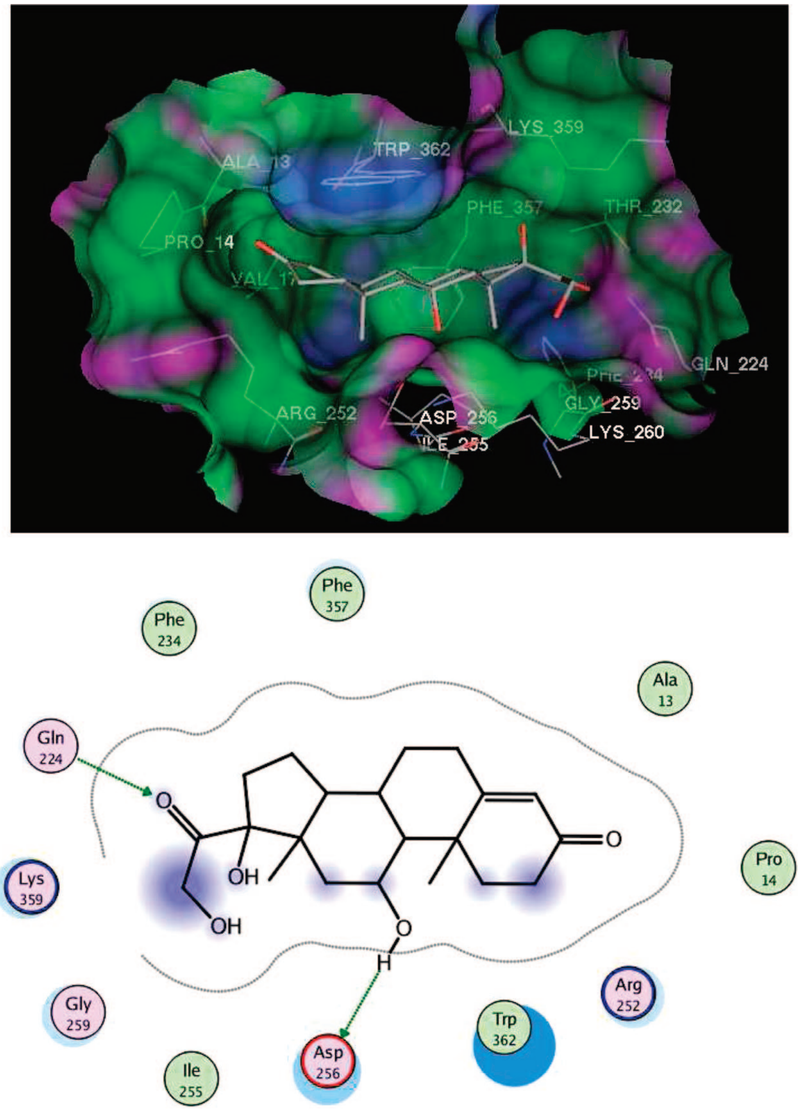

Figure 7. 3D- and 2D-representations of the interaction of the most active (training) ligand of the CBG data set with its receptor.

molecules. SigmaVAL213C, SigmaCYS191N, and RsTyr228OH descriptors have relatively high importance during the optimization of the model: 100%, 83%, and 71%, relatively. SigmaVAL213C descriptor has an extremely low regression coefficient on the model (−21.95) and indicates that a polar interaction with that nonpolar residue is not favorable (eq 10), whereas SigmaCYS191N and RsTyr228OH descriptors have highly positive regression coefficients (+11.81 and +14.46, respectively). That indicates that the presence of polar groups close to those residues would be needed. However, the steric 'induced' nature (Rs) of RsTyr228OH can be a hint regarding precaution regarding larger groups close to Tyr-228. For the most active ligand of this data set, the docking calculations showed the formation of two hydrogen bonds: with Thr-98 and Gln-175. Descriptors for those two residues are not present on the model, but some attention would be taken for potential hydrogen bonds to them.

3D-QSAR and Docking Models for the SHBG Data Set. Human sex hormone-binding globulin (SHBG)[33] is a glycoprotein, which transports sex steroids in blood and regulates their access to target tissues. Those steroids play key roles in the regulation of fertility, reproduction, and sexual behavior. The steroid binding pocket is predominantly hydrophobic. The main binding contributions are from Phe-67, Met-107, and Met-139. The QSAR model (eq 11) shows a steric 'induced' descriptor regarding Met-107 (RsMET107CB) among those with the highest regression coefficients (+22.4). Moreover it is highly relevant during the optimization of the model (100%). As shown in Figure 6 Met-107 is in relative proximity to the steroid, which confirms its ability to form hydrophobic interactions. Other residues whose descriptors are also highly contributing to the optimization of the model are as follows: Val-127 (94% on RsVAL127CA), Thr-40 (93% on SigmaTHR40OG1), Lys-106 (92% on RsLYS106CA), Ser-41 (65% on SigmaSER41O), and Thr-66 (56% on RsTRP66CH2). The regression coefficient of RsVAL127CA, RsLYS106CA, and SigmaSER41O (−17.60, −17.72, and −13.11, respectively) indicates that groups capable of interacting with those residues should be avoided. Whereas SigmaTHR40OG1 and RsTRP66CH2, both with high regression coefficients (+23.41 and +9.40, respectively), indicate that a polar group close to the former residue as well as a nonpolar group capable of interacting with the second residue should be considered. Figure 6 shows the characteristic hydrogen bonds between Ser-42, Asp-65, and Asn-82 of SHBG and the steroid. However, on the QSAR model (eq 11) just the descriptor regarding the Ser-42 is present (+2.92 SigmaSER42CA). Probably, descriptors regarding those three residues are highly

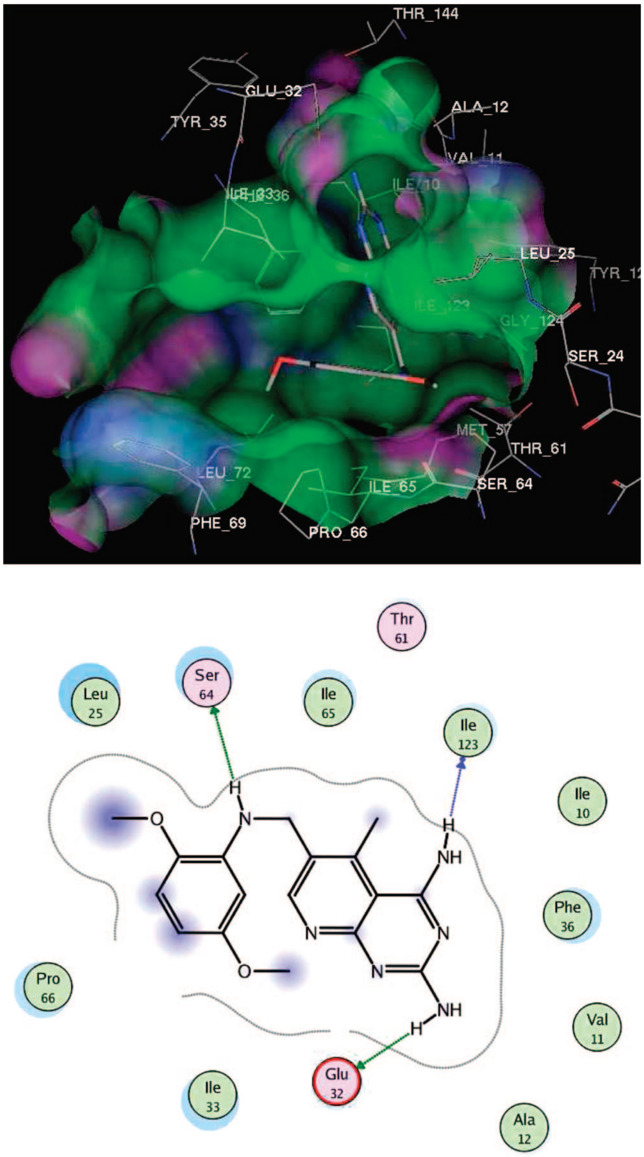

Figure 8. 3D- and 2D-representations of the interaction of the most active (training) ligand of the *P. carinii* DHFR data set with its receptor.

correlated to each other, and just the SigmaaSER42CA descriptor was "captured" by the genetic algorithm during the optimization of the model.

3D-QSAR and Docking Models for the CBG Data Set. Human corticosteroid-binding globulin (CBG), also a serine protease, binds corticosteroids in plasma, delivering them to sites of inflammation to modify the inflammatory response. CBG is an effective drug carrier for genetic manipulation, and hence there is immense biological interest in the location of the hormone binding site. The crystal structure of human CBG has not been determined, and a homology model was used in this study. As expected, the binding pocket is predominantly hydrophobic. Figure 7 shows such a hydrophobic binding pocket. Moreover, Gln-224 and Asp-256 form hydrogen bonds to the steroid. No descriptor regarding these two residues was captured during the optimization of the model. However, as for SHBG, some anchoring hydrogen bonds are expected for this system. Equation 12 shows the negative contribution of descriptors regarding polar residues and the relatively positive contribution for the hydrophobic groups. Consequently, in principle, the hypothesis for drug design for lead molecules interacting with CBG is similar to those ones described for SHBG.

3D-QSAR and Docking Models for the *P. carinii* DHFR Data Set. Figure 8 shows the active site of *P. carinii* DHFR[34] with the most active ligand docked into it. The active site is predominantly hydrophobic, but a couple of polar residues form hydrogen bonds with the ligand. Descriptors regarding those polar residues are not present on the constructed QSAR model (eq 13). However, SigmaaTYR129OH seems to be regarding to the hydrogen bond between the hydroxyl group of Tyr-129 and the cofactor NADPH. The highly negative regression coefficient of SigmaaLEU72N indicates that a polar group close to Leu-72 would disturb the favorable hydrophobic interactions that residue is able to form with the ligand, whereas SigmaaLYS37N shows how stable the interaction of Lys-37 seems to be and the presence of the polar group on the ligand. The high regression coefficient of RsTHR61N seems to indicate that a bulky group close to this small residue would result in unfavorable steric hindrance.

CONCLUSION

The results presented in this study illustrate the complex scenario common to any drug design project, where distinct tools are complementarily used in order to explore multiple aspects of the chemical space. The presented models are interpretable, with high statistical and predictive significance, and can be used for guiding ligand modification for the development of potential new inhibitors for several targets.

It was shown how docking and QSAR analysis can be used together for the construction of drug design hypothesis, when working with structurally diverse data sets. Moreover, the applicability of molecular fingerprint based cluster analysis for screening and resampling that kind of data set was also described. It was shown that the method is reliable and accurate enough for calculating QSAR models for structurally diverse ligand data sets.

The advantages of the proposed 3D-QSAR approach over other ones are as follows: (a) No hypothetical receptor structure is used, as in the case of 6D-QSAR. Instead, actual 3D-structures are used. (b) No ambiguous alignments are needed, since extensive docking simulations are carried out, so that the orientations of the ligands are the optima docked poses. (c) 'Induced' descriptors "capture" intermolecular interactions between ligands and receptors. (d) As compared to other methods, the computational cost is not so high.

Not including additional conformational of the ligands seems to be the main limitation of the approach. Probably, the proposition of "active" conformations for the ligands, based on conformational ensemble, would improve still more the statistical significance of the models. However, conformational ensemble calculation would certainly increase the computational cost.

ACKNOWLEDGMENT

O.A.S.-F. and A.C. were supported by the Genome Canada-funded PREPARE project (http://www.prepare.med.ubc.ca).

Supporting Information Available: Cross-correlation of both training and test sets with data for all of the data sets (Figures 1−8) and data sets selected as test sets (Tables 1−8).

This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Eposito, E. X.; Hopfinger, A. J.; Madura, J. D. Methods for applying the quantitative structure-activity relationship paradigm. *Methods Mol. Biol.* **2004**, *275*, 131–213.

(2) 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.

(3) 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.

(4) Robinsin, 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.

(5) Glen, W. G.; Dunn, W. J., III; Scot, D. R. Principal components analysis and partial least squares regression. *Tetrahedron Comput. Meth.* **1989**, *2*, 349–376.

(6) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509–10524.

(7) Vedani, A.; Dobler, M.; Lill, M. A. Combining protein and 6D-QSAR. Simulating the binding of structurally diverse ligands to estrogen receptor. *J. Med. Chem.* **2005**, *48*, 3700–3703.

(8) Tokarki, J. S; Hopfinger, A. J. Prediction of ligand-receptor thermodynamics by free energy force field (FEFF) 3D-QSAR analysis: application to a set of peptidometic rennin inhibitors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 792–811.

(9) Lill, M. A.; Vedani, A.; Dobler, M. Raptor: Combining Dual-Shell Representation, Induced-Fit Simulation, and Hydrophobicity Scoring in Receptor Modeling: Application toward the Simulation of Structurally Diverse Ligand Sets. *J. Med. Chem.* **2004**, *47*, 6174–6186.

(10) Gohlke, H.; Klebe, G. DrugScore Meets CoMFA: Adaptation of Fields for Molecular Comparison (AFMoC) or How to Tailor Knowledge-Based Pair-Potentials to a Particular Protein. *J. Med. Chem.* **2002**, *45*, 4153–4170.

(11) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2006**, *00(Database issue)*, D1−D4http://www.bindingdb.org/bind/index.jsp(accessed month year).

(12) 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.

(13) 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.

(14) 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.

(15) Sutherland, J. J.; O’Brien, L. A.; Weaver, D. F. Spline-Fitting with a Genetic Algorithm: A Method for Developing Classification Structure-Activity Relationships. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1906–1915.

(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. *Nucleic Acids Res.* **2000**, *28*, 235–242; http://www.pdb.org(accessed month year).

(17) *Maestro 8.5*; Schrödinger Inc.: San Diego, CA, 2007.

(18) *Glide; Version 4.5*; Schrödinger Inc.: San Diego, CA, 2007.

(19) Cherkasov, A. ‘Inductive’ Descriptors. 10 Successful Years in QSAR. *Curr. Comput.-Aided Drug Des.* **2005**, *1*, 21–42.

(20) 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.

(21) 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.

(22) 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.

(23) 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.

(24) Karakoc, A.; Cherkasov, A.; Sahinalp, S. C. Distance Based Algorithms for Small Biomolecule Classification and Structural Similarity Search. *Bioinformatics* **2006**, *22*, e243−251.

(25) 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.

(26) Cherkasov, A. Can ‘Bacterial-Metabolite-Likeness’ Model Improve Odds of in silico Antibiotic Discovery. *J. Chem. Inf. Model.* **2006**, *46*, 1214–1222.

(27) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared near neighbors. *IEEE Trans. Comput.* **1973**, *C-22*, 1025–1034.

(28) SVL exchange. http://svl.chemcomp.com/viewcat.php (accessed month year).

(29) 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.

(30) Holland J. H. Adaptation in Natural and Artificial Systems; Ann Arbor, MI, 1975.

(31) Rogers D. G/SPLINES: A Hybrid of Friedman’s Multivariate Adaptive Regression Splines (MARS) Algorithm with Holland’s Genetic Algorithm. *The Proceedings of the Fourth International Conference on Genetic 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) Grishkovskaya, I.; Avvakumov, G. V.; Sklenar, G.; Dales, D.; Hammond, G. L.; Muller, Y. A. Crystal structure of human sex hormone-binding globulin: steroid transport by a laminin G-like domain. *EMBO J.* **2000**, *19*, 504–512.

(34) Gagjee, A.; Yu, J.; McGuire, J. J.; Cody, V.; Galitsky, N.; Kisliuk, R. L.; Queener, S. F. Design, synthesis, and X-ray crystal structure of a potent dual inhibitor of thymydilate synthase and dihydrofolate reductase as an antitumor agent. *J. Med. Chem.* **2000**, *43*, 3837–3851.

(35) Panchal, R. G.; Hermone, A. R.; Nguyen, T. L.; Wong, T. Y.; Schwarzenbacher, R.; Schmidt, J.; Lane, D.; McGrath, C.; Turk, B. E.; Burnett, J.; Aman, M. J.; Little, S.; Sausville, E. A.; Zaharevitz, D. W.; Cantley, L. C.; Liddington, R. C.; Gussio, R.; Bavari, S. Identification of small molecule inhibitors of anthrax lethal factor. *Nat. Struct. Biol.* **2004**, *11*, 67–72.

(36) Ren, J.; Esnouf, R. M.; Hopkins, A. L.; Stuart, D. I.; Stammers, D. K. Crystallographic analysis of the binding modes of thiazoloisoindolinine non-nucleoside inhibitors to HIV-1 reverse transcriptase and comparison with modeling studies. *J. Med. Chem.* **1999**, *42*, 3845–3851.

(37) Lou, M. Antiviral drugs fit for a purpose. *Nature* **2006**, *443*, 37–38.

(38) Qiu, X.; Padmanabhan, K. P.; Carperos, V. E.; Tulinsky, A.; Kline, T.; Maraganore, J. M.; Fenton, J. W., II. Structure of the hirulog 3-thrombin complex and nature of the S’ subsites of substrates and inhibitors. *Biochemistry* **1992**, *31*, 11689–11697.

(39) Maignan, S.; Guilloteau, J.-P.; Pouzieux, S.; Choi-Sledeski, Y. M.; Becker, M. R.; Klein, S. I.; Ewing, W. R.; Pauls, H. W.; Spada, A. P.; Mikol, V. Crystal structures of human factor Xa complexed with potent inhibitors. *J. Med. Chem.* **2000**, *43*, 3226–3232.

(40) *MacroModel 9.5*; Schrödinger Inc.: San Diego, CA, 2007.

(41) *MOE: Molecular Operational Environment; Version 2004.03*; Chemical Computation Group Inc.: Montreal, Canada, 2004.

(42) Edelsbrunner, H.; Facello, M.; Fu, R.; Liang, J. Measuring Proteins and Voids in Proteins Proceedings of the 28th Hawaii International Conference on Systems Science; 1995; pp 256−264.

(43) Klein, C. D. P.; Muller, M. K.; Schellinski, C.; Landmann, S.; Hauschild, S.; Heber, D.; Mohr, K.; Hopfinger, A. J. Synthesis, Pharmacological and Biophysical Characterization, and Membrane-Interaction QSAR Analysis of cationic amphiphilic model compounds. *J. Med. Chem.* **1999**, *42*, 3874–3888.

(44) Kulkarni, A.; Han, Y.; Hopfinger, A. J. Predicting Caco-2 cell permeation coefficients of organic molecules using membrane-interaction QSAR analysis. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 331–342.

CI8001952