2D QSAR Consensus Prediction for High-Throughput Virtual Screening. An Application to COX-2 Inhibition Modeling and Screening of the NCI Database

Nicolas Baurin,[†] Jean-Christophe Mozziconacci,[†] Eric Arnoult,[†] Philippe Chavatte,[‡]
Christophe Marot,[†] and Luc Morin-Allory*,[†]

Institut de Chimie Organique et Analytique, UMR 6005, Université d'Orléans, BP 6759,
F-45067 Orléans Cedex 2, France, and Institut de Chimie Pharmaceutique Albert Lespagnol,
Université de Lille 2, BP 83, F-59006 Lille Cedex, France

Received July 25, 2003

Using classification (SOM, LVQ, Binary, Decision Tree) and regression algorithms (PLS, BRANN, k-NN, Linear), this paper details the building of eight 2D-QSAR models from a 266 COX-2 inhibitor training set. The predictive performances of these eight models were subsequently compared using an 88 COX-2 inhibitor test set. Each ligand is described by 52 2D descriptors expressed as van der Waals Surface Areas (P_VSA) and its COX-2 binding IC_{50}. One of our best predictive models is the neural network model (BRANN), which is able to select a subset, from the 88 ligand test set, that contains 94% COX-2 active inhibitors ($pIC_{50} > 7.5$) and detects 71% of all the actives. We then introduce a QSAR consensus prediction protocol that is shown to be more predictive than any single QSAR model: our C3 consensus approach is able to select a subset from the 88 ligand test set that contains 94% active inhibitors and 83% of all the actives. The 2D QSAR consensus protocol was finally applied to the high-throughput virtual screening of the NCI database, containing 193 477 organic compounds.

INTRODUCTION

To accelerate the discovery of new drugs, many biological events can be modeled by a cheminformatician from the data generated during a research project. Building QSAR models from Binding data, as well as Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADME/TOX) data can help to capture the knowledge from the experimental data and speed up the discovery of safe and efficient new chemical entities. For an efficient support to drug discovery project, the QSAR models must be able to be updated easily from the information flow produced by the synthesis/biological testing activities and able to virtually screen large structural databases in a short time. Those “reactivity and speed” characteristics for QSAR modeling are of utmost importance if one considers the volume of data produced by high-throughput screening and combinatorial synthesis. Another challenge that a cheminformatician has to face with QSAR models is the ability to manage various QSAR models on various targets. This challenge is requiring the same “reactivity and speed” characteristics if we consider the potential number of protein targets that is estimated from 2000 to 5000 from the human genome sequencing.[1]

The essential characteristic of any QSAR model is, of course, its predictive power. We detail in this paper how several 2D-QSAR models were built following a “reactive and fast” protocol. 266 COX-2 inhibitors were used to build eight QSAR models using various regression and classification techniques. We then looked at the predictive powers of the obtained models relative to an 88 ligand test set.

We chose to use 2D surface descriptors, because, among all the 2D descriptors available, we wanted the descriptors to be implicitly linked to the ligand binding: here, we used the P_VSA descriptors which express various physicochemical properties (electrostatic, lipophilic, steric, pharmacophoric) in terms of van der Waals Surface.[2] This is a good compromise between the speed required to calculate the descriptors and the molecular information we want to be related strongly to the ligand binding.

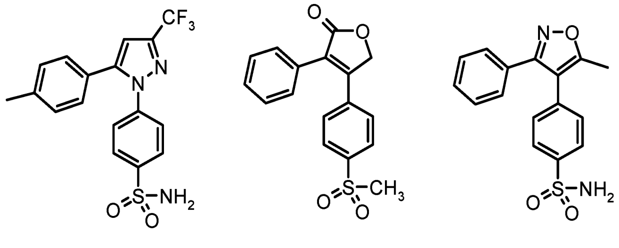
COX-1 and COX-2 are the two isoforms of cyclooxygenases, which are involved in the metabolism of prostaglandins by transforming arachidonic acid into PGH_2.[3] The isoform 2 of human cyclooxygenase (COX-2) is a major therapeutic target for inflammatory diseases since its selective inhibition has been shown to prevent prostaglandin synthesis at the inflammatory sites while producing reduced gastrointestinal and renal side effects compared to classical Nonsteroidal-Anti Inflammatory Drugs (NSAIDs).[4] Indeed, the nonselective COX-2 NSAIDs are inhibiting both the constitutive COX-1 isoform, which is involved in the gastrointestinal and renal homeostatic functions, and the inducible COX-2 enzyme expressed specifically during the inflammatory events. Human COX-1 and COX-2 are 60% structurally homologous, but in the active site only one amino acid differs (Ile523 for COX-1 is Val523 for COX-2).[5] The free space around Val523 in COX-2 is occupied by selective COX-2 inhibitors and confers a specific binding mode to the selective COX-2 inhibitors compared to classical NSAIDs such as ibuprofen.[5] The ongoing research efforts to design selective COX-2 inhibitor have produced several antiinflammatory drugs such as celecoxib, rofecoxib, and, more recently, valdecoxib (Figure 1). Other therapeutic applications of selective COX-2 inhibitors are now under investigation,

* Corresponding author phone: +33 (0) 2 38 41 70 42; e-mail: luc.morin-allory@univ-orleans.fr.

[†] Université d'Orléans.

[‡] Université de Lille 2.

10.1021/ci0341565 CCC: $27.50
Published on Web 10/23/2003



1 2 3

Figure 1. Selective COX-2 inhibitors on the market: **1** celecoxib, **2** rofecoxib, **3** valdecoxib.

including prevention of cancer[6,7] (celecoxib is used to treat the Familial Adenomatous Polyposis (FAP) syndrome) and prevention of Alzheimer disease.[8,9]

Four regression techniques were applied on the COX-2 training set data. We used one nonlinear technique (Bayesian Regularized Artificial Neural Network), two linear techniques (Partial Least Squares and MOE_LINEAR), and one similarity-based technique (Genetic Algorithm coupled to k Nearest Neighbors). Four classification algorithms were also applied to build QSAR models with which the prediction output is a class, i.e. either active or inactive. We used one supervised nonlinear technique (Learning Vector Quantization), one nonsupervised nonlinear technique (Self-Organizing Map), one supervised probabilistic technique (MOE_BINARY), and one supervised Recursive Partitioning technique (MOE_DT). This is not an exhaustive list of all the available QSAR techniques, but it is representative of the various fast approaches developed. Some of these tools are available in the MOE software, used here as our core environment, while the others are freely available for academics. These several diverse techniques were investigated to compare their predictive performances on this target in order to select the best methods for virtual screening. After learning, our eight 2D-QSAR COX-2 models will not give similar type of predictions, since four of them will be quantitative and four will be qualitative. To compare their predictive powers, we examined their ability to speed up the discovery of active COX-2 inhibitors, i.e. their COX-2 discriminative power that enables the computational chemist to enrich the molecular set predicted as active with real active COX-2 inhibitors. As an additional validation, we have built the same models with the scrambled pIC_{50} activity data to check that the quality of the COX-2 QSAR models obtained was not due to chance correlation. The detailed observations of how well each technique performs in finding true positives, with as few false positives as possible, made us introduce a consensus prediction rather than choose one best technique and leave the others out. Consensus scoring has been shown in docking studies to decrease the false positives compared to single scoring.[10−12] The same tendencies are observed in this paper with QSAR consensus prediction, which uses several QSAR models to optimize the hit rate in virtual screening. To illustrate how the whole 2D QSAR consensus approach can be used as a fast virtual screening protocol, we finally mined a freely accessible version of the open NCI database.

MATERIALS AND METHODS

Selection of the Ligands. We selected from the literature a specific series of selective COX-2 inhibitors, the diarylheterocycles, to build our 2D-QSAR models. The three selective COX-2 inhibitors on the market (Figure 1) belong to this series that was already used for 3D-QSAR CoMFA[13] studies and recently for a 2D-QSAR study.[14] The COX-2 inhibition powers of the 354 ligands finally selected have all been tested following a common protocol:[15] this test uses a human recombinant COX-2 isoform and the detection of the transformed arachidonic acid into PGE_2 with an immunoenzymatic test ELISA (Enzyme Linked ImmunoSorbent Assay). The 354 COX-2 inhibitors belong to nine chemical families (Figure 2) and their COX-2 inhibition power is expressed as pIC_{50}, i.e. the log $(1/IC_{50})$ where IC_{50} is the inhibitor concentration needed to inhibit 50% of the COX-2 enzymatic activity.

The nine chemical families differ with respect with their central heterocyclic scaffold. These scaffolds were spiroheptene (30 ligands[16]), pyrrole (22 ligands[17]), imidazole (127 ligands[18,19]), cyclopentene (40 ligands[20,21]), benzene (44 ligands[22]), spiroheptadiene (2 ligands[16]), thiophene (1 ligand[23]), isoxazole (2 ligands[24]), and pyrazole (86 ligands[23]).

Building of the Ligands. The molecular structures of the 354 COX-2 inhibitors were built and stored in a database of the MOE software.[25] The database was further used to centralize the other data (COX-2 pIC50, P_VSA descriptors, and QSAR predictions).

P_VSA Descriptors. The calculation of P_VSA descriptors is implemented in the MOE software.[2,25] Those descriptors are based on the approximation at the atomic level of the molecular van der Waals surface area, VSA_i, along with some other molecular property, P_i. VSA_i is calculated using parameters from the MMFF94[26] force field, and P_i is an atomic contribution to logP(o/w),[27] molar refractivity,[27] electrostatics,[28] and pharmacophoric properties. Thus, the P_VSA descriptors only need 2D molecular connectivity information. Each descriptor in a series is defined to be the sum of the VSA_i over all atoms i for which P_i is in a specified range [a,b]. The ranges had been determined by percentile subdivision over the Maybridge database.[29] 52 descriptors were calculated (Table 1) including 10 SlogP_VSA descriptors describing hydrophobobic and hydrophilic interactions, 26 PEOE_VSA descriptors describing direct electrostatic interactions, 8 SMR_VSA descriptors describing polarizability, and 7 P_VSA descriptors describing pharmacophoric features and the total molecular surface area.

These descriptors specifically express structural information as molecular surfaces. Because the molecular surface is the interface through which intermolecular interactions happen, we think this is an advantage for P_VSA descriptors compared to other 2D descriptors expressing molecular information in ways much less related to the physicochemical reality of the ligand-macromolecule recognition event. Furthermore, even a small number of P_VSA descriptors has been shown to contain much of the information encoded by larger sets of popular descriptors.[2] Nevertheless, the P_VSA descriptors are impossible to retro-visualize as specific chemical features, and then interpreting a COX-2 QSAR model using these 2D descriptors is neither straightforward nor of great help to guide the synthesis in an explicit way. Their main disadvantage is the lack of information about conformations or configurations of the products. Being based on the connectivity, they can be computed on a standard PC for 10^6-10^7 compounds a day. Again, the true aim of this

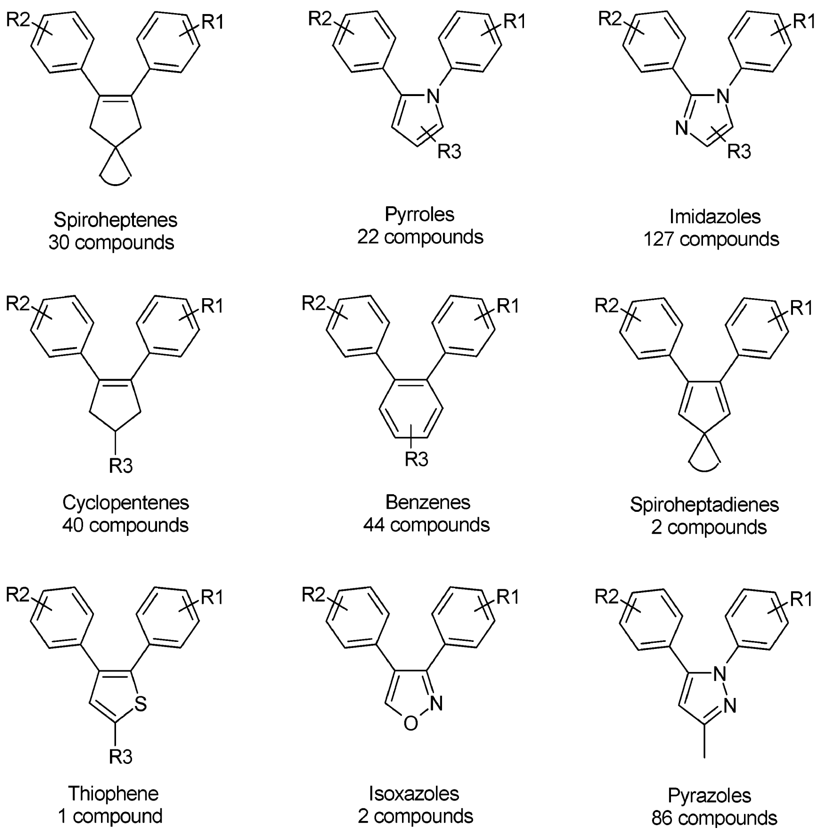


Figure 2. Definition of the nine families of COX-2 inhibitors studied.

Table 1. Description of the 52 P_VSA Descriptors Initially Calculated for Each of the 354 COX-2 Inhibitors Studied

descriptors	description
SlogP_VSA0, SlogP_VSA1, ..., SlogP_VSA9	sum of VSAi with LogPi ⇐ −0.4, ∈]−0.4;−0.2], ∈]−0.2;0], ∈]0;0.1], ∈]0.1;0.15], ∈]0.15;0.20], ∈]0.20;0.25], ∈]0.25;0.30], ∈]0.30;0.40], > 0.40
PEOE_VSA+0, PEOE_VSA+1, ..., PEOE_VSA+6	sum of VSAi with qi ∈]0.00;0.05], ∈]0.05;0.10], ∈]0.10;0.15], ∈]0.15;0.20], ∈]0.20;0.25], ∈]0.25;0.30], > 0.3
PEOE_VSA-0, PEOE_VSA-1, ..., PEOE_VSA-6	sum of VSAi with qi ∈]−0.05;0.00], ∈]−0.10;−0.05], ∈]−0.15;−0.10], ∈]−0.20;−0.15], ∈]−0.25;−0.20], ∈]−0.30;−0.25]], < −0.30
PEOE_VSA_POS	sum of VSAi with qi > 0
PEOE_VSA_NEG	sum of VSAi with qi < 0
PEOE_VSA_PPOS	sum of VSAi with qi > 0.2
PEOE_VSA_PNEG	sum of VSAi with qi < −0.2
PEOE_VSA_HYD	sum of with |qi | ⇐ 0.2
PEOE_VSA_POL	sum of VSAi with |qi | > 0.2
PEOE_VSA_FPOS	sum of VSAi with qi > 0 divided by the total surface area
PEOE_VSA_FNEG	sum of VSAi with qi < 0 divided by the total surface area
PEOE_VSA_FPPOS	sum of VSAi with qi > 0.2 divided by the total surface area
PEOE_VSA_FPNEG	sum of VSAi with qi < −0.2 divided by the total surface area
PEOE_VSA_FHYD	sum of VSAi with |qi| ⇐ 0.2 divided by the total surface area
PEOE_VSA_FPOL	sum of VSAi with |qi| > 0.2 divided by the total surface area
SMR_VSA0, SMR_VSA1, ..., SMR_VSA7	sum of VSAi with SMRi ∈ [0;0.11], ∈]0.11;0.26], ∈]0.26;0.35], ∈]0.35;0.39], ∈]0.39;0.44], ∈]0.44;0.485], ∈]0.485;0.56], > 0.56
vsa_acc	sum of VSAi for hydrogen bond acceptors
vsa_acid	sum of VSAi for acidic atoms
vsa_base	sum of VSAi for basic atoms
vsa_don	sum of VSAi for hydrogen bond donors
vsa_hyd	sum of VSAi for hydrophobic atoms
vsa_pol	sum of VSAi for polar atoms
vsa_other	sum of VSAi for atoms typed as other
vdw_area	sum of all VSAi in the molecule

kind of approach is to quickly build QSAR models, which, if statistically validated to be predictive, will quickly screen in silico large databases of small molecules to identify more active molecules.

Building of the Training and Test Sets. We rationally selected 3/4 of all the available molecules to be part of the training set. This rational selection is a distance-based experimental design,[30] which chose 266 molecules with an optimal diversity from the 354 molecules available. A C++/wxpython program developed in-house implementing this experimental design algorithm was used. Those 266 molecules were chosen to have the largest Tanimoto distances

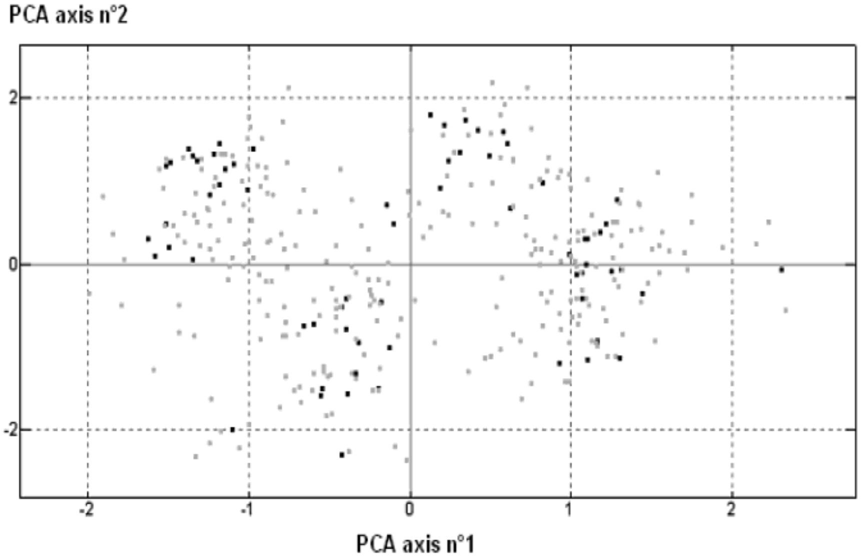


Figure 3. Plane created by the two first axes from PCA on the 52 P_VSA descriptors of the 354 COX-2 inhibitors. The 266 ligands of the training set are shown as gray squares, and the 88 ligands of the test set are shown as black squares. The PCA axes no. 1 and no. 2 account for, respectively, 21% and 15% of the whole molecular information.

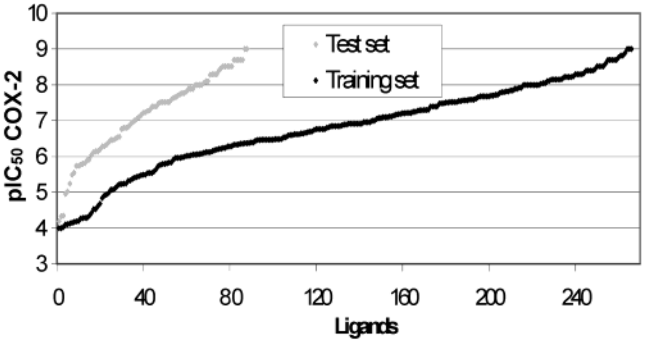


Figure 4. COX-2 pIC50 distributions of the 266 ligand training set and the 88 ligand test set.

from each other. Each Tanimoto distance d_{ij} between molecule i and molecule j is calculated from 53 variables (52 P_VSA descriptors + COX-2 pIC_{50}) and is expressed as

$$d_{ij} = 1 - \frac{\sum_{k=1}^{n}(a_{ik} * a_{jk})}{\sum_{k=1}^{n} a_{ik^2} + \sum_{k=1}^{n} a_{jk^2} - \sum_{k=1}^{n}(a_{ik} * a_{jk})} \quad (n = 53)$$

where a_{ik} is the value of the kth variable for molecule i.

We included the COX-2 pIC_{50} in addition to the P_VSA descriptors for the distance-based experimental design in order to obtain the most diverse 266 ligand training set with respect to both structural and biological data. In Figure 3, we can see that the training set is widely spread in the plane defined by the two main Principal Component Analysis (PCA) axes calculated from the 52 P_VSA descriptors describing the 354 COX-2 inhibitors.

On the other hand, the test set whose structural diversity has not been optimized during the experimental design process is plotted on discrete portions of this same plane. In Figure 4, we can see that the COX-2 pIC_{50} distributions are homogeneous within both the training set and test set. This is an essential prerequisite for the training set if we want our QSAR techniques to efficiently discriminate the active COX-2 inhibitors (high pIC_{50}) from the inactive COX-2 inhibitors (low pIC_{50}).

Concerning the test set, this homogeneous pIC_{50} distribution will make sure that we test the predictive power of the QSAR models with inactive and active COX-2 inhibitors, i.e under realistic conditions where we will have to screen virtual molecular sets with both types of compounds mixed.

QSAR Methodology. Once the 266 COX-2 inhibitors of the training set were rationally chosen, the structural training data (a 266*52 matrix) was centered (training set means) and scaled (training set standard deviations). The same means and standard deviations were used to modify the structural test data (an 88*52 matrix). We then eliminated those of the 52 P_VSA descriptors that were invariant for this 266 ligand set because a descriptor that is constant for all the structures of the training set is not able to explain the structural differences between the active and inactive structures. We also eliminated the descriptors whose values were strictly correlated with values of another descriptor to avoid useless redundancy. Three descriptors were eliminated: PEOE_VSA_FPOL (identical with PEOE_VSA_FPOS), PEOE_VSA_POL (identical with PEOE_VSA_POS), and vsa_base (invariant). The quantitative QSAR models are built from the 266 ligands of the training set described each with 49 P_VSA descriptors (scaled and centered) and one COX-2 pIC_{50} value. The qualitative QSAR models are built from the same training set, each ligand being described by the 49 P_VSA descriptors and one COX-2 inhibition class, either the active class or inactive class if the ligand's COX-2 pIC_{50} is, respectively, greater (or equal) than 7.5 or lower than 7.5. For the scrambling validation, each COX-2 inhibitor was attributed a randomly chosen pIC_{50} from the real pIC_{50}'s, and the eight corresponding QSAR models were built. Once the QSAR models were built, each molecule of the training and test sets has its COX-2 activity predicted; these predictions are the basis of the results presented in this paper.

Below we give a brief description of each of the eight QSAR techniques used in this study. The fully detailed protocols are given as Supporting Information.

BRANN. Bayesian Regularized Artificial Neural Network is a nonlinear regression technique using a neural network to establish a relationship between the P_VSA descriptors (inputs) and the COX-2 pIC_{50} (output). This neural network differs from classical neural nets in that each weight is replaced by a distribution of weights whose modulations during the training is ruled by Bayesian inference.[31] We used a neural network implementing a procedure called Automatic Relevance Determination (ARD),[23] which automatically modulates the strength of each input in the network with respect to its relevance for the quality of the prediction during the training. Burden et al. first introduced BRANN-ARD for QSAR modeling.[33] R. M. Neal's software for Flexible Bayesian Modeling was used.[34]

GA-kNN. Genetic Algorithm coupled to k Nearest Neighbors[35−37] is a similarity-based prediction technique. Independently from the GA variable selection procedure, the COX-2 pIC_{50} prediction of a molecule is made by searching in the training set the k nearest neighbors, i.e. k COX-2 inhibitors in the training set with the smallest Euclidean distance from this molecule. The Euclidian distance is calculated from the 49 P_VSA available descriptors. Then, the predicted COX-2 pIC_{50} is the mean of the k known pIC_{50} values. A C++ program developed in-house, using the M. Wall "GAlib" library,[38] was used.

PLS. Partial Least Squares is a widely used linear regression technique.[39] PLS determines the coefficients of the linear relationship between the COX-2 pIC_{50} and the 49 P_VSA descriptors, via the extraction of p latent variables. Each latent variable is expressed as a linear combination of the 49 P_VSA initial descriptors, whose variance is highly correlated with the COX-2 pIC_{50} variance, and each latent variable is independent from the others. The optimal number p of latent variables extracted was the one with the best q^2 determined by a Leave-One-Out cross-validation procedure, to prevent overtraining. A C++ program developed in-house, using the property-based PLS algorithm,[40] was used. The final PLS model retained after training on the 266 ligand training set has five latent variables.

MOE_LINEAR. MOE_LINEAR is a linear regression technique implemented in the MOE software.[25] PLS determines the coefficients (b_0, b_1, ..., b_{49}) of the linear relationship, between the COX-2 pIC_{50} and the 49 P_VSA descriptors, via the solving of the conditions expressed by the derivation of the Mean Squared Error parameter (MSE) with respect to the b_0 and b_i parameters. MSE is defined as

$$\mathrm{MSE}(b_0, b_i) = \frac{\sum_{k=1}^{n}(\mathrm{pIC}_{50k} - (b_0 + \sum_{i=1}^{49} b_i{*}\mathrm{P\_VSA}_i))^2}{n} \qquad (n = 266)$$

No cross-validation is used, i.e. all the nonorthogonal structural information of the 266 ligand training set is used for the regression.

LVQ. Learning Vector Quantization is a classification technique using a special kind of supervised neural network.[41] LVQ handles a set of labeled vectors (active and inactive), here of dimension 49. There coordinates will "learn" to be characteristic of either one of the two classes found among the 266 COX-2 inhibitors. During the training, the vectors labeled "active" will learn to adapt their coordinates to be characteristic of the 87 COX-2 "active" reference vectors found in the 266 ligand training set ($pIC_{50} \geq 7.5$). In the same time, the vectors labeled "inactive" will learn to adapt their coordinates to be characteristic of the 179 inactive reference vectors ($pIC_{50} < 7.5$). The learning is basically as following: For each reference vector, the vector that is closest is detected, with respect to its Euclidean distance. The coordinates of this winning vector are then adapted. The direction of the adaptation depends on whether the class of the winning vector is the same as the class of the reference vector. If the classes are the same, the vector is moved closer to the reference vector, otherwise it is moved farther away (the details of the movement are specific of LVQ procedures, named oLVQ1, LVQ3, and LVQ2.1). We used the Kohonen's team LVQ_PAK software[42] to implement a procedure fully described in the Supporting Information.

SOM. The Self-Organizing Map algorithm uses a neural network, whose neurons (vectors) will learn to adopt a 2D topology related to the similarity of the high-dimensional inputs.[43] The SOM algorithm is very similar to the LVQ algorithm by using a vector quantization approach, but SOM is unsupervised because it does not take into account the classes of the references vectors during learning. After learning, the topology of the map obtained is only related to the similarity between the sets of 49 P_VSA descriptors among the 266 reference vectors found in the 266 COX-2 ligand training set. The 2D map obtained after learning might be used as a classification tool by labeling the vectors of the map with the class of the closest reference vector in the training set with respect to its Euclidean distance (calibration step). The learning process is similar to the LVQ learning since, for each reference vector, the vector that is closest is detected. The coordinates of this winning vector are then adapted to be closer to the reference vector. In a lesser extent, the coordinates of the vectors in the close neighborhood of the winning neuron are also adapted. We used the Kohonen's team SOM_PAK software [44] to implement a procedure fully described in the Supporting Information.

MOE_BINARY. This qualitative technique called Binary QSAR estimates the probability density Pr ($Y = 1 \mid X = x$), where Y is a binary variable ($Y = 1$ for an active COX-2 inhibitor, $pIC_{50} \geq 7.5$, and $Y = 0$ for an inactive COX-2 inhibitor, $pIC_{50} < 7.5$) and X is an n-vector containing the P_VSA values ($n = 49$) for a molecule for which we want to predict the COX-2 pIC_{50} value. We used the MOE software[25] that implements the Binary QSAR procedure to estimate the above probability density by applying the Bayes theorem.[45] The expression of this probability density assumes the mutual independency of the variables, and thus it is based upon the transformed descriptors, Z_i, calculated with a Principal Component Analysis (PCA) on the 49 P_VSA descriptors of the 266 ligand training set (we used here all the significant principal components extracted with the PCA, so that all the structural information contained in the training set was used). The expression of the probability density Pr ($Y = 1 \mid X = x$) is calculated from histograms approximating the density functions $P(Z_i = z_i \mid Y = 1)$ and $P(Z_i = z_i \mid Y = 0)$.

MOE_DT. This classification technique, implemented in the MOE software,[25] is based upon a Decision Tree algorithm, also called Recursive Partitioning (RP).[46] Our Decision Tree model is built from the 266 ligand training set and their known COX-2 classes (active, $pIC_{50} \geq 7.5$, or inactive, $pIC_{50} < 7.5$). We will predict the class of a molecule by running it through the MOE_DT COX-2 tree: the molecule enters the tree at the root node and falls down the tree from node to node. At each node, a rule based upon a P_VSA descriptor is applied to the item, and, depending on the outcome, the item passes into one branch or another. The item ultimately drops into a leaf node of the tree, which has an associated class assignment (active or inactive). For the training of the tree, we use a cross-validation procedure (five groups) and all default options. Our final MOE_DT model has only one rule: ≪ SlogP_VSA1 ≤ 49.1 ≫ for active COX-2 inhibitors.

Presentation of the Results. In this paper we compare the predictive powers of classification and regression models. That is why we do not report the classical regression models parameters (r^2, q^2) but two parameters (**%Actives**, **%Actives detected**) which can be calculated from the quantitative predictions made by regression models as well as from the qualitative predictions made by the classification models. A molecule will be predicted as active if a quantitative model predicts its COX-2 pIC_{50} greater than 7.5, or if a qualitative model classifies this molecule into the active category. A COX-2 inhibitor can then be classified into four types depending on its predicted and real COX-2 activity: True Positive (TP), False Positive (FP) when an inactive product is predicted positive, True Negative (TN), or False Negative (FN) when an active compound is predicted inactive. Using the TP/FP/TN/FN classification of predictions, we then calculated the two parameters, **%Actives** and **%Actives detected**. **%Actives** is the percentage of real active inhibitors in the subset predicted as active, and it is defined as **%Actives = TP/(TP + FP)**. **%Actives detected** is the percentage of real active inhibitors detected among all the real active inhibitors, and it is defined as **%Actives detected = TP/(TP + FN)**. A perfect model would detect all the real active inhibitors (**%Actives detected** = 100%) without mistakes (**%Actives** = 100%).

Virtual Screening Application. To illustrate the mining of large molecular databases with such QSAR models, we used the open part of the NCI structural database. It is a valuable database and one of the largest datasets of diverse organic compounds freely available. Hence this molecular database is a library well suited to test virtual screening algorithms.[47]

In this work, we used an SD file[48] containing 249 081 structures from the year 2000, publicly and freely available data of NCI's Developmental Therapeutics Program (DTP).[49] The SD file downloaded was stored in a MOE molecular database. The nonorganic compounds (containing atoms different from oxygen, nitrogen, carbon, sulfur, hydrogen, fluorine, chlorine, or bromine) and those with problems in the structure during the importation phase were removed. Then, on the basis of a nonstereospecific SMILES code generated by MOE for each structure, we eliminated the duplicates. The 52 descriptors were calculated on the 193 477 resulting unique organic compounds.

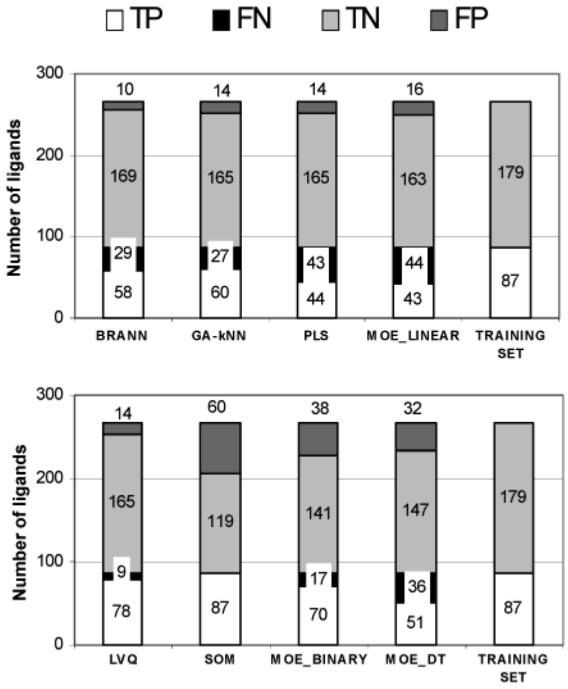


Figure 5. Classification of the predictions made by the four quantitative COX-2 models (top) and the four qualitative COX-2 models (bottom) on the 266 ligand training set.

Hardware. All programs needed for this study were run on a desktop PC (Windows 2K, Pentium III 866 MHz, 256Mo RAM) except the SOM_PAK, LVQ_PAK, and FBM programs that were run on an SGI workstation (IRIX 6.5, MIPS 225 MHz, 512Mo RAM).

RESULTS AND DISCUSSION

Learning Performances on the 266 Ligand Training Set. In Figures 5 and 6, we can see in details how well the eight COX-2 models have been trained by looking at the distribution of the predictions for the 266 molecules of the training set. Figure 6 shows the difficulty for models to enrich the selected subset with active inhibitors, while detecting as many as possible of all the active COX-2 inhibitors.

The BRANN and GA-kNN models are the best models among the 4 quantitative models since they predict as active a subset that contains, respectively, 85% and 81% of active inhibitors and that contains, respectively, 67% and 69% of all the active inhibitors from the training set. The LVQ model is the best model among the four qualitative models since it predicts as active a subset that contains 85% of active inhibitors and that contains 90% of all the active inhibitors from the training set. We see that the LVQ model is the best of the eight models since, with equal performances concerning the proportion of True Positives in the subset predicted as active (**%Actives** = 85%), it detects more active inhibitors from the 266 COX-2 inhibitors of the training set (**%Actives detected** = 90%, against **%Actives detected** = 67% for the quantitative model BRANN).

Predictive Performances on the Test Set. In Figures 7 and 8, we can check the predictive performance of each of the eight COX-2 models by looking at the distribution of the predictions for the 88 molecules of the test set.

As those molecules were not used for the training of the models, we simulated here the practical application of these

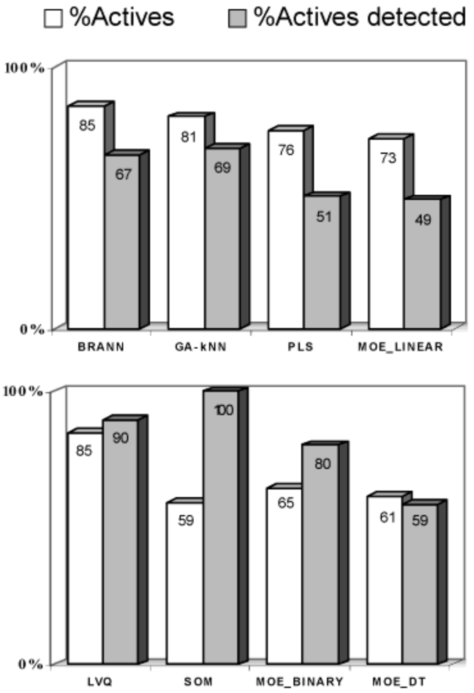


Figure 6. Composition of the subset selected from the 266 ligand training set by one of the four quantitative COX-2 models (top) or one of the four quantitative COX-2 models (bottom).

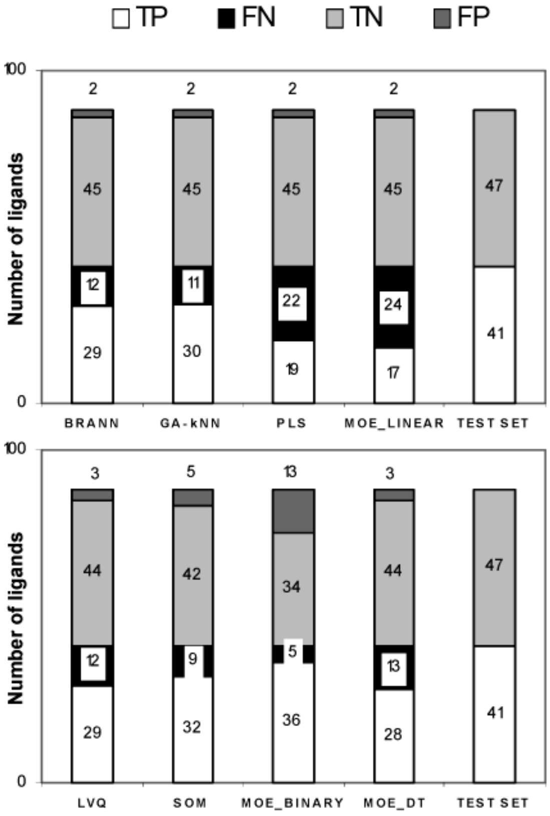


Figure 7. Classification of the predictions made by the four quantitative COX-2 models (top) and the four qualitative COX-2 models (bottom) on the 88 ligand test set.

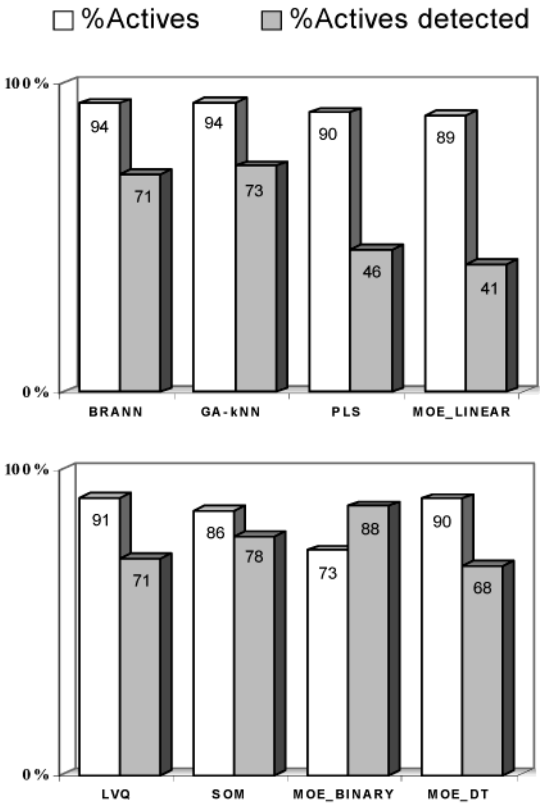


Figure 8. Composition of the subset selected from the 88 ligand test set by one of the four quantitative COX-2 models (top) or one of the four quantitative COX-2 models (bottom).

models during a virtual screening when they must detect a maximum of active COX-2 inhibitors while being wrong as rarely as possible. Interestingly, the trends are similar to what was observed when looking at the learning performances on the 266 ligand training set. The BRANN and GA-kNN models are the most predictive among the 4 quantitative models since they predict as active a subset that contains 94% of active inhibitors and that contains, respectively, 71% and 73% of all the active inhibitors from the test set. The LVQ model selects the most active enriched subset among the 4 qualitative models (**%Actives** = 91%).

To further validate the predictive power of our eight COX-2 models, we built the corresponding scrambled models, and we looked at their enrichment factors when predicting the nonscrambled 88 ligand test set. Globally, all the scrambled models do not succeed in enriching a molecular set better than a random selection (enrichment factor = 1). This confirms that the link established by the nonscrambled COX-2 models between the COX-2 inhibition power and the P_VSA descriptors is not due to chance correlation. In addition, a further validation was carried out to check the efficiency of the genetic algorithm part of the GA-kNN algorithm. The predictions calculated with the optimized model are really better than those obtained with all the descriptors (30 versus 19 True Positives for the model with all the descriptors).

At this stage, we could choose to screen any virtual library with the BRANN or the GA-kNN models, which were shown as the best ones to enrich a molecular subset with active COX-2 inhibitors. Instead, we carefully inspected the good predictions (True Positives) made by our eight models on the 88 ligand test set. Table 2 displays the results of the comparisons of the populations of True Positives obtained by two COX-2 models. We see that each model detects True Positives in common with another model but also that each model detects True Positives that another model does not detect.

Table 2. Percentage of True Positives Detected Both by Two COX-2 Models from the 88 Ligand Test Set (Boldface) and Detected by Only One of Two COX-2 Models (Italic)[a]

	BRANN (%)	GA-kNN (%)	PLS (%)	MOE_LINEAR (%)	LVQ (%)	SOM (%)	MOE_BINARY (%)	MOE_DT (%)
BRANN	100	**79**	**66**	**59**	**66**	**69**	**67**	**73**
GA-kNN	*21*	100	**63**	**57**	**79**	**72**	**69**	**81**
PLS	*34*	*37*	100	**80**	**50**	**50**	**41**	**52**
MOE_LINEAR	*41*	*43*	*20*	100	**59**	**44**	**39**	**61**
LVQ	*34*	*21*	*50*	*41*	100	**69**	**67**	**78**
SOM	*31*	*28*	*50*	*56*	*31*	100	**70**	**67**
MOE_BINARY	*33*	*31*	*59*	*61*	*33*	*30*	100	**68**
MOE_DT	*27*	*19*	*48*	*39*	*22*	*33*	*32*	100

[a] Thus, 79% of the True Positives detected by the BRANN and GA-kNN models are the same molecules. Conversely, 21% of the True Positives detected by the BRANN and GA-kNN models are specific of the BRANN model or GA-kNN model.

Table 3. Distribution of the 193 477 NCI Organic Molecules, with Respect to the QSAR Consensus Protocol Which Predicted Them as COX-2 Active Compounds ($pIC_{50} \geq 7.5$)

QSAR consensus protocol	C1	C2	C3	C4	C5	C6	C7	C8
no. of molecules predicted as COX-2 active inhibitors	193 117	170 319	67 596	25 113	7 641	1 835	274	22

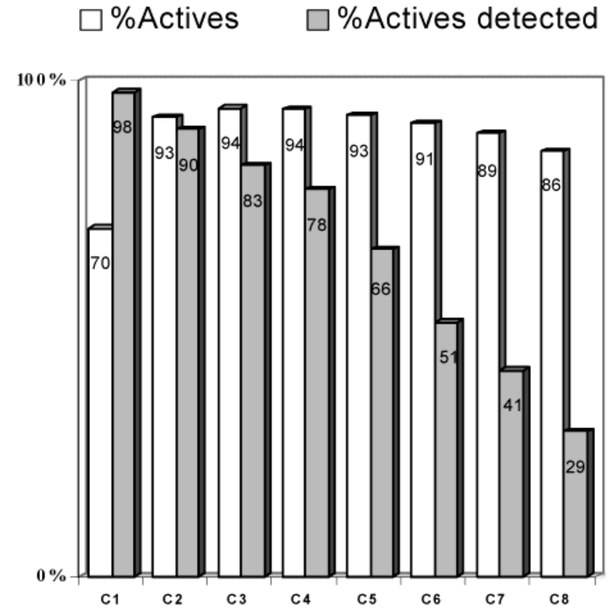


Figure 9. Composition of the subset selected from the 88 ligand test set by one of the eight Consensus prediction procedures.

Consensus Prediction. This observation led us to apply a QSAR-based screening procedure named "Consensus prediction": We have thus screened the 88 ligand test set using the eight COX-2 models, and we then chose to select a subset composed of molecules which were predicted as active COX-2 inhibitors by at least one of the eight COX-2 models. This procedure is called consensus prediction C1. We then applied consensus prediction C2 (selection of molecules predicted active COX-2 inhibitors by at least two out of our eight COX-2 models), C3, ..., and C8 (selection of molecules predicted as active COX-2 inhibitors by all eight COX-2 models). Figure 9 shows the predictive performances of the eight consensus prediction procedures for the 88 ligand test set.

The consensus prediction C1 detects more active inhibitors than the best model alone (**%Actives detected** = 98% against 88% for the MOE_BINARY model). This underlines the fact that the models explored subspaces of all the solutions (True Positives), and when combining all their predictions, we are able to detect more solutions. At the same time, combining all their predictions results in an increase of the False Positives rate (**%Actives** = 70%) that is worse than when using only one model. An evolution of the predictive performances is seen when the conditions become progressively stricter concerning the choice of the selected molecules: From C1 to C8, fewer and fewer active COX-2 inhibitors are detected (**%Actives detected** falls from 98% to 29%), and the enrichment with active inhibitors increases from 70% to 86% with an optimum at 94%. This underlines that, on one hand, a subset of True Positives is detected by all eight COX-2 models (29% of all the active inhibitors in the 88 ligand test set) but, on the other hand, a subset of False Positives exists which is composed of molecules that all our eight COX-2 wrongly predict. This is the reason, despite the very thorough conditions imposed by the C8 consensus prediction procedure, we do not select a molecular set composed only of active COX-2 inhibitors. Nevertheless, the molecular subset selected by the consensus prediction C8 is composed of 12 True Positives and two False Positives: those two False Positives have COX-2 pIC_{50} of 7.28 and 7.48, which means that they are close to the limit between active and inactive inhibitors. Those two False Positives are not reducing the value of the C8 procedure since those molecules might be considered as interesting COX-2 inhibitors. Moreover, considering those two molecules as active inhibitors, the **%Actives** parameter would reach 74% for the C1 procedure, then 98% for the C2 procedure, and 100% for the C3 to C8 procedures.

The C1 and C8 procedures show two extreme characteristics ranging from an optimal detection of the active inhibitors (C1, but with a nonnegligible number of False Positives) to an optimal enrichment with active inhibitors (C8, but with a nonnegligible missing of True Positives). It is noteworthy that between these two extremes, the C3 and C4 procedures show good general predictive performances compared to any single model.

Virtual Screening of the NCI Database. The 193 477 compounds from the NCI molecular database were virtually screened by calculating predictions with the eight models described above. Then we applied the QSAR consensus protocol. Preparation of this molecular database and calculation of the descriptors took a few hours. The calculation of the predictions is also very fast: fewer than 10 minutes for each of the eight models. These times are compatible with the high-throughput virtual screening purpose of this work. In Table 3, the predictions are analyzed using consensus approaches C1 to C8.

The C1 consensus protocol selected 99.8% of the whole database. This can be explained by the crudeness of at least one out of our eight models: The MOE-DT method uses only one condition to discriminate between active and inactive compounds. A large number of the NCI compounds (86%) are predicted as active by this model. C6 and C7 consensus predictions led to a more restrictive selection of compounds. 1835 and 274 compounds are predicted as active COX-2 inhibitors by, respectively, at least six models and at least seven models. The set of 7641 compounds selected by the C5 consensus protocol comprises 93 compounds with sulfonamide group, 40 compounds with sulfoxydes, and 232 compounds with sulfones or sulfonates. These chemical groups are present in the training and test set COX-2 inhibitors and their role is well-known for the binding process with COX-2. That is why we clearly expect a lot of False Positives if compounds were only selected by this QSAR consensus. The ultimate validation would imply either to experimentally test the COX-2 inhibition potency of the compounds selected by the C6 or C7 consensus protocol or to use a virtual docking approach with the compounds selected by C3 or C4 consensus. We are currently working on this second way.

CONCLUSION

Predictive QSAR models of the COX-2 inhibition were developed using a large set of selective COX-2 inhibitors. Compared to already published 3D-QSAR COX-2 work[13] using CoMFA, the 2D-QSAR COX-2 models we developed are applicable for rapid virtual screening of large databases. By using eight QSAR techniques which can be seen as a sample of the approaches available today for deriving a QSAR model, the complementary predictivity of these models was shown. We further introduce QSAR consensus prediction as a better predictive protocol compared to the arbitrary choice of a single QSAR model. Several QSAR consensus strategies are then possible: A low level consensus procedure enables one to maximize the detection of active compounds, such as the C1 procedure applied on the COX-2 test set. Higher level consensus procedures can maximize the enrichment with active compounds, as reported for the C3 to C8 procedures applied on the COX-2 test set. The application of this methodology was illustrated by virtually screening the NCI database. This approach, based on rapidly computed but low-information-content descriptors, uses the complementarities of various statistical methods to enrich a set of products with putative active ones and may be a fruitful preliminary step for a docking approach.

ACKNOWLEDGMENT

The authors are grateful to Dr. M. Afshar for his help in the redaction of the final version of the paper. Special thanks are due to Dr. Marc C. Nicklaus for his highly valuable comments about the manuscript.

Supporting Information Available: Tables of the 354 COX-2 inhibitor structures and respective pIC_{50} values as well as a complete description of the statistic methods. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Caron, P. R.; Mullican, M. D.; Mashal, R. D.; Wilson, K. P.; Su, M. S.; Murcko, M. A. Chemogenomic approaches to drug discovery. *Curr. Opin. Chem. Biol.* **2001**, *5*, 464−470.

(2) Labute, P. A widely applicable set of descriptors. *J. Mol. Graph. Model.* **2000**, *18*, 464−477.

(3) Gilroy, D. W.; Colville-Nash, P. R. New insights into role of COX-2 in inflammation. *J. Mol. Med.* **2000**, *78*, 121−129.

(4) Futaki, N.; Yoshikawa, K.; Hamasaka, Y.; Arai, I.; Higuchi, S.; Iizuka, H.; Otomo, S. NS-398, a novel nonsteroidal antiinflammatory drug with potent analgesic and antipyretic effects, which causes minimal stomach lesions. *Gen. Pharmacol.* **1993**, *24*, 105−110.

(5) Kurumbail, R. G.; Stevens, A. M.; Gierse, J. K.; McDonald, J. J.; Stegeman, R. A.; Pak, J. Y.; Gildehaus, D.; Miyashiro, J. M.; Penning, T. D.; Seibert, K.; Isakson, P. C.; Stallings, W. C. Structural basis for selective inhibition of cyclooxygenase-2 by antiinflammatory agents. *Nature* **1996**, *384*, 644−648.

(6) Dempke, W.; Rie, C.; Grothey, A.; Schmoll, H.-J. Cyclooxygenase-2: a novel target for cancer chemotherapy. *J. Cancer Res. Clin. Oncol.* **2001**, *127*, 411−417.

(7) Steinbach, G.; Lynch, P. M.; Phillips, R. K. S. The effect of celecoxib, a cyclooxygenase-2 inhibitor, in familial adenomatous polyposis. *N. Engl. J. Med.* **2000**, *29*, 1946−1952.

(8) Stewart, W. F.; Kawas, C.; Corrada, M.; Metter, E. J. Risk of Alzheimer's disease and duration of NSAID use. *Neurology* **1997**, *48*, 626−632.

(9) Pasinetti, G. M. Cyclooxygenase and Alzheimer's disease: implications for preventive initiatives to slow the progression of clinical dementia. *Arch. Gerontology Geriatrics* **2001**, *33*, 13−28.

(10) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit-rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1961**, *42*, 5100−5109.

(11) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.

(12) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.

(13) Chavatte, P.; Yous, S.; Marot, C.; Baurin, N.; Lesieur, D. Three-Dimensional Quantitative Structure−Activity Relationships of Cyclooxygenase-2 (COX-2) Inhibitors: A Comparative Molecular Field Analysis, *J. Med. Chem.* **2001**, *44*, 3223−3230.

(14) Kauffman, G. W.; Jurs, P. C. QSAR and k-Nearest Neighbor Classification Analysis of Selective Cyclooxygenase-2 Inhibitors Using Topologically-Based Numerical Descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1553−1560.

(15) Gierse, J. K.; Hauser, S. D.; Creely, D. P.; Koboldt, C.; Rangwala, S. H.; Isakson, P. C.; Seibert, K. Expression and Selective Inhibition of the Constitutive and Inducible Forms of Human Cyclo-oxygenase. *Biochem. J.* **1995**, *305*, 479−484.

(16) Huang, H. C.; Li, J. J.; Garland, D. J.; Chamberlain, T. S.; Reinhart, E. J.; Manning, R. E.; Seibert, K.; Koboldt, C. M.; Gregory, S. A.; Anderson, G. D.; Veenhuizen, A. W.; Zhang, Y.; Perkins, W. E.; Burton, E. G.; Cogburn, J. N.; Isakson, P. C.; Reitz, D. B. Diarylspiro-[2.4]heptenes as Orally Active, Highly Selective Cyclooxygenase-2 inhibitors: Synthesis and Structure−Activity Relationships. *J. Med. Chem.* **1996**, *39*, 253−266.

(17) Khanna, I. K.; Weier, R. M.; Yu, Y.; Collins, P. W.; Miyashiro, J. M.; Koboldt, C. M.; Veenhuizen, A. W.; Currie, J. L.; Seibert, K.; Isakson, P. C. 1,2-Diarylpyrroles as Potent and Selective Inhibitors of Cyclooxygenase-2. *J. Med. Chem.* **1997**, *40*, 1619−1633.

(18) Khanna, I. K.; Weier, R. M.; Yu, Y.; Xu, X. D.; Koszyk, F. J.; Collins, P. W.; Koboldt, C. M.; Veenhuizen, A. W.; Perkins, W. E.; Casler, J. J.; Masferrer, J. L.; Zhang, Y. Y.; Gregory, S. A.; Seibert, K.; Isakson, P. C. 1,2-Diarylimidazoles as Potent, Cyclooxygenase-2 Selective, and Orally Active Antiinflammatory Agents. *J. Med. Chem.* **1997**, *40*, 1634−1647.

(19) Khanna, I. K.; Yu, Y.; Huff, R. M.; Weier, R. M.; Xu, X.; Koszyk, F. J.; Collins, P. W.; Cogburn, J. N.; Isakson, P. C.; Koboldt, C. M.; Masferrer, J. L.; Perkins, W. E.; Seibert, K.; Veenhuizen, A. W.; Yuan, J.; Yang, D. C.; Zhang, Y. Y. Selective cyclooxygenase-2 inhibitors: heteroaryl modified 1,2-diarylimidazoles are potent, orally active antiinflammatory agents. *J. Med. Chem.* **2000**, *43*, 3168−3185.

(20) Reitz, D. B.; Li, J. J.; Norton, M. B.; Reinhart, E. J.; Collins, J. T.; Anderson, G. D.; Gregory, S. A.; Koboldt, C. M.; Perkins, W. E.; Seibert, K.; Isakson, P. C. Selective Cyclooxygenase Inhibitors: Novel

1,2-Diarylcyclopentenes are Potent and Orally Active COX-2 Inhibitors. *J. Med. Chem.* **1994**, *37*, 3878−3881.

(21) Li, J. J.; Anderson, G. D.; Burton, E. G.; Cogburn, J. N.; Collins, J. T.; Garland, D. J.; Gregory, S. A.; Huang, H. C., Isakson, P. C.; Koboldt, C. M.; Logusch, E. W.; Norton, M. B.; Perkins, W. E.; Reinhart, E. J.; Seibert, K.; Veenhuizen, A. W.; Zhang, Y.; Reitz, D. B. 1,2-Diarylcyclopentenes as Selective Cyclooxygenase-2 Inhibitors and Orally Active Antiinflammatory Agents. *J. Med. Chem.* **1995**, *38*, 4570−4578.

(22) Li, J. J.; Norton, M. B.; Reinhart, E. J.; Anderson, G. D.; Gregory, S. A.; Isakson, P. C.; Koboldt, C. M.; Masferrer, J. L.; Perkins, W. E.; Seibert, K.; Zhang, Y.; Zweifel, B. S.; Reitz, D. B. Novel Terphenyls as Selective Cyclooxygenase-2 Inhibitors and Orally Active Antiinflammatory Agents. *J. Med. Chem.* **1996**, *39*, 1846−1856.

(23) Penning, T. D.; Talley, J. J.; Bertenshaw, S. R.; Carter, J. S.; Collins, P. W.; Docter, S.; Graneto, M. J.; Lee, L. F.; Malecha, J. W.; Miyashiro, J. M.; Rogers, R. S.; Rogier, D. J.; Yu, S. S.; Anderson, G. D.; Burton, E. G.; Cogburn, J. N.; Gregory, S. A.; Koboldt, C. M.; Perkins, W. E.; Seibert, K.; Veenhuizen, A. W.; Zhang, Y. Y.; Isakson, P. C. Synthesis and Biological Evaluation of the 1,5-Diarylpyrazole Class of Cyclooxygenase-2 Inhibitors: Identification of 4-[5-(4-Methylphenyl)-3-(trifluoromethyl)-1H-pyrazol-1-yl]benzenesulfonamide (SC-58635, Celecoxib). *J. Med. Chem.* **1997**, *40*, 1347−1365.

(24) Talley J. J.; Brown, D. L.; Carter, J. S.; Graneto, M. J.; Koboldt, C. M.; Masferrer, J. L.; Perkins, W. E.; Rogers, R. S.; Shaffer, A. F.; Zhang, Y. Y.; Zweifel, B. S.; Seibert, K. 4-[5-Methyl-3-phenylisoxazol-4-yl]-benzenesulfonamide, Valdecoxib: a Potent and Selective Inhibitor of COX-2. *J. Med. Chem.* **2000**, *43*, 775−777.

(25) MOE 2000.02, Chemical Computing Group Inc., Montreal, H3A 2R7 Canada, http://www.chemcomp.com.

(26) Halgren, T. A. MMFF94 The Merck force field. *J. Comput. Chem.* **1996**, *17*, 490−755.

(27) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.

(28) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity−A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3228.

(29) Maybridge Chemical Company Ltd., Cornwall, PL34 OHW England. http://www.maybridge.co.uk.

(30) Marengo, E.; Todeschini, R. A new algorithm for optimal, distance-based experimental design. *Chemom. Intell. Lab. Syst.* **1992**, *16*, 37−44.

(31) Neal, R. M. In *Bayesian Learning for Neural Networks*, Lecture Notes in Statistics No. 118; Bickel, P., Diggle, P., Fienberg, S., Krickeberg, K., Olkin, I., Wermuth, N., Zeger, S., Eds.; Springer-Verlag: New York, 1996.

(32) MacKay, D. J. C. Bayesian methods for back-propagation networks. In *Models of Neural Networks III*; Domany, E., van Hemmen, J. L., Schulten, K., Eds.; Springer: New York, 1994; pp 211−254.

(33) Burden, F. R.; Ford, M. G.; Whitley, D. C.; Winkler, D. A. Use of Automatic Relevance Determination in QSAR Studies Using Bayesian Neural Networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1423−1430.

(34) Neal, R. M. Software for flexible Bayesian modeling (version of 2000-08-21), available at http://www.cs.toronto.edu/∼radford.

(35) Zheng, W.; Tropsha, A. Novel variable selection quantitative structure−property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185−194.

(36) Hoffman, B.; Cho, S. J.; Zheng, W.; Wyrick, S.; Nichols, D. E.; Mailman, R. B.; Tropsha, A. Quantitative structure−activity relationship modeling of dopamine D(1) antagonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and K nearest neighbor methods. *J. Med. Chem.* **1999**, *42*, 3217−3226.

(37) Zheng, W.; Cho, S. J.; Tropsha, A. Rational Combinatorial Design. 1. Focus 2-D: A New Approach to the Design of Targeted Combinatorial Chemical Libraries. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 251−258.

(38) Wall, M. GAlib: A C++ Genetic Algorithm Library. Available at http://lancet.mit.edu/ga.

(39) Hellberg, S.; Sjöstrom, M.; Skagerberg, B.; Wold, S. Peptide QSAR, a multivariate approach. *J. Med. Chem.* **1987**, *30*, 1126−1135.

(40) Bush, B.; Nachbar, R. B. Sample-distance partial least squares: PLS optimized for many variables, with application to CoMFA. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 587−619.

(41) Kohonen, T.; Kangas, J.; Laaksonen, J.; Torkkola, K. LVQ_PAK: a program package for the correct Application of Learning Vector Quantization algorithms. *Proceedings of the international joint conference on Neural Networks*; Baltimore, 1992; pp 725−730.

(42) Kohonen, T.; Hynninen, J.; Kangas, J.; Laaksonen, J. *SOM_PAK: The Self-Organizing Map Program Package*; Technical Report A31; Helsinki University of Technology, Laboratory of Computer and Information Science, FIN-02150 Espoo, Finland, 1996. Available at http://www.cis.hut.fi/research/som-research/nnrc-programs.shtml.

(43) Kohonen, T. Self-Organized Formation of Topologically Correct Feature Maps. *Biol. Cybern.* **1982**, *43*, 59−69.

(44) Kohonen, T.; Hynninen, J.; Kangas, J.; Laaksonen, J.; Torkkola, K. *LVQ_PAK: The Learning Vector Quantization Program Package*; Technical Report A30; Helsinki University of Technology, Laboratory of Computer and Information Science, FIN-02150 Espoo, Finland, 1996. Available at http://www.cis.hut.fi/research/som-research/nnrc-programs.shtml.

(45) Valleron, A.-J. In *Introduction à la biostatistique*; Masson: Paris, 1998.

(46) Elder, J. F.; Pregibon, D. A statistical perspective on knowledge discovery in databases. In *Advances in Knowledge Discovery and Data Mining*; Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., Eds.; The AAAI Press: U.S.A., 1996.

(47) Voigt, J. H.; Bienfait, B.; Wang, S.; Nicklaus, M. C. Comparison of the NCI Open Database with Seven Large Chemical Structural Databases. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 702−712.

(48) http://cactus.nci.nih.gov/ncidb2/download.html. The file NCI_aug00_2D.sdz was downloaded in November 2001.

(49) Milne, G. W. A.; Nicklaus, M. C.; Driscoll, J. S.; Wang, S.; Zaharevitz, D. National Cancer Institute Drug Information System 3D Database. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1219−1224.

CI0341565