Journal of
Medicinal Chemistry

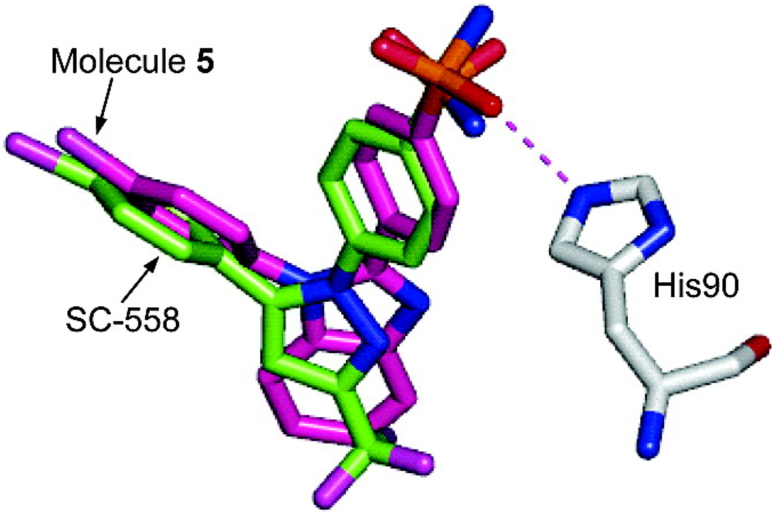
Article

Extraction and Visualization of Potential Pharmacophore Points Using Support Vector Machines: Application to Ligand-Based Virtual Screening for COX-2 Inhibitors

Lutz Franke, Evgeny Byvatov, Oliver Werz, Dieter Steinhilber, Petra Schneider, and Gisbert Schneider

J. Med. Chem., **2005**, 48 (22), 6997-7004• DOI: 10.1021/jm050619h • Publication Date (Web): 01 October 2005

Downloaded from http://pubs.acs.org on March 29, 2009



Extraction and Visualization of Potential Pharmacophore Points Using Support Vector Machines: Application to Ligand-Based Virtual Screening for COX-2 Inhibitors

Lutz Franke,[†,||] Evgeny Byvatov,[†,||] Oliver Werz,[‡] Dieter Steinhilber,[‡] Petra Schneider,[§] and Gisbert Schneider*,[†]

Institut für Organische Chemie und Chemische Biologie and Institut für Pharmazeutische Chemie, Johann Wolfgang Goethe-Universität, Marie-Curie-Strasse 9, D-60439 Frankfurt, Germany, and Schneider Consulting GbR, George-C.-Marshall Ring 33, D-61440 Oberursel, Germany

Received June 30, 2005

Support vector machines (SVM) were trained to predict cyclooxygenase 2 (COX-2) and thrombin inhibitors. The classifiers were obtained using sets of known COX-2 and thrombin inhibitors as “positive examples” and a large collection of screening compounds as “negative examples”. Molecules were encoded by topological pharmacophore-point triangles. In retrospective virtual screening, 50−90% of the known active compounds were listed within the first 0.1% of the ranked database. To check the validity of the constructed classifiers, we developed a method for feature extraction and visualization using SVM. As a result, potential pharmacophore points were weighted according to their importance for COX-2 and thrombin inhibition. Known thrombin and COX-2 pharmacophore points were correctly recognized by the machine learning system. In a prospective virtual screening study, several potential COX-2 inhibitors were predicted and tested in a cellular activity assay. A benzimidazole derivative exhibited significant inhibitory activity with an IC_{50} of 0.2 μM, which is better than Celecoxib in our assay. It was demonstrated that the SVM machine-learning method can be used in virtual screening and be analyzed in a human-interpretable way that results in a set of rules for designing novel molecules.

Introduction

A pharmacophore is defined as “the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response” (IUPAC Glossary of Terms Used in Medicinal Chemistry, URL: http://www.chem.qmul.ac.uk/iupac/medchem/). Pharmacophoric descriptors that are used to define a pharmacophore can be used in different ways in drug design programs: (i) as a query tool in virtual screening to identify potential new compounds from databases of “drug-like” molecules with patentable structures that are different from those already discovered; (ii) to predict the activities of a set of new compounds yet to be synthesized; (iii) to help understand the possible mechanism of action; or (iv) to extract potential privileged (sub)structures.[1−3] Currently several algorithms are known that construct pharmacophore models from a set of available active compounds employing potential pharmacophore points (PPP), e.g., CATALYST[4] and DISCO,[5] ligand−receptor interaction patterns derived from protein-structure,[6] or field-based approaches.[7,8] The quality of the extracted models usually relies on the quality of the initial three-dimensional (3D) ligand alignment. This influence of the initial alignment on the quality of the resulting pharmacophore model can be modulated by considering multiple ligand conformations (ensembles) and applying “fuzzy” pharmacophore point definitions.[6,9,10] Here we present a complementary method for PPP identification that is grounded on a topological three-point pharmacophore (3PP) concept.[11] The motivation was to avoid a strict dependency of the pharmacophore model on a 3D alignment. Each molecule was represented as a binary vector, where each feature corresponds to the presence or absence of a particular pharmacophore triangle.[11] These alignment-free feature vectors were used for construction of a classifier predicting molecules to have certain biological activity. Subsequent visualization of features with respect to their contribution to the model allowed us to find patterns of potential pharmacophore points. For the present study, Support vector machines (SVM) were used for both classification and feature extraction.[12−17] Molecules were represented by fingerprints that contained ~10^4 potential 3PP triangles, and we expected SVM to efficiently discriminate between important and unimportant features. This approach is motivated by the fact that SVM classifiers have been shown to be well-suited for first-pass virtual screening purposes.[18−20] Two test cases were selected to evaluate our new approach, namely the development of SVM classifiers for cyclooxygenase 2 (COX-2) and thrombin (Factor IIa) inhibitors.

Methods

Data Sets and Feature Fingerprint. Two subsets of the COBRA (version 2.11) collection of pharmacologically active reference compounds were used for SVM

[*] To whom correspondence should be addressed. Tel:+49 (0)69 798 29821. Fax:+49 (0)69 798 29826. E-mail: gisbert.schneider@modlab.de.

[†] Institut für Organische Chemie und Chemische Biologie, Johann Wolfgang Goethe-Universität.

[‡] Institut für Pharmazeutische Chemie, Johann Wolfgang Goethe-Universität.

[§] Schneider Consulting GbR.

[||] These authors contributed equally to this study.

10.1021/jm050619h CCC: $30.25 © 2005 American Chemical Society
Published on Web 10/01/2005

training:[21] 188 thrombin and 94 COX-2 inhibitors. We used these subsets as a reference for ranking ~2.7 million substances that are commercially available from different vendors. A similar screening library has been compiled by Irwin and Shoichet recently.[22] Each compound was represented by a 3PP fingerprint using the fingerprint generator available from the software suite MOE (version 2004.05; MOE, Molecular Operating Environment, Chemical Computing Group Inc., Montreal, Canada; URL: www.chemcomp.com). The individual 3PP feature is a triangle. We considered all possible triangles with their vertexes located at the atom centers of a molecule. Presence or absence of a certain triangle defines the "on" (i.e., bit is set) or "off" state of the corresponding bit in the fingerprint. We distinguished triangles by the type of atom at vertexes and by the length of their edges. The vertex can be either donor (D), acceptor (A), polar (P), donor and planar (D=), acceptor and planar (A=), hydrophobic (H), and hydrophobic and planar (H=) as defined by the rule-based atom-typer implemented in MOE which follows the PATTY atom-type definition.[23] Lengths of the edges were calculated along the molecular graph, so no estimation of the 3D structure of molecule was performed. The graph distance was defined as the number of bonds in the shortest path between the atoms in the chemical graph. Distances were binned into six categories {1,2,3,4,5−9,10-} yielding higher resolution for smaller distances and less for larger distances. Distances greater than 10 bonds were pooled. As a result, a fingerprint is the set of all tuples of the form (a_1, a_2, a_3, d_1, d_2, d_3), where a_1, a_2, and a_3 are atom-types, and d_1, d_2, and d_3 are graph distances between the respective atoms.

Support Vector Machine. The SVM constructs a surface in the n-dimensional space that separates active from inactive compounds.[24] Here, n is the number of 3PP that were used to describe a molecule. Prior to construction of the separating surface the data is mapped to a very high-dimensional space, where the separating surface is found in a form of a hyperplane. This hyperplane is then mapped backed to the original space.[25] The result of SVM training can be given by the following equation (eq 1).

$$f(\mathbf{x}) = \sum_i \alpha_i K(\mathbf{x},\mathbf{x}_i^{sv}) + b, \text{ where } K(\mathbf{x},\mathbf{y}) = ((\mathbf{x}\cdot\mathbf{y})s + 1)^5 \quad (1)$$

Here, $f(\mathbf{x})$ gives the prediction of the molecule to belong to the class "active", i.e., the greater the value of f the higher is the predicted probability to be active. $\mathbf{x}$ and $\mathbf{y}$ are molecular fingerprint vectors, $\mathbf{x}^{sv}$ are support vectors, i.e., molecular fingerprints that define the exact shape of the separating hyperplane. The kernel function K defines the complexity of the surface that will be constructed. Different standard kernels can be used during SVM training.[24] The influence of a particular Kernel on classification accuracy remains a matter of debate. Our own previous virtual screening studies using SVM indicated that the choice of a particular Kernel is of limited importance for the overall virtual screening process.[15,19] Other studies come to different conclusions.[16] In the present work we used a fifth-order polynomial for all SVM models. This Kernel represents a compromise between complexity and computational efficiency. For the purpose of this study, namely to provide proof-of-principle for the feature selection approach, the outcome of SVM training was not compared for different Kernel functions. Kernel parameter s was optimized to achieve better ranking of compounds as described.[18,19]

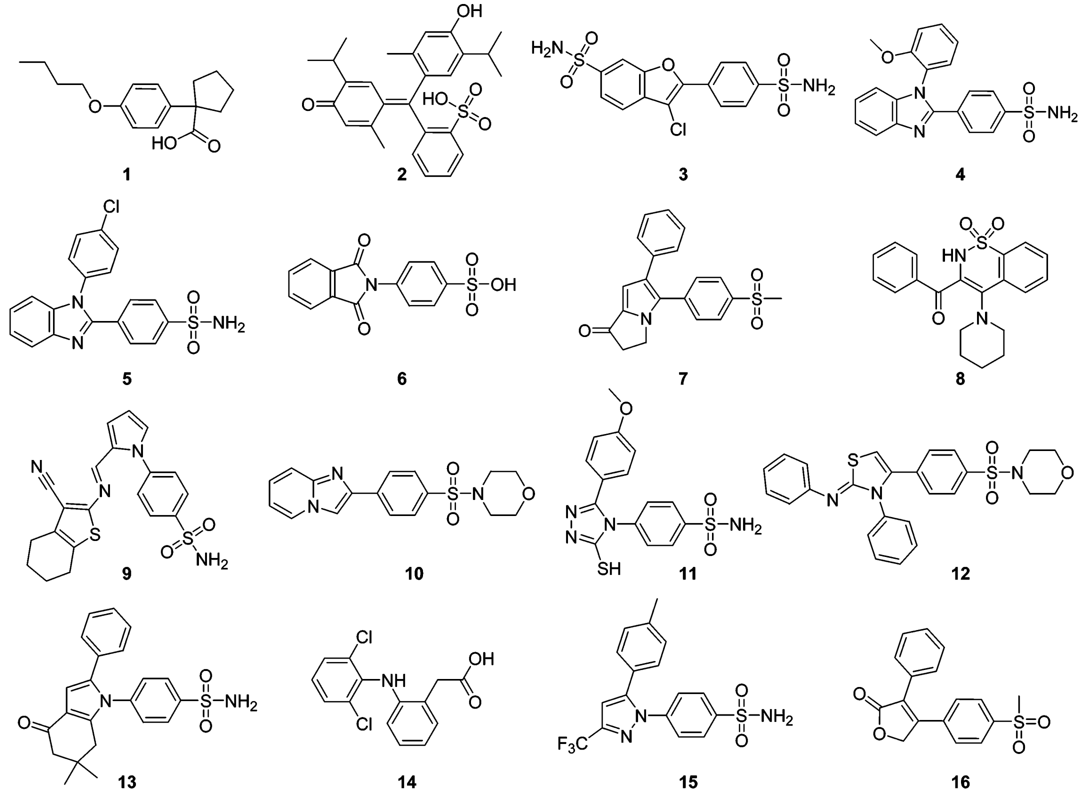
For database screening, we sorted all available compounds with respect to predicted f. The sum in eq 1 is over support vectors, they are part of the training set. Note that the ranking function depends only on the support vectors. Parameters α_i and b were determined during SVM training as described.[26] For constructing SVM models we used the SVM-light package.[27]

Training of the SVM, "Active Learning" Optimization, and Compound Selection. During SVM training we tried to optimize the percentage of active compounds found within the top 0.1% of the ranked screening database. To achieve this, we used a standard 4-fold cross-validation procedure:[28] The reference set of active compounds was divided into four equal parts. Each part in turn was mixed with the ~2.7 million screening compounds. The remainder of the set of known actives and the set of molecules to be virtually screened were assigned "class" (known inhibitors) and "nonclass" (all other molecules) labels for SVM training. Note that during SVM training "mixed" active compounds were marked as "nonclass" molecules. After training, the nonclass compounds were sorted with respect to the f value computed by the SVM classifier. Molecules with higher f values are expected to be similar to the active compounds. The parameters of SVM were optimized to yield a maximum number of "mixed" (i.e. reference molecules that were added to the pool of screening compounds) active molecules within the top 0.1% of the ranked data.

Actual SVM training was performed in two steps to reduce computing time and focus on the class/nonclass boundary in the descriptor space: instead of using all nonclass molecules of a training set, first an SVM was trained with a randomly selected subset containing only 10^5 compounds from the nonclass set. We are well aware that in this case the region near the active compounds might be insufficiently sampled. We therefore employed the "active learning" approach to focus on "relevant" portions of descriptor space:[28,29] After obtaining the first ranked list of the compounds the SVM training procedure was repeated, now with a sample set consisting of the top-ranking 10^5 compounds. By this two-step training process, a more fine-tuned SVM classifier focusing on the class/nonclass boundary was obtained. This concept has been shown to be useful in related virtual screening applications recently.[15,20] Predictions were made for the whole data set of ~2.7 million molecules. Since parts (2×10^5 compounds = 7.4%) of the nonclass data were used for training, the predictions do not represent true validation results.

The two-step active learning procedure was followed by final SVM training, which was performed with all available actives ("class") and all molecules with unknown activity as "nonclass" samples. Optimized SVM parameters were used. The resulting ranking of these compounds by the trained SVM was used to cherry-pick molecules for in vitro activity testing: Beginning at rank

Chart 1. Compounds **1**−**13** Were Cherry-Picked from the Virtual Screening Results and Tested for COX-2 Inhibition. Reference Compounds Diclofenac **14**, Celecoxib **15**, and Rofecoxib **16**



1 of the final SVM-ranked list of commercially available compounds, 13 molecules were cherry-picked. We excluded several compounds that contain certain reactive groups and potentially insoluble molecules by visual inspection. At this stage, only compounds available from Specs (Delft, The Netherlands; www.specs.net) were considered (Chart 1), since the aim of our study was not to find as many novel COX-2 inhibitors as possible, but to provide proof-of-principle for our approach. In this prospective study, the complete set of ∼2.7 million compounds with unknown activity was employed to span a wide chemical space for SVM training.

Pharmacophore Point Visualization. Potential pharmacophore points of the inhibitors were visualized by highlighting atoms that contribute to the most important features. The importance R_i of each 3PP feature was calculated based on the change of SVM prediction for a molecule when this feature is removed (eq 2).

$$R_i = f(\mathbf{x}(F_i = 1)) - f(\mathbf{x}(F_i = 0)) \qquad (2)$$

where $\mathbf{x}$ is a fingerprint representation of a molecule with presence or absence of feature F_i. Each atom contributing to feature F_i receives the weight R_i. It is reasonable to assume that the importance of atoms in a 3PP differs. To take this into account the importance of every atom in the reference set of actives was estimated (Figure 1). The individual weight w of an atom was estimated as the average weight of all 3PP triangles that contain this atom as a vertex. Averaging was done twice, first over the triangles of each molecule (Figure 1d) and finally over the whole set of actives.

Visualization contrast was enhanced by diminishing the atom weights $\mathbf{w}$ in every 3PP except for the most important one by w^n. Prior to diminishing the weight of every feature, weights were normalized, so that the maximum w is equal to 1. We choose $n = 10$ empirically, so that the weight of the most important atom equals 1 and all other weights diminished.

Docking of COX2 Inhibitors. For docking of compounds into the COX-2 active site cavity MOE software was used. The complex of COX-2 with a selective inhibitor SC-558 (PDB-identifier: 1CX2) served as reference. Only one of the four identical domains of the COX-2 complex was considered. Prior to docking hydrogen atoms were added to the protein complex, and its structure was energy minimized keeping positions of all atoms fixed except for the added hydrogen atoms. Partial charges of the atoms were calculated using MMFF estimation.[30] Docking was performed using MOE molecular dynamics approximation and Tabu-search as described.[31] The results were evaluated by

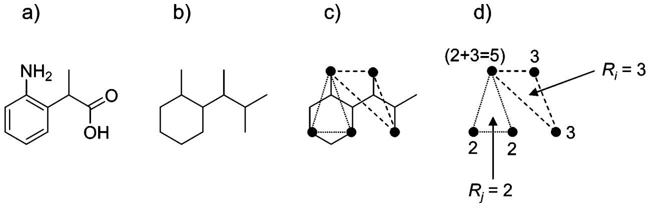


Figure 1. Calculation of atom weights for feature visualization. The two-dimensional molecular structure (a) is converted to the hydrogen-depleted molecular graph representation (b). Then topological 3PP triangles are assigned (the length of each edge is calculated as the number of bonds in the molecular graph connecting the two vertexes along the shortest path) (c), and the importance R of each triangle is determined by eq 2. Individual atoms are weighted proportional to the sum of the R_i values of contributing 3PP features (d).

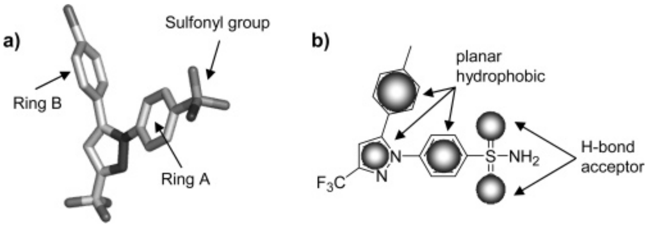


Figure 2. (A) Crystal structure conformation of the inhibitor SC-558 bound to COX-2 (PDB identifier: 1DWC). According to Palomer et al., essential interactions for specific COX-2 inhibitors are the aromatic rings A and B and the sulfonyl group.[33] (B) Potential pharmacophore points (shaded circles) identified by SVM for the structurally similar Celecoxib **15**. The sizes of the potential pharmacophore points reflect their relative contribution (weight) to the SVM classifier.

comparison with the binding mode of the reference inhibitor SC-558.

Results and Discussion

SVM Training and Feature Visualization. SVM classifiers were trained to predict thrombin and COX-2 inhibitors. The accuracy of the predictions was assessed by retrospective database screening. In the case of the COX-2 classifier, 81 ± 6% of the 22 test compounds were retrieved within the first 0.1% of the ranked database in a 4-fold cross-validation study (a priori probability: 0.1%). The retrieval of thrombin ligands was less accurate, yielding 55 ± 14% of the 46 test compounds from the first 0.45% of the ranked database (a priori probability: 0.45%). The small standard deviations in both cases indicate robust prediction models. With further optimization by active learning we yielded 81 ± 6% of the test compounds in 0.0031% of the ranked database for COX-2 inhibitors and 55 ± 14% of the test compounds from the first 0.083% of the ranked database for thrombin ligands. This difference in performance might be explained not only by differences of the two reference sets and SVM classifier shortcomings but also by the structural diversity of chemotypes that are present in the screening database.[32] Overall, we concluded that the two SVM classifiers might be useful for generating focused libraries with significant enrichment of actives compared to a random selection of compounds, as indicated by the low a priori probabilities for finding an active molecule among a random selection of molecules from the screening database.

To further validate the constructed SVM models we visualized PPPs that were predicted as relevant. For COX-2 inhibitors a well-known pharmacophoric pattern was highlighted: a constellation of aromatic rings with a sulfonamide group attached to one of them (Figure 2a).[33] Interestingly, it has been reported that the aryl moiety B can be replaced by alkoxy groups, still retaining COX-2 potency and selectivity (for review, see ref 34). For interpretation of the SVM feature extraction results, this should be kept in mind. For our SVM-based PPP constellation it is important to note that for the example of Celecoxib the oxygen atoms of the sulfonamide were marked as important for COX-2 inhibition and not the amino group (Figure 2b). Sulfonyl and sulfonamide groups are present in many specific COX-2 inhibitors.[33,34] They are known to interact with Arg-513 in the hydrophilic side-pocket of the COX-2 active site.[35] This confirms that the SVM was able to extract relevant pharmacophore points from the set of all potential pharmacophore points that are present in a molecule. The predicted "planar hydrophobic" PPP of the pyrazole in Celecoxib (Figure 2b) should be regarded as an artifact arising from a bias toward diaryl heterocycles in the training data. There are several examples of selective COX-2 inhibitors lacking this predicted PPP, e.g. 1,2 diaryl-stilbenes or −alkenes.[34,36−39] This finding shows that although the feature extraction and visualization technique may be suited for finding relevant features, it reflects the diversity of chemotypes provided for SVM training. It is evident that fully generalizing features will not be found if biased training data are used.

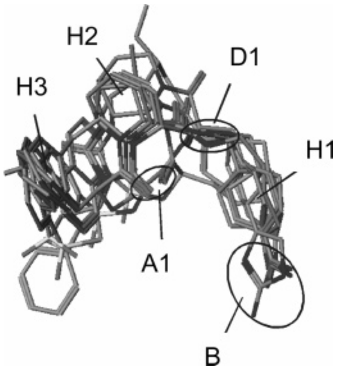


Figure 3. Three-dimensional alignment of thrombin inhibitors based on PDB structures 1C4V, 1D4P, 1D6W, 1D9I, 1DWD, 1FPC, and 1TOM (adapted from ref 10). The molecules were aligned by superposition of their appropriate protein structures. Essential interaction points according to Patel et al. are indicated. B is a basic interaction, H1, H2, and H3 are hydrophobic interactions, A1 is a hydrogen-bond acceptor, and D1 is a hydrogen-bond donor.

In contrast to the comparably simple pharmacophore of COX-2 inhibitors, thrombin inhibitors represent more complex molecules, and the respective pharmacophore contains more interaction points (Figure 3).[40,41] According to Patel et al., the major interactions are B, H1, H2, and H3, where B is a basic interaction which interacts with the carboxylic group of an aspartate.[42] H1, H2, and H3 are hydrophobic interactions; less conserved interactions are D1 and A1, where D1 is a hydrogen-bond donor and A1 is a hydrogen-bond acceptor.[42] Figure 4 shows suggested thrombin pharmacophore points extracted by the corresponding SVM classifier for one of the compounds that were selected from the screening database. We can find the basic guanidinium moiety potentially binding to Asp189 at the bottom of the specificity pocket P1 of thrombin.[43] It is interesting to note that not all atoms of the arginine side-chain are considered important by SVM. This is exactly what one would expect, as several arginine-analogues have been identified that bind in the same or a similar mode to the P1 pocket.[44] This result was probably achieved by selecting 3PP feature triangles with amines at the vertexes and relatively long edges which correspond to the arginine side chain. H-Bonding to the Gly216 backbone was also accurately predicted as one of the crucial interaction sites (Figure 4d). The hydrophobic interactions H1, H2, and H3 in the model of Patel et al. (Figure 3) were not appropriately recognized by the SVM, although several known inhibitors containing these interaction points

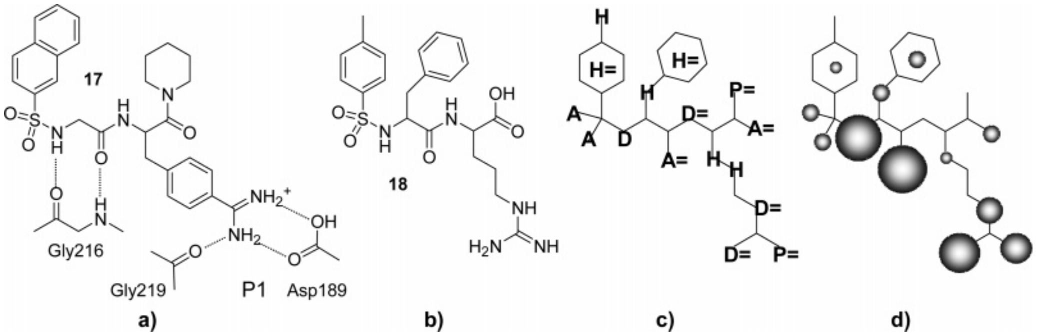


Figure 4. (a) Most important interactions between NAPAP[45] **17** and thrombin; (b) structure of a NAPAP-like compound **18** that was predicted to be a potential thrombin inhibitor by the SVM classifier; (c) with all potential pharmacophore points, and (d) the corresponding weights assigned by the SVM feature extraction procedure. The crucial pharmacophore pattern of the NAPAP–thrombin complex was automatically identified by the feature extraction method.

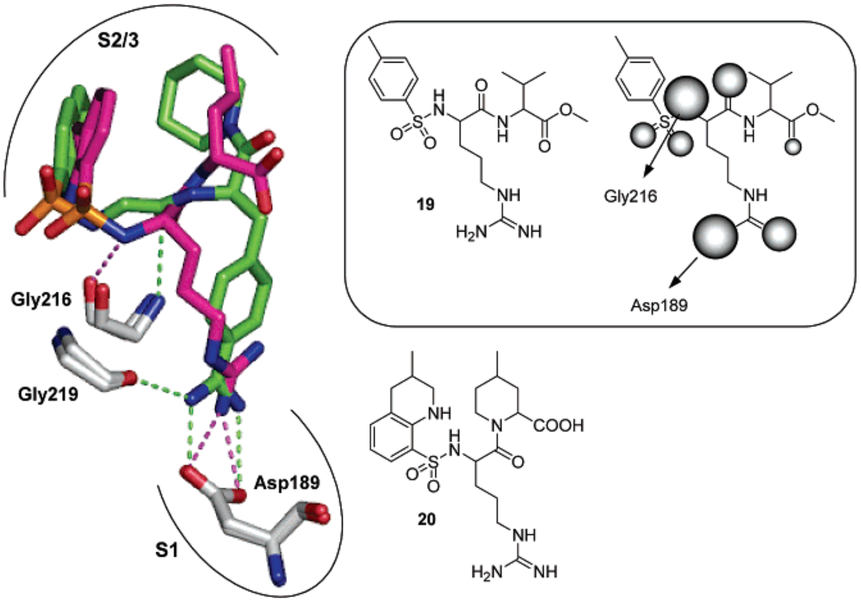


Figure 5. Complex of NAPAP **17** (green; PDB identifier: 1DWD) and Argatroban **20** (magenta; PDB identifier: 1DWC) with thrombin. Structures were superimposed according to their C_α coordinates. Hydrogen bonds are drawn as dotted lines, approximate locations of the S1, S2/3 active site pockets are indicated. Molecule **19** represents a predicted thrombin inhibitor. Pharmacophore features are highlighted that were considered "important" by the SVM classifier. "Importance" is indicated by the size of the circles.

were present in the training data. This observation reveals a limitation of our feature extraction method.

An interesting property of the SVM model for thrombin is illustrated by analyzing another compound that was predicted to be a potential thrombin inhibitor. Structure **19** contains a pattern of potential interaction points that might correspond to a different binding mode than that of NAPAP-inhibitors. The binding modes of NAPAP[45] **17** and argatroban[46] **20** are shown in Figure 5. It seems reasonable to assume that compound **19** adopts an Argatroban-like binding mode. The guanidinium group has the potential to form hydrogen-bonds with Asp189; and binding to Gly216 could be similar as for Argatroban. This assumption is supported by the SVM model which considers essential known pharmacophore points as most important (Figure 5). From this entirely theoretical consideration we concluded that both binding patterns, Argatroban-like and NAPAP-like, were contained in the SVM model resulting in an interpretable prediction of functional groups that might form key interactions with the target enzyme.

Virtual Screening for COX-2 Inhibitors. As a first practical validation of our prediction results and the validity of the SVM approach, we tested potential COX-2 inhibitors in an in vitro binding study. We chose this application because the SVM model of COX-2 inhibitors was more accurate than our thrombin classifier. Structures **1**–**13** were tested for COX-2 inhibition with Diclofenac **14**, Celecoxib **15**, and Rofecoxib **16** as positive references. We selected a set of compounds which contain known motifs of COX-2 ligands and potentially novel structural features in order to allow for "scaffold hopping" (Chart 1). Compounds **4**, **5**, and **7** exhibited an inhibitory effect in the activity assay. The most potent compound was **5** with $IC_{50} = 0.2 \pm 0.3$ μM. For molecule **4** we determined an IC_{50} of 8 ± 2 μM. Concentrations of **4** above 20 μM resulted in an increase of remaining enzyme activity. A similar observation was made for **7** above 5 μM concentration with approximately 60% remaining COX-2 activity. IC_{50} determination was thus not achievable for **7**. This observation might be a consequence of poor solubility. IC_{50} values of Diclofenac (5 ± 1 nM), Celecoxib (6 ± 3 μM), and Rofecoxib (15 ± 3 μM) were determined in our assay to obtain appropriate reference values. Compared to IC_{50} values (Diclofenac: 26 nM, Celecoxib: 6.8 μM,

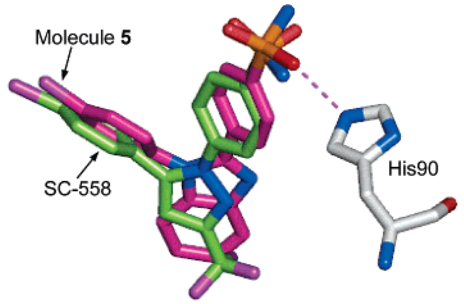
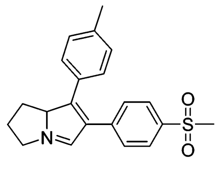
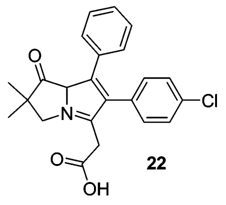


Figure 6. Superposition of the selective COX-2 inhibitor SC-558 (green; from PDB entry 1CX2) and molecule **5** (magenta), which was docked into the COX-2 active site. The two molecules have essentially the same binding conformation. A potential hydrogen-bond between **5** and His90 is indicated by the dotted line.

Rofecoxib: 25 μM)[47] that were obtained in a study by Kato et al. who used a similar assay system, we observed slightly lower IC_{50} values for all three reference compounds, with comparable relative inhibition, which can probably be attributed to individually different assay conditions. Molecule **5** exhibits an inhibitory effect on COX-2 that is even stronger than the effect of the two coxibs. Structures **4** and **5** contain a benzimidazole scaffold which, to the best of our knowledge, has not been employed for COX inhibitor development. The higher IC_{50} value of **4** might result from the methoxy group introducing steric hindrance, as deduced from extensive SAR studies performed for Celecoxib by Penning et al.[48] The two phenyl substituents of **5** are similar to the respective moieties of Celecoxib **15** and its potent derivatives.[48] Docking of **5** into the active site pocket of COX-2 essentially revealed a similar potential binding mode to SC-558,[49] a selective COX-2 inhibitor gaining selectivity over COX-1 interaction through interaction with residues (Ile517, Phe518, Gln192, and His(90)) forming a secondary pocket that is not present in COX-1 (Figure 6).[34,49] Note that due to the comparably low resolution of 3 Å of the SC-558-COX-2 cocrystal structure, the existence of a hydrogen-bond between **5** and His90 is speculative.

The pyrrolizine derivative **7** extends the class of known pyrrolizine-based COX-2 inhibitors. Molecule **21** has been described as a low nanomolar COX-2 inhibitor revealing micromolar 5-lipoxygenase (5-LO) inhibition.[50] Researchers at Merckle (Germany) described a family of 1-oxo-pyrrolizines **22** as potent COX/5-LO dual inhibitors.[51] The altered substitution pattern of **7** may thus be worthwhile testing for 5-LO inhibition, although it lacks possibly essential substructure elements of known 5-LO ligands, e.g. an aliphatic chain with a polar headgroup like the carboxylic acid function in **22**. Dual inhibition of COX-2 and 5-LO provides a new strategy to provide safer nonsteroidal antiinflammatory drugs, which is of interest to avoid potential side-effects of the coxib stuctural family.[52,53]


21


22

Although the benzimidazole **5** is less active than Diclofenac, but showed to have higher potency than Celecoxib and Rofecoxib, it might be suited for further optimization. A promising observation is the strong similarity between the PPP pattern of compound **5** and the interaction patetrns described for the known COX-2 inhibitors Celecoxib **15** and Rofecoxib **16**.[54,55] Certainly, selectivity for COX-2 over COX-1 inhibition should be investigated and considered for future designs. Again, for this task our SVM approach could be used to develop an additional COX-1 classifier and employed for (i) cherry-picking potential COX-2 selective inhibitors, and (ii) identification of enzyme subtype-specific pharmacophore points.

Conclusions

This study demonstrated that SVM can be employed for identification of promising screening candidates that exhibit significant biological activity. The extracted potential pharmacophore patterns coincided with known binding models of thrombin and COX-2 inhibitors. The SVM classifiers produced a quantitative ranking of substructure elements which can guide further hit and lead structure profiling. It was demonstrated that this machine-learning method is suitable for virtual screening and can be analyzed in a human-interpretable way that results in a set of rules for designing novel molecules. The method complements the suite of modeling techniques that have been employed for designing selective COX-2 inhibitors previously.[56] We employed the SVM method for our study since it offers particular advantages over other machine-learning approaches: [13,14,18,57] (i) the SVM class/nonclass boundary is constructed as the maximum margin classifier, i.e., it does not represent an arbitrary solution; (ii) it relies only on the so-called “support-vectors”, i.e., those molecules that define the classifier function (eq 1). This means, that in contrast to machine-learning methods employing an error function that is calculated from all data points (e.g., the mean-squared-error, mse), it is less affected by outliers and the overall shape of the data distribution; (iii) SVM training was shown to be more robust than, for example, training of radial-basis-function and multilayer-feedforward networks.[19,57] Still, most important to the particular task of molecular fingerprint weighting and potential pharmacophore-point visualization is the fact that SVM training can cope with high-dimensional molecular descriptors.[18,58] The feature-weighting procedure used in our study could be applicable to other “black box” prediction systems too, e.g. artificial neural networks predicting molecular properties. It complements similar techniques that are grounded on, e.g., genetic algorithms or ensemble weighting.[10,59] Since SVM training was not rigorously optimized in this study (e.g., choice of Kernel function, parameter optimization), we think that further optimization of classification accuracy might be possible. Together with an augmented set of reference compounds that are less biased toward individual chemotypes this will lead to further refined pharmacophore point hypotheses.

Experimental Section

COX-2 Assay. A COX-2 inhibition assay was performed to evaluate compound activity with Diclofenac **14**, Celecoxib **15**, and Rofecoxib **16** as positive controls.[52] Celecoxib **15** and Rofecoxib **16** samples were obtained from the Department of Clinical Pharmacology, Goethe-University of Frankfurt (Germany). Diclofenac was obtained from Sigma (Deisenhofen, Germany). The human monocytic cell line Mono Mac 6 was differentiated with transforming growth factor beta (TGFβ, 1 ng/mL) and calcitriol (50 nM) for 96 h as described.[60] Six hours prior to harvest, lipopolysaccharide (100 ng/mL) was added to induce COX-2 expression. Then, cells were harvested, washed twice, resuspended in PGC buffer (phosphate buffered saline at pH 7.4 containing 1 mg/mL glucose and 1 mM $CaCl_2$) (5×10^6 cells/ml), preincubated with the test compounds at the indicated concentrations for 15 min at 37°C, and then incubated with arachidonic acid (30 μM) for 15 min at 37 °C. The reaction was stopped on ice for 10 min. Cells were centrifuged (300*g*, 5 min, 4°C), and the amount of 6-Keto $PGF_{1\alpha}$ released was assessed by ELISA using a monoclonal antibody against 6-keto $PGF_{1\alpha}$ according to the protocol described by Yamamoto and co-workers.[61,62] For the ELISA, the monoclonal antibody (0.2 μg/200 μl) was coated to microtiter plates via a goat anti-mouse-IgG antibody. 6-Keto $PGF_{1\alpha}$ (15 μg) was linked to bacterial β-galactosidase (0.5 mg, Calbiochem), and the enzyme activity bound to the antibody was determined in an ELISA reader at OD550 nm (reference wavelength: 630 nm) using chlorophenol-red-β-D-galactopyranoside (CPRG, Roche Diagnostics GmbH, Germany) as substrate.

Calculation of IC_{50} Values. The raw ELISA readout was calibrated using 1 pg, 100 pg, 1 ng, 5 ng, 10 ng, 100 ng, and 1 μg 6-Keto $PGF_{1\alpha}$ ($N = 4$). Maximal enzyme activity was determined with 30 μM arachidonic acid. Background activity was determined without arachidonic acid. From the ELISA readouts the amounts of 6-Keto $PGF_{1\alpha}$ product was calculated. For IC_{50} value determination, 10 different concentrations of the test compounds were used ($N = 4$).

Acknowledgment. We wish to thank Dana Albert and Ilse Zündorf for experimental support. This research was supported by the Beilstein-Institut zur Förderung der Chemischen Wissenschaften, Frankfurt am Main, Germany. E. Byvatov conceived of the feature visualization method and performed SVM model development.

References

(1) Dror, O.; Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Predicting molecular interactions in silico: I. A guide to pharmacophore identification and its applications to drug design. *Curr. Med. Chem.* **2004**, *11*, 71–90.

(2) S. Pickett, The biophore concept. In *Protein–Ligand Interactions*; Böhm, H.-J., Schneider, G., Eds.); Wiley-VCH: Weinheim, 2003; pp 73–105.

(3) Guner, O. F. History and evolution of the pharmacophore concept in computer-aided drug design. *Curr. Top. Med. Chem*. **2002**, *2*, 1321–1332.

(4) Kurogy, Y.; Guner, O. F. Pharmacophore modeling and three-dimensional database searching for drug design using Catalyst. *Curr. Med. Chem.* **2001**, *8*, 1035–1055.

(5) Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J.; Lico, I.; Pavlik, P. A. A fast approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 83–102.

(6) Wolber, G.; Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model*. **2005**, *45*, 160–169.

(7) Cruciani, G.; Watson, K. A. Comparative molecular field analysis using GRID force-field and GOLPE variable selection methods in a study of inhibitors of glycogen phosphorylase b. *J. Med. Chem.* **1994**, *37*, 7, 2589–2601.

(8) Sippl, W. Receptor-based 3D QSAR analysis of estrogen receptor ligands – merging the accuracy of receptor-based alignments with the computational efficiency of ligand-based methods. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 559–572.

(9) Horvath, D.; Jeandenans, C. Neighborhood behavior of in silico structural spaces with respect to in vitro activity spaces-a novel understanding of the molecular similarity principle in the context of multiple receptor binding profiles. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 680–690.

(10) Renner, S.; Schneider, G. Fuzzy pharmacophore models from molecular alignments for correlation-vector based virtual screening. *J. Med. Chem*. **2004**, *47*, 4653–4664.

(11) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical similarity using geometric atom pair descriptors. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 128–136.

(12) McGregor, M. J.; Muskal, S. M. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569–574.

(13) Vapnik, V., The Nature of Statistical Learning Theory. In Ed. 1995: Springer

(14) Byvatov, E.; Schneider, G. Applications of support vector machines in bioinformatics. *Appl. Bioinf.* **2003**, 2, 67–77.

(15) Byvatov, E.; Sasse, B. C.; Stark, H.; Schneider, G. From virtual to real screening for novel D3 dopamine receptor ligands. *ChemBioChem* **2005**, *6*, 997–999.

(16) Müller, K. R.; Ratsch, G.; Sonnenburg, S.; Mika, S.; Grimm, M.; Heinrich, N. Classifying ′drug-likeness' with Kernel-based learning methods. *J. Chem. Inf. Model*. **2005,** 45, 249–253.

(17) Zernov, V. V.; Balakin, K. V.; Ivaschenko, A. A.; Savchuk, N. P.; Pletnev, I. V. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 2048–2056.

(18) Byvatov, E.; Schneider, G. SVM-based feature selection for characterization of focused compound collections, *J. Chem. Inf. Comput. Sci.,* **2004**, *44*, 993–999.

(19) Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1882–1889.

(20) Warmuth, M. K.; Liao, J.; Ratsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active learning with support vector machines in the drug discovery process. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 667–673.

(21) Schneider, P.; Schneider, G. Collection of bioactive reference compounds for focused library design. *QSAR Comb. Sci.* **2003**, *22*, 713–718.

(22) Irwin, J. J.; Shoichet, B. K. ZINC- -a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–82.

(23) Sheridan, R. P.; Bush, B. L. PATTY: A programmable atom typer and language for automatic classification of atoms in molecular databases. *J. Chem. Inf. Comput. Sci*. **1993**, *33*, 756–762.

(24) Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press: Cambridge, 2000.

(25) Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discovery* **1998**, *2*, 121–167.

(26) Cortes, C.; Vapnik, V. Support-vector networks. *Machine Learning* **1995**, *20*, 273–297.

(27) Joachims, T. Making large-Scale SVM learning practical, in*: Advances in Kernel Methods – Support Vector Learning* (Schölkopf, B., Burges, C., Smola, A., Eds.), MIT-Press: Cambridge, MA, 1999; pp 41–56.

(28) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*; Wiley-Interscience: New York, 2000.

(29) Tong, S.; Koller, D. Support vector machine active learning with applications to text classification. In *Proceedings of the Seventeenth International Conference on Machine Learning*; Morgan Kaufmann: San Francisco, CA, 2000.

(30) Halgren, T. A., The Merck force field, *J. Comput. Chem.* **1996**, *17*, 490–512.

(31) Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D., Flexible docking using tabu search and an empirical estimate of binding affinity. *Proteins Struct. Funct. Genet.* **1998**, *33*, 367–382.

(32) Schneider, P.; Schneider, G. Navigation through chemical space: ligand-based library design of focused compound libraries. In *Chemogenomics in Drug Discovery*; Kubinyi, H., Müller, G., Eds.); Wiley-VCH: Weinheim, 2004; pp 341–376.

(33) Palomer, A.; Cabre, F.; Pascual, J.; Campos, J.; Trujillo, M. A.; Entrena, A.; Gallo, M. A.; Garcia, L.; Mauleon, D.; Espinosa, A. Identification of novel cyclooxygenase-2 selective inhibitors using pharmacophore models. *J. Med. Chem.* **2002**, *45*, 1402–1411.

(34) Kalgutkar, A. S.; Zhao, Z. Discovery and design of selective cyclooxygenase-2 inhibitors as nonulcerogenic, antiinflammatory drugs with potential utility as anti-cancer agents. *Curr. Drug Targets* **2001**, *2*, 79–106.

(35) Kozak, K. R.; Prusakiewicz, J. J.; Rowlinson, S. W.; Schneider, C.; Marnett, L. J. Amino acid determinants in cyclooxygenase-2 oxygenation of the endocannabinoid 2-arachidonylglycerol. *J. Biol. Chem*. **2001**, *276*, 30072−30077.

(36) Merck Frosst Canada, Inc. US5849943, 1998.

(37) Merck Frosst Canada, Inc. US5789413, 1998.

(38) Merck Frosst Canada, Inc. US5733909, 1998.

(39) Lab. UPSA, PCT WO9815528, 1998.

(40) Banner, D. W. Principles of enzyme−inhibitor design. In: *Protein−Ligand Interactions*; Böhm, H.-J., Schneider, G. Eds., Wiley-VCH: Weinheim, 2003; pp 163−185.

(41) Hilpert, K.; Ackermann, J.; Banner, D. W.; Gast, A.; Gubernator, K.; Hadvary, P.; Labler, L.; Müller, K.; Schmid, G.; Tschopp, T. B.; van de Waterbeemd, H. Design and synthesis of potent and highly selective thrombin inhibitors. *J. Med. Chem*. **1994**, *37*, 3889−3901.

(42) Patel, Y.; Gillet, V. J.; Bravi, G.; Leach, A. R. A comparison of the pharmacophore identification programs: Catalyst, DISCO and GASP. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 653−681.

(43) Berliner, L. J. *Thrombin: Structure and Function*; Plenum Press: New York, 1992.

(44) Kimball, S. D. Challenges in the development of orally bioavailable thrombin active site inhibitors. *Blood Coagulation Fibrinolysis* **1995,** *6*, 511−519.

(45) Kikumoto, R.; Tamao, Y.; Tezuka, T.; Tonomura, S.; Hara, H.; Ninomiya, K.; Hijikata, A.; Okamoto, S. Selective inhibition of thrombin by (2*R*,4*R*)-4-methyl-1-[N2-[(3-methyl-1,2,3,4-tetrahydro-8-quinolinyl)sulfonyl]-l-arginyl)]-2-piperidinecarboxylic acid. *Biochemistry* **1984**, *23*, 85−90.

(46) Sturzebecher, J.; Markwardt, F.; Viogt, B.; Wagner, G.; Walsmann, P. Cyclic amides of N alpha-arylsulfonylaminoacylated 4-amidinophenylalanine−tight binding inhibitors of thrombin. *Thromb. Res.* **1983**, *29*, 635−642.

(47) Kato, M.; Nishida, S.; Kitasato, H.; Sakata, N.; Kawai, S. Cyclooxygenase-1 and cyclooxygenase-2 selectivity of nonsteroidal antiinflammatory drugs: investigation using human peripheral monocytes. *J. Pharm. Pharmacol*. **2001**, *53*, 1679−1685.

(48) Penning, T. D.; et al. Synthesis and biological evaluation of the 1,5-diarylpyrazole class of cyclooxygenase-2 inhibitors: identification of 4-[5-(4-methylphenyl)-3-(trifluoromethyl)-1*H*-pyrazol-1-yl]benzenesulfonamide (SC-58635, celecoxib). *J. Med. Chem.* **1997**, *40*, 1347−1365.

(49) 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 anti-inflammatory agents. *Nature* **1996**, *384*, 644−648.

(50) Ulbrich, H.; Fiebich, B.; Dannhardt, G. Cyclooxygenase-1/2 (COX-1/COX-2) and 5-lipoxygenase (5-LOX) inhibitors of the 6,7-diaryl-2, 3-1H-dihydropyrrolizine type. *Eur. J. Med. Chem.* **2002**, *37*, 953−959.

(51) Laufer, S.; Neher, K.; Striegel, H. G. Anti-inflammatory oxo and hydroxy derivatives of pyrrolizines and their pharmaceutical use. WO-00105792, 2001.

(52) Albert, D.; Zündorf, I.; Dingermann, T.; Müller, W. E.; Steinhilber, D.; Werz, O. Hyperforin is a dual inhibitor of cyclooxygenase-1 and 5-lipoxygenase. *Biochem. Pharmacol.* **2002**, *64*, 1767−1775.

(53) Charlier, C.; Michaux, C. Dual inhibition of cyclooxygenase-2 (COX-2) and 5-lipoxygenase (5-LOX) as a new strategy to provide safer nonsteroidal anti-inflammatory drugs. *Eur. J. Med. Chem*. **2003,** 38, 645−659.

(54) Flower, R. J.; Vane, J. R. Inhibition of prostaglandin synthetase in brain explains the anti-pyretic activity of paracetamol (4-acetamidophenol). *Nature* **1972**, *240*, 410−411.

(55) Smith, W. L.; Garavito, R. M.; Dewitt, D. L. Prostaglandin endoperoxide H synthases (cyclooxygenases)-1 and -2. *J. Biol. Chem.* **1996**, *271*, 33157−33160.

(56) Trummlitz, G.; van Ryn, J. Designing selective COX-2 inhibitors: molecular modeling approaches. *Curr. Opin. Drug Discovery Dev*. **2002**, *5*, 550−561.

(57) Burbidge R, Trotter M, Buxton B, Holden S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5−14.

(58) Liu, Y. A comparative study on feature selection methods for drug discovery. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1823−1828.

(59) Jones, G.; Willett, P.; Glen, R. C. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 532−549.

(60) Brungs, M.; Rådmark, O.; Samuelsson, B.; Steinhilber, D. Sequential induction of 5-lipoxygenase gene expression and activity in Mono Mac 6 cells by transforming growth factor-beta and 1,25-dihydroxyvitamin D3. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 107−111.

(61) Brune, K.; Reinke, M.; Lanz, R.; Peskar, B. A. Monoclonal antibodies against E- and F-type prostaglandins. High specificity and sensitivity in conventional radioimmunoassays. *FEBS Lett.* **1985**, *186*, 46−50.

(62) Yamamoto, S.; Yokota, K.; Tonai, T.; Shono, F.; Hayashi, Y. *Enzyme Immunoassay. Prostaglandins and Related Substances − A Practical Approach*. IRL Press: Oxford, 1987.

JM050619H