In Silico Functional Profiling of Small Molecules and Its Applications

Tomohiro Sato,[†,‡] Yo Matsuo,[‡,§,||] Teruki Honma,[‡] and Shigeyuki Yokoyama*[,†,‡]

Department of Biophysics and Biochemistry, Graduate School of Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan, RIKEN Systems and Structural Biology Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan, and Department of Supramolecular Biology, International Graduate School of Arts and Sciences, Yokohama City University, 1-7-29 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan

Received May 1, 2008

In silico screening is routinely used in the drug discovery process to predict whether each molecule in a database has a function of interest, such as inhibitory activity for a target protein. However, drugs generally have multiple functions including adverse effects. In order to obtain small molecules with desirable physiological effects, it is useful to simultaneously predict as many functions as possible. We employed Support Vector Machine to build classification models for 125 molecular functions, derived from the MDDR database, which showed higher kappa statistics (0.775 on average) than those of predictions by Tanimoto similarity (0.708). By analyzing the patterns of the predicted values (functional profiles) of 871 marketed drugs, we demonstrated its applications to indication discovery, clustering of drugs, and detection of molecular actions related to adverse effects. The results showed that functional profiling can be a useful tool for identifying the multifunctionality or adverse effects of small molecules.

Introduction

Since the 1990s, *in silico* screening has been widely used to accelerate the drug discovery process in the pharmaceutical industry. The screening predicts whether each molecule in a database has a function of interest, such as affinity for a target protein. The prediction methods are classified into 2D-based methods, using descriptors calculated from 2D structures, and 3D-based methods, such as docking with a target protein or a pharmacophore search. Among them, the 2D-based methods are more or less based on the detection of structural similarities to molecules that are known to have the target function, assuming that similar molecules have similar functions.[1,2] The comparison of molecular fingerprints, which represent the structural features of individual molecules, is a popular approach to similarity detection.[3] Comparisons of descriptors based on the topological structures of small molecules are also often used.[4,5] However, in cases where structurally different compounds have the same functions, a simple similarity analysis does not work well. In addition to the classical structural similarity, machine-learning algorithms, such as an artificial neural network and Support Vector Machine (SVM),[a,6] have recently been used to develop methods for recognizing the functionally important structural patterns shared by a set of known active compounds.[7−14]

In the *in silico* screening using the above methods, predictions are usually made for only one or a few target function(s) that the new drug candidates should have. However, a small molecule drug can have multiple significant functions, including unexpected side effects, as shown by many examples.[15−17] As a typical example, tricyclic antidepressants are known to have some adverse effects, due to their interactions with off-target proteins. The treatment with the tricyclic antidepressants causes dry mouth, constipation, and ocular side effects due to their anticholinergic activity;[18] weight gain due to their antihistaminergic activity;[19] and hypotension through α-adrenergic blockade.[20] To assess the multiple functionalities of drug candidates, it is desirable to make predictions on many possible functions, including both target and off-target functions, in the first *in silico* screening stage. Searching for molecules with desirable "patterns of predicted functions" (called "functional profiles" here) should facilitate the triage of better chemical classes for further drug development.

Recently, several approaches have been reported to comprehensively identify the target proteins of small molecules.[21] Fliri et al. associated the molecular structures of small compounds with their biological activity profiles, using 92 biological assays to represent a cross section of the drugable proteome, called Biospectra.[22,23] Muller et al. screened 2150 active sites from the Protein Data Bank to identify the putative targets of five small molecules using high-throughput docking.[24] Screening by molecular docking can predict the affinities and the binding modes of ligands complexed with their target proteins, unlike screening based on molecular fingerprints or descriptors of the ligands. However, molecular docking requires information about the 3D structures and the appropriate binding sites of the target proteins. In recent years, the throughput of X-ray and NMR analyses of proteins has been largely improved; however, there are still many target proteins with unsolved structures, and there are also many diseases and adverse effects with molecular mechanisms and target proteins that are not fully understood. On the other hand, ligand-based approaches using machine-learning algorithms are capable of predicting such functions without complete knowledge of their mechanisms.[10] Furthermore, the use of ligand-based approaches can save the cost and calculation time required for *in silico* screening, as compared with structure-based approaches.

* Address correspondence to this author at The University of Tokyo. Telephone: +81-3-5841-4395. Fax: +81-3-5841-8057. E-mail: yokoyama@biochem.s.u-tokyo.ac.jp.

† Department of Biophysics and Biochemistry, The University of Tokyo.

‡ Systems and Structural Biology Center, RIKEN.

§ Department of Supramolecular Biology, Yokohama City University.

|| Current address: OncoTherapy Science, Inc.

[a] Abbreviations: SVM, Support Vector Machine; MDDR, MDL Drug Data Report; RBF, radial basis function; TP, the number of true positives; FP, the number of false positives; FN, the number of false negatives; TN, the number of true negatives; SGOT, serum glutamic oxaloacetic transaminase; SGPT, serum glutamic pyruvic transaminase; LDH, lactate dehydrogenase; GGT, γ-glutamyl transferase; CNS, central nervous system.

10.1021/jm800504q CCC: $40.75
Published on Web 11/25/2008

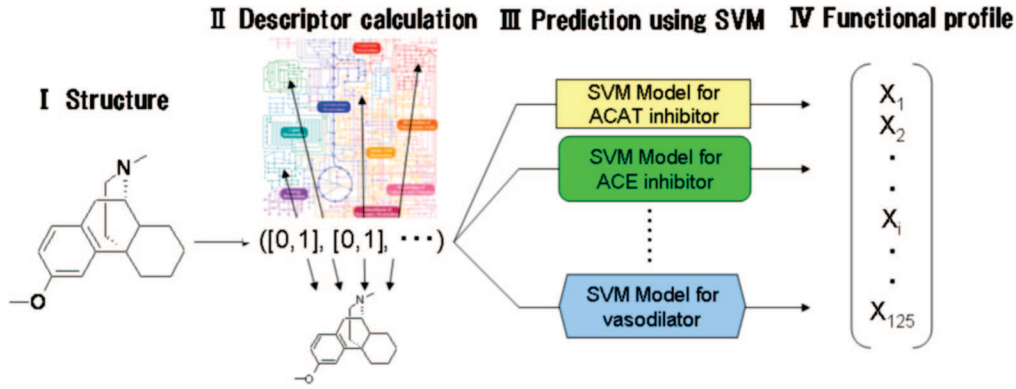


Figure 1. Overview of the functional profiling. The 2D structure of a compound (I) was compared with each of the reference compounds selected from KEGG (II). The obtained vector was used as input for the prediction of the 125 functions using SVM (III). The functional profile of a compound was defined as the array of the 125 predictions (IV).

In the present work, we have pursued this idea of "*in silico* functional profiling" using ligand-based information and the machine-learning technique. Classification models using SVM were built to make predictions for 125 functions observed in existing active compounds and/or marketed drugs in the MDL Drug Data Report (MDDR) database. The functional profile of each small molecule is defined as a 125-dimensional vector, with components representing the results from the SVM prediction models.

In addition, the potential use of *in silico* functional profiling in drug discovery is discussed. The first application of functional profiling is indications discovery. Since existing drugs approved by regulatory agencies for human use have acceptable pharmacokinetics and safety profiles in many cases, identifications of existing drugs for new therapy are expected to reduce the cost and accelerate the drug discovery process.[25,26] In this study, we created functional profiles of 871 known drugs and investigated compounds predicted to have diverse functions, including new drug opportunities.

Second, we tried to classify the existing drugs based on the functional profiles. Recently, the prospect that drugs that interact with several molecular targets simultaneously will lead to new and more effective medications was reported.[27] However, it is difficult to identify good combinations of targets for a particular therapy. The clustering analysis based on the functional profiles can contribute the discovery of better candidate combinations for the therapy.

Finally, functional profiling was applied to detect the molecular actions that are relevant to the adverse effects. Previously, Fliri et al. tried to associate Biospectra and the adverse effects of drugs,[28,29] and Bender et al. compared Bayesian-based predictions about target proteins and adverse effects.[30] In this study, we focused on the human liver adverse effects and introduced a mathematical measure to evaluate the relevancies of the target proteins to the adverse effects. The functional profiles of the existing drugs were compared to the information about their human liver-related adverse effects. High propensities were found in the components of the functional profiles of drugs causing hepatotoxicity on some molecular actions that are known to be relevant to the occurrence of the adverse effects.

Methods

A schematic diagram of our functional profiling is shown in Figure 1. A compound was structurally compared to each of the predefined reference compounds. The vector consisting of the resulting structural similarities, representing the structural properties of the compound, was used as the input for SVM. The predictions of the 125 functions were performed one by one, using SVM models learned from the MDDR database. The details of each calculation step are described below.

Representation of Small Molecules. To apply SVM learning, the structure of each small molecule was represented by a multidimensional vector whose components were similarity measures against reference compounds. The reference compounds were taken from the KEGG COMPOUND database, which is a chemical structure database of metabolic compounds, macromolecules, and other chemical substances, such as inhibitors of metabolic pathways as well as drugs and xenobiotic chemicals that are relevant to biological systems (10932 compounds as of April 2004).[31,32] The similarity between two small molecules was calculated by our in-house method, as described below. The reference compounds were selected as follows. First, a compound was arbitrarily chosen from the KEGG database as the first reference compound (r_1). Next, from the remaining compounds in the database, another compound was randomly chosen as the second reference compound (r_2), such that its similarity against r_1 was less than 0.75. In the same way, the ith reference compound, r_i, was selected, such that its similarities against r_1, r_2,..., r_{i-1} were all less than 0.75. The selection process continued until there were no more compounds left with similarities against the reference compounds selected thus far that were less than 0.75. As a result, 173 reference compounds were obtained. For a given small molecule compound, a 173-dimensional vector was defined, such that its ith component was the similarity between the compound and the ith reference compound. The vectors defined in this way for small molecules were used as the inputs for making the SVM models.

Structural similarities between two small molecules were calculated by the following procedure. The atoms of the small molecules were classified into seven pharmacophore features according to the PATTY algorithm:[33] "cation", "anion", "donor", "acceptor", "polar", "hydrophobic" and "other". In addition, the atoms of an aromatic ring were labeled as "aromatic". For each atom classified as "cation", "anion", "donor", "acceptor", or "polar", a pharmacophore distance matrix $A = (a_{i,j})$ was defined, where $a_{i,j}$ is the number of other atoms of type i at a distance j from the atom ($i \in$ {"cation", "anion", "donor", "acceptor", "polar", "hydrophobic", "other", "aromatic"}, $1 \le j \le 9$). The distance between two atoms was defined by the length of the shortest path of covalent bond counts between them in the chemical structure of the molecule. In some cases, an ether oxygen ("acceptor") of a molecule is converted into a hydroxyl group ("polar") *in vivo*. Therefore, it was treated as 0.5 "acceptor" atom plus 0.5 "polar" atom in defining $a_{i,j}$. In order to avoid oversensitivity to small differences in distance, $a_{i,j}$ was modified to give $a'_{i,j}$: $a'_{i,j} = 0.5a_{i,j-1} + a_{i,j} + 0.5a_{i,j+1}$ ($a_{i,j-1} = 0$ if $j - 1 < 0$, $a_{i,j+1} = 0$ if $j + 1 > 9$). Next, $a''_{i,j}$ was defined to give more weight to polar atoms, because polar interactions, such as hydrogen bonds and their geometry, are more important than hydrophobic interactions for molecular recognition: $a''_{i,j} = u_i w_{i,j} a'_{i,j}$,

where $u_i = 10$, $w_{i,j} = 1 - j/9$ if i = {"cation", "anion", "donor", "acceptor", "polar"} and $u_i = 3$, $w_{i,j} = \max\{1 - j/7, 0\}$ if i = {"hydrophobic", "other", "aromatic"}. For each pair of atoms from two molecules, the Tanimoto coefficient (Tc) was calculated if the atoms were of the same type and if they were "cation", "anion", "donor", "acceptor", or "polar":

$$\mathrm{Tc} = \frac{\sum_{i,j} a''_{1,i,j} a''_{2,i,j}}{\sum_{i,j} \left(a''_{1,i,j}{}^2 + a''_{2,i,j}{}^2 - a''_{1,i,j} a''_{2,i,j} \right)} \quad (1)$$

where $a''_{1,i,j}$ and $a''_{2,i,j}$ are as defined above for the two atoms. Finally, the similarity between two molecules was measured by the largest of all of the Tc values. If no Tc value was available, then the similarity was set to 0. This method is based on counts of atom pairs characterized by the graph distances and their types of atoms, as with CATS[4] (atom pairs) and Similog[5] (atom triplets). In both CATS and Similog, a molecule is represented by one vector. However, a molecule is represented by several vectors in our method. Each vector reflects the topological structure around each atom in the molecule. The similarity between two molecules is defined as the highest Tc among those between all possible pairs of vectors representing atoms of the same type.

SVM Learning. SVM is regarded as one of the most popular and effective machine-learning algorithms for pattern recognition and classification. SVM models nonlinearly discriminate two classes of compounds, by mapping data vectors to a very high dimensional descriptor space and finding a hyperplane that separates the two classes with the largest margin. The most significant difference between SVM and simple linear discrimination is the so-called "kernel trick". Using a kernel function, such as a radial basis function (Gaussian kernel), SVM can obtain a complicated nonlinear separating hyperplane and is particularly effective for a difficult classification problem, such as the prediction of biological activity. A full description of the use of SVM for classification was reported by Cristianini et al.[34]

For each of the 125 functions, the data set of small molecules containing both positive and negative cases for the target function was prepared using the MDDR database. All of the compounds with each target function in MDDR were treated as positive cases for the target. As negative cases, 5000 compounds without the function in the MDDR database were randomly selected. Each of these data sets was split for 5-fold cross-validation. Using the sets, 125 SVM models were built. The radial basis function (RBF) kernel was adopted, because it has been used in previous studies with good results.[7,10] The gamma parameter in the RBF kernel was optimized so as to maximize the results of the cross-validations. Here, the results of the prediction were evaluated using balanced accuracy and kappa statistics. In general, accuracy (also called concordance, (TP + TN)/(TP + FP + TN + FN)) is not an appropriate measure to assess a data set with an uneven ratio of positive and negative compounds. In such cases, instead of accuracy, balanced accuracy and kappa statistics are often used as measures. Balanced accuracy is an average of positive and negative accuracies and is defined as 0.5(TP/(TP + FN) + TN/(TN + FP)). Kappa statistics means the true accuracy, by which the agreement by chance is corrected. A value higher than 0.4 is desirable.[35] It is defined as

$$\mathrm{kappa} = \frac{P_\mathrm{o} - P_\mathrm{c}}{1 - P_\mathrm{c}} \quad (2)$$

P_o indicates the probability of observed agreement, (TP + TN)/(TP + FN + FP + TN), and P_c indicates the probability of chance agreement, [(TP + FP)/(TP + FP + FN + TN)][(TP + FN)/(TP + FP + FN + TN)] + [(TN + FN)/(TP + FP + FN + TN)][(TN + FP)/(TP + FP + FN + TN)].

The SVM$^{\mathrm{light}}$ software[36] was used for model building and predictions. Cost factor, by which training errors on positive examples outweigh errors on negative examples, was defined as the ratio between the number of positive and negative compounds, in order to accommodate the difference in the number of positive and negative cases. Other parameters, such as C, regulating the tradeoff between minimization of training error and maximization of margin, were set to the default values in SVM$^{\mathrm{light}}$.

The performances of the SVM predictions were compared with those obtained by similarity searching, using the Tanimoto coefficient between the MDL public keys (MACCS keys) fingerprint (166 bits). The threshold of Tanimoto similarity, to distinguish the positive and negative predictions, was determined so as to optimize the mean of the balanced accuracies of the similarity-based predictions about the 125 functions upon cross-validation.

Definition of the Functional Profile. In this study, the functional profile of a molecule was defined as a 125-dimensional vector (x_1, x_2,..., x_{125}), where its ith component x_i was set to 1 when the SVM model predicted the molecule is positive (the prediction value f_i from the SVM model for the ith function was ≥ 0); otherwise, x_i was set to 0. When a function of the molecule associated with the ith component was already annotated in MDDR, the component was also set to 1.

The 125 functions, as found in the annotations of MDDR, were used as the prediction targets. These functions consist of 70 molecular actions, such as COX inhibitor or NMDA receptor antagonist, and 55 therapeutic areas, such as antiinflammatory or antihypertensive. The list of 125 functions is shown in Table 1. Minor variations in the function annotations were ignored. For example, inhibitors of different subtypes of a protein were treated as a single functional category. Agonists and antagonists of a protein were not discriminated in this study. Since MDDR is a hand-curated database, in which the annotations describe only the presence or absence of functions for each compound, the lack of a controlled dictionary, the overlap of functions, and the incompleteness of annotations are inevitable on some level. In order to address these problems, we analyzed the results about molecular actions and therapeutic areas separately or developed a statistical measure to correct the bias from the overlap of the functions. The details of each procedure are described in the corresponding sections.

Detecting Molecular Actions Closely Related to Adverse Effects. The Human Liver Adverse Effects Database[37] is a database containing data of the adverse effects for 490 pharmaceuticals developed by the FDA. Among the 490 pharmaceuticals, 314 are also included in MDDR. The database contains data about the activity of the liver enzyme composite module, alkaline phosphatase increase, serum glutamic oxaloacetic transaminase (SGOT) increase, serum glutamic pyruvic transaminase (SGPT) increase, lactate dehydrogenase (LDH) increase, and γ-glutamyl transferase (GGT) increase. The activity of each compound is labeled as either "active", "marginally active", or "inactive". In this study, the "active" and "marginally active" compounds were combined into the "active" class.

The relevancies of molecular actions to the liver-related adverse effects were calculated by comparing the functional profiles of the 314 drugs and the information about their adverse effects. We excluded 22 out of 70 molecular actions from the calculation, because only two or fewer drugs out of the 314 drugs were predicted to have the functions by our SVM models. The relevance of the ith function to the jth adverse effect was defined by the following procedure. The ratio of positive prediction of the ith function for drugs that are active for the jth adverse effect ($P_{i,j,\mathrm{active}}$) was defined as

$$P_{i,j,\mathrm{active}} = \frac{N_{i,j,\mathrm{active}}}{N_{j,\mathrm{active}}} \quad (3)$$

where $N_{j,\mathrm{active}}$ represents the number of compounds that are active for the jth adverse effect and $N_{i,j,\mathrm{active}}$ represents the number of compounds that are predicted or already known to have the ith function and are active for the jth adverse effect. $P_{i,j,\mathrm{inactive}}$ was also defined as follows:

Table 1. 125 Drug Functions Used as the Prediction Targets

function	no. of active compounds	function	no. of active compounds
(A) 70 Functions about Molecular Actions[a]			
ACAT inhibitor	1385	K^+ channel activator	860
ACE inhibitor	493	K^+ channel blocker	256
acetylcholinesterase inhibitor	720	leukotriene agonist/antagonist	1481
adenosine agonist/antagonist	516	LHRH agonist/antagonist	179
adrenoceptor agonist/antagonist	398	lipid peroxidation inhibitor	595
aldose reductase inhibitor	916	lipoxygenase inhibitor	2699
AMPA receptor antagonist	506	MMP inhibitor	449
angiotensin 2 agonist/antagonist	2244	muscarinic agonist/antagonist	1112
antiestrogen	256	neurokinin agonist/antagonist	385
antioxidant	407	nitric oxide synthase inhibitor	417
aromatase inhibitor	556	NMDA receptor antagonist	1429
benzodiazepine	336	oxazolidinone	364
Ca^{2+} channel blocker	1610	oxytocin antagonist	198
cAMP phosphodiesterase inhibitor	200	PAF analogue/antagonist	1440
carbapenem	1275	phosphodiesterase inhibitor	2131
carbonic anhydrase inhibitor	269	phospholipase inhibitor	720
CCK	820	PKC inhibitor	448
cephalosporin	1418	prolylendopeptidase inhibitor	300
cholinergic	289	prostaglandin	446
collagenase inhibitor	523	protease inhibitor	487
COX inhibitor	1282	quinolone	1075
dopamine agonist/antagonist	1838	renin inhibitor	671
elastase inhibitor	605	reverse transcriptase inhibitor	539
endothelin agonist/antagonist	906	serotonin agonist/antagonist	4313
estrogen	195	sigma antagonist	438
factor Xa inhibitor	508	squalene synthetase inhibitor	456
farnesyl protein transferase inhibitor	944	steroid reductase inhibitor	936
gastrin antagonist	327	substance P antagonist	1253
GP2b3a receptor antagonist	1255	thrombin inhibitor	931
growth hormone release promoting agent	280	thromboxane antagonist	835
H^+/K^+ ATPase inhibitor	720	TNF inhibitor	807
histamine agonist/antagonist	297	topoisomerase inhibitor	241
HIV protease inhibitor	650	tyrosine kinase inhibitor	963
HMG-CoA reductase inhibitor	1023	vasopressin antagonist	276
IL-1 inhibitor	343	vitamin D analogue	355
(B) 55 Functions about Therapeutic Areas			
agent for cognition disorders	6073	antineoplastic	12057
agent for pancreas disorders	310	antiobesity	1025
agent for pulmonary emphysema	575	anti-Parkinsonian	1176
agent for restenosis	671	antiprotozoal	269
agent for urinary incontinence	739	antipsoriatic	1907
analgesic nonopioid	2855	antipsychotic	3943
analgesic opioid	1057	antiulcerative	1816
anthelmic	374	antiviral	2817
antiacne	1253	antiviral AIDS	2679
antiallergic	8133	anxiolytic	4976
antianginal	2620	bone resorption inhibitor	551
antiangiogenic	546	bronchodilator	2172
antiarrhythmic	2160	cardiotonic	2294
antiarthritic	5769	gastric antisecretory	1641
antibacterial	2601	hair growth promoter	394
antibiotic	1763	hypolipidemic	4919
anticoagulant	1573	immunomodulator	1076
anticonvulsant	2485	immunostimulant	359
antidepressant	3986	immunosuppressant	1166
antidiabetic	2601	neural injury inhibitor	4393
antiemetic	1078	platelet antiaggregatory	4114
antifungal	1739	sedative/hypnotic	540
antiglaucoma	1095	signal transduction inhibitor	429
antihypertensive	9455	stimulant peristaltic	657
antiinflammatory	5217	treatment for osteoporosis	1475
antiischemic	1849	treatment for septic shock	991
antimalarial	256	vasodilator	948
antimigraine	1382		

[a] The numbers represent the number of drugs registered in MDDR. At least 179 drugs were found in MDDR for each function.

$$P_{i,j,\text{inactive}} = \frac{N_{i,j,\text{inactive}}}{N_{j,\text{inactive}}} \quad (4)$$

where $N_{j,\text{inactive}}$ and $N_{i,j,\text{inactive}}$ represent the number of compounds that are inactive for the jth adverse effect and those predicted or known to have the ith function, respectively. The relevance of the ith function to the jth adverse effect ($R_{i,j}$) was defined as the ratio of $P_{i,j,\text{active}}$ and $P_{i,j,\text{inactive}}$.

$$R_{i,j} = \frac{P_{i,j,\text{active}}}{P_{i,j,\text{inactive}}} \quad (5)$$

$R_{i,j}$ values are the "odds ratio" of functions and liver-related adverse

effects. If the $R_{i,j}$ value of a particular function-adverse effect pair is significantly higher than 1.0, then the function should be closely related to the occurrence of the adverse effect. In this study, the functions with a relevance ($R_{i,j}$) to any adverse effect that exceeded 2.0 were investigated.

Results and Discussion

Prediction Performances of the SVM Models. Classification models by SVM and Tanimoto similarity were built for the 125 functions and were validated by 5-fold cross-validation. The means of the balanced accuracies (the average of the positive and negative accuracies) of the 125 SVM models showed a slightly higher value (0.912 (standard deviation: 0.093)), as compared with the value (0.896 (standard deviation: 0.052)) of the similarity-based models. The slight difference of the balanced accuracy between the SVM models and the similarity-based models mainly arose from the negative accuracies. The means of the negative accuracies were 0.948 for the SVM models and 0.902 for the similarity-based models, while the means of the positive accuracies were 0.875 and 0.890, respectively. The negative accuracies of the similarity-based models significantly decreased, particularly when more than 3000 active compounds were associated for the functions in MDDR. For example, in the prediction models for antineoplastic activity (12057 active compounds in MDDR), the negative accuracy of the similarity-based model showed a much lower value (0.565) than that of the SVM model (0.845). It is reasonable that the SVM models maintained high prediction performances for negative accuracies in those cases, because machine learning can learn information about negative compounds in addition to positive compounds. According to the previous studies using the SVM method for protein targets, the balanced accuracies were 0.891 (carbonic anhydrase II),[9] approximately 0.95 (factor Xa), and 0.85 (kinase inhibitors).[14] Although the performances of the models could not be directly compared, because of the differences in the data sets and the activity criteria, the balanced accuracies of our SVM models for carbonic anhydrase inhibitors, factor Xa inhibitors, and tyrosine kinase inhibitors showed higher values (0.974, 0.963, and 0.908, respectively) than those of the reported models.

When the ratios of positive and negative compounds are largely biased, it is difficult to compare the prediction performances of classification models by accuracy. In such cases, kappa statistics is widely used to assess the prediction performances of classification models. Kappa statistics subtracts the probability of random agreements from accuracy (eq 2). According to the equation, the kappa statistics of a random model (compounds are randomly predicted by the ratio of positive and negative compounds in the training set) or a one-side model (all compounds are positively predicted or all compounds are negatively predicted) results in 0, where perfect predictions result in 1.0. For example, consider a data set consisting of 10 positive and 90 negative compounds. If a model predicts all 100 compounds as negative, then the accuracy has a very high value (0.900). However, the model cannot predict true positives at all. In contrast, the balanced accuracy and the kappa statistics of the model are 0.500 and 0. Obviously, the balanced accuracy and the kappa statistics more adequately assess the prediction performance of the one-side model, which cannot predict positive compounds, than the simple accuracy, and the kappa statistics show the true performance of models over accidental coincidence. In our data sets, the ratios of positives and negatives are also uneven (1:7.70 on average). Therefore, kappa statistics is expected to assess the true performance of prediction models for our data sets. The 125 SVM models recorded higher kappa statistics than those yielded by the similarity-based models in 108 out of the 125 functions (86.4%) (Figure 2). The kappa statistics of vasopressin antagonist models by the SVM and similarity-based methods were 0.851 and 0.534, respectively, which represented the largest difference among the 125 functions. On the other hand, the balanced accuracies of the predictions by the two methods yielded the almost same values (0.926 and 0.939). This distinction arose from the difference in the precision (true positive rate of all positively predicted compounds, TP/(TP + FP)). On the 5-fold cross-validation, each vasopressin antagonist test set contained about 55 positive samples and 1000 negative samples, on average. The SVM-based model predicted 55 compounds as positive, and 47 of these compounds were actually positive (precision: 0.860). However, the similarity-based model predicted 130 compounds as positive, and less than half of them (53 compounds) were actually positive (precision: 0.408). The kappa statistics of the vasopressin antagonist models appropriately reflected the differences in the precision, as compared with the balanced accuracy. The means of the kappa statistics of the predictions by the SVM and similarity-based models about the 125 functions were 0.775 and 0.708, respectively. In 100 of the 125 functions (80.0%), the kappa statistics exceeded 0.700, and those about molecular actions (0.816 on average) were higher than those about therapeutic areas (0.721 on average). The difference in the kappa statistics indicated that the therapeutic areas consisting of drugs with multiple mechanisms of action included more diverse chemical structures and were difficult to predict correctly. The high balanced accuracies and kappa statistics indicated that our SVM models have excellent predictability for the 125 functions of the MDDR compounds. The structures and CAS numbers of the five most positively predicted compounds, which were not annotated in MDDR, for 70 molecular actions and 55 therapeutic areas are provided in the Supporting Information.

Functional Profiling of Existing Drugs. To investigate the multifunctionality of existing small molecule drugs, 871 drugs were selected from the MDDR database for functional profiling. They were all "launched" drugs and had 10−35 non-hydrogen atoms, including at least one charged or polar atom. The 125 SVM prediction models were applied to each of these small molecule drugs. The functional profiles of the 871 drugs are shown in the Supporting Information.

Among the 1200 annotations of drug−function relationships already registered in MDDR, 986 annotations were correctly predicted. Furthermore, 6058 novel drug−function pairs were predicted through the functional profiling in addition to the already known functions.

The number of positively predicted functions (d) of each molecule can reflect the functional diversity of a molecule:

$$d = \sum_{i=1}^{125} x_i \tag{6}$$

The values for d ranged from 0 to 27 among the 871 molecules, and the mean and standard deviation were 8.28 and 4.62, respectively. The mean value of d (8.28) is much larger than expected, because the number of annotations per molecule in MDDR is only 1.37, on average.

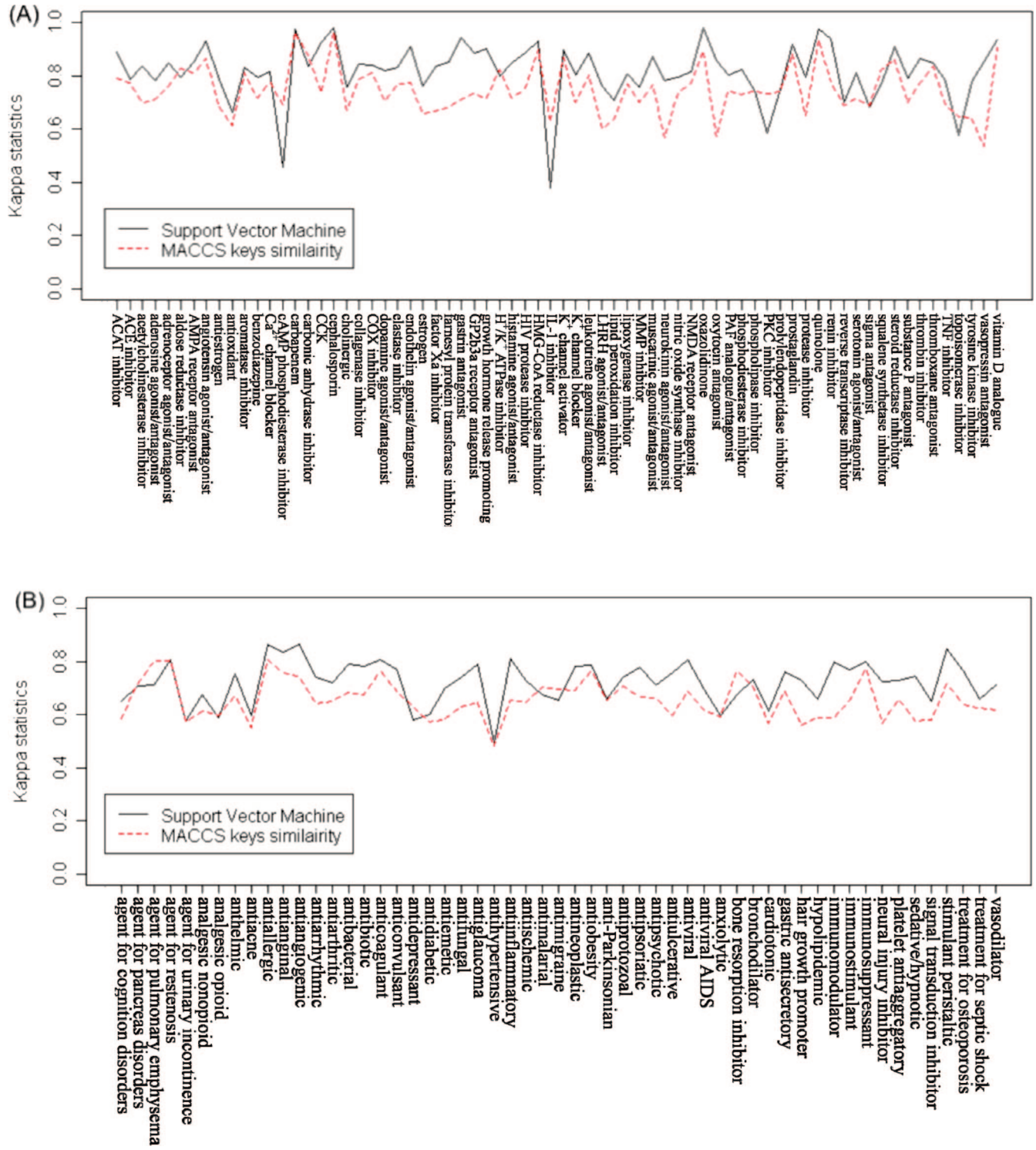


Figure 2. Kappa statistics of the 5-fold cross-validation of the SVM models (black solid line) and similarity searching using the Tanimoto coefficient of the MACCS keys fingerprint (red dotted line) on 70 molecular actions (A) and 55 therapeutic areas (B).

Some of the 125 functions were closely related to each other (e.g., "serotonin agonist/antagonist" and "antipsychotic") and therefore gave positively correlated prediction results. The correlation coefficients between the SVM outputs for the ith and jth functions, f_i and f_j, were calculated ($1 \leq i,j \leq 125$, $i \neq j$). For the 7750 ($=125 \times 124/2$) pairs of functions, the mean and standard deviation of the correlation coefficients were 0.022 and 0.152, respectively. The distribution of the correlation coefficients is shown in Figure 3. Although most of the pairs showed no significant correlations, the following eight pairs gave correlation coefficients of ≥ 0.7: "agent for pulmonary emphysema" and "elastase inhibitor" (0.874); "antipsychotic" and "dopamine agonist/antagonist" (0.816); "antipsychotic" and "serotonin agonist/antagonist" (0.804); "antidepressant" and "antipsychotic" (0.785); "estrogen" and "antiestrogen" (0.769); "antidepressant" and "anxiolytic" (0.762); "antidepressant" and "serotonin agonist/antagonist" (0.738); and "dopamine agonist/antagonist" and "serotonin agonist/antagonist" (0.716).

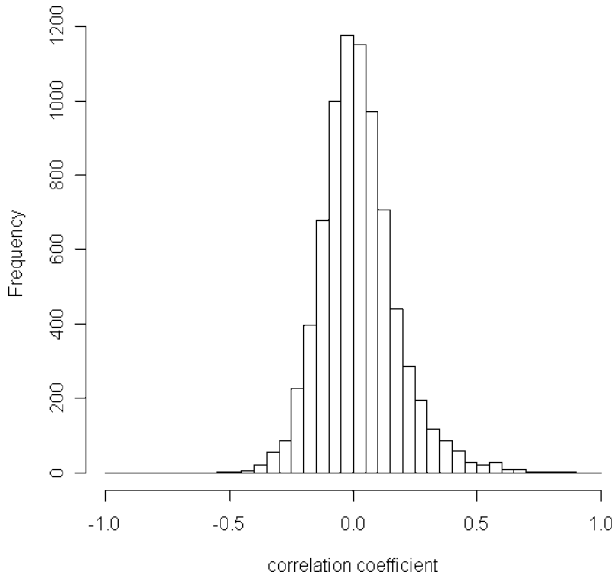


Figure 3. The histogram of correlation coefficients between SVM predictions about the 125 functions.

Table 2. The 20 Most Functionally Diverse Drugs[a]

rank	name	CAS number	*d*	*d′*
1	**dexmedetomidine**	113775-47-6	15	12.49
2	droxicam	90101-16-9	18	12.47
3	**midazolam**	59467-70-8	20	12.38
4	phenindione	83-12-5	18	12.16
5	emorfazone	38957-41-4	17	12.14
6	iobenguane	77679-27-7	16	11.66
7	ondansetron	99614-02-5	18	11.64
8	flunitrazepam	1622-62-4	18	11.61
9	imiquimod	99011-02-6	16	11.60
10	croconazole	77174-66-4	17	11.56
11	ibudilast	50847-11-5	18	11.46
12	caffeine	69-22-7	16	11.44
13	afloqualone	56287-74-2	16	11.31
14	riluzole	1744-22-5	15	10.78
15	disulfiram	97-77-8	16	10.78
16	ormeloxifene	31477-60-8	20	10.77
17	oxandrolone	53-39-4	16	10.75
18	exalamide	53370-90-4	13	10.32
19	cinnoxicam	87234-24-0	14	10.28
20	clobazam	22316-47-8	17	10.14

[a] The top 20 among the 871 existing drugs are listed in the order of the functional diversity, *d′*. *d*, the number of positively predicted functions. *d′*, the functional diversity value.

To avoid counting highly correlative or synonymous functions separately, the functional diversity, *d′*, of each molecule was defined as follows:

$$d' = \sum_{i=1}^{125} \frac{x_i}{\sum_{j=1}^{125} x_j R'_{ij}{}^2} \tag{7}$$

where $R'_{ij} = \max\{0, R_{ij}\}$ and R_{ij} is the correlation coefficient between f_i and f_j. The mean and standard deviation of *d′* were 5.26 and 2.28, respectively. Table 2 lists the top 20 molecules in terms of the functional diversity, *d′*. Among the listed compounds, dexmedetomidine (*d′* = 12.49), and midazolam (*d′* = 12.38) were already well investigated in terms of their side effects in a clinical study and are useful to test our prediction results. These two compounds are highlighted in bold font in Table 2. We compared the predicted functional profiles of these two compounds with their reported side effects and off-target functions.

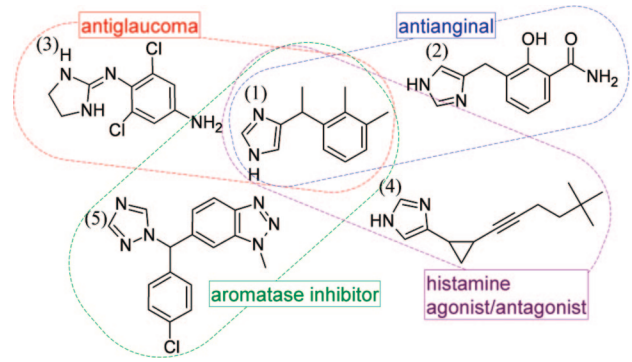


Figure 4. Dexmedetomidine and drugs with potentially related functions: (1) dexmedetomidine (CAS number, 112775-47-6), (2) mivazerol (125472-02-8), (3) apraclonidine (73218-79-8), (4) perceptin (213027-19-1), and (5) vorozole (118949-22-7). Dexmedetomidine was predicted to share some functions with the drugs (2)−(5). The shared functions are shown in the boxes around dexmedetomidine and the associated drugs.

Dexmedetomidine. Dexmedetomidine (Figure 4 (1)) showed the highest functional diversity (*d* = 15, *d′* = 12.49) among the 871 drugs. According to the MDDR database, it is an adrenergic α2 agonist with the known functions of (i) sedative/hypnotic and (ii) analgesic nonopioid. These two functions were correctly predicted in the functional profiling. In addition to these, the following functions were positive in the functional profile: (iii) antianginal, (iv) antiglaucoma, (v) platelet antiaggregatory, (vi) antihypertensive, (vii) antiobesity, (viii) histamine agonist/antagonist, (ix) agent for cognition disorders, (x) aromatase inhibitor, (xi) lipoxygenase inhibitor, (xii) antidepressant, (xiii) antidiabetic, (xiv) antineoplastic, and (xv) H^+/K^+ ATPase inhibitor.

The predicted functions other than sedative/hypnotic activity and analgesic activity, which were previously annotated for dexmedetomidine in MDDR, were validated. Some other adrenergic α2 agonists similar to dexmedetomidine are known to have the functions iii (Figure 4 (2)) and iv (Figure 4 (3)). There are some studies reporting that dexmedetomidine has the functions v[38] and vi.[39] Adrenergic α2 antagonists are known to have the function vii,[40] although dexmedetomidine is an agonist. A histamine agonist/antagonist (viii) can have an effect on appetite through its interaction with histamine H1 receptor, and therefore may be involved in the function vii. Perceptin (Figure 4 (4)), which is a histamine H3 antagonist with the function ix, shares an imidazole ring with dexmedetomidine. Vorozole (Figure 4 (5)), which is structurally similar to dexmedetomidine, is known to have the function x. Dexmedetomidine reportedly has an antiinflammatory effect in rats during endotoxemia,[41] which supports the predictions of the functions viii and xi. By the inhibition of histamine receptor or lipoxygenase, or both of them, many antidepressants (xii) are known to be related to the adrenergic α2 agonist action as well as the functions i, ii, and viii.

Functional profiling of dexmedetomidine indicated not only its major functions annotated in MDDR but also minor functions reported in some papers. Therefore, the functional profile could well represent the features of the more comprehensive effects on humans by dexmedetomidine. In particular, the detection of the functions of drugs structurally similar to dexmedetomidine and the potential mechanisms of its antiinflammatory activity provided insight into the potential use of functional profiling for indication discovery.

Midazolam. Midazolam (Figure 5 (1)) showed the second largest functional diversity value (*d* = 20, *d′* = 12.38).

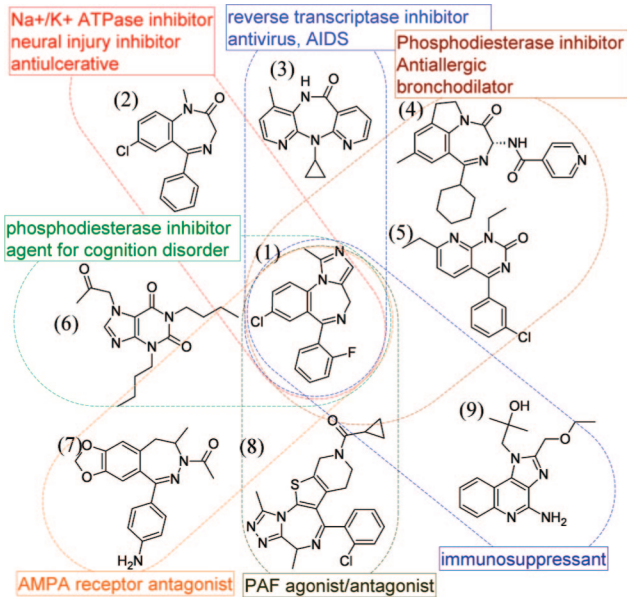


Figure 5. Midazolam and drugs with potentially related functions: (1) midazolam (CAS number, 59467-70-8), (2) diazepam (439-14-5), (3) nevirapine (129618-40-2), (4) **1** (179024-48-7), (5) **2**, (6) denbufylline (57076-71-8), (7) talampanel (161832-65-1), (8) **3** (131614-02-3), and (9) resiquimod (144875-48-9). Midazolam was predicted to share some functions with the drugs (2)−(9). The shared functions are shown in the boxes around midazolam and the associated drugs.

According to the MDDR database, it is known to be (i) sedative/hypnotic and (ii) benzodiazepine. These two functions were correctly predicted in the functional profiling. In addition to these functions, the following functions were positive in the functional profile: (iii) anxiolytic, (iv) anticonvulsant, (v) H^+/K^+ ATPase inhibitor, (vi) neural injury inhibitor, (vii) antiulcerative, (viii) PAF agonist/antagonist, (ix) antiinflammatory, (x) antiallergic, (xi) antiarthritic, (xii) immunomodulator, (xiii) immunosuppressant, (xiv) reverse transcriptase inhibitor, (xv) antiviral AIDS, (xvi) phosphodiesterase inhibitor, (xvii) bronchodilator, (xviii) agent for cognition disorders, (xix) AMPA receptor antagonist, and (xx) gastric antisecretory.

In fact, midazolam is often used for the functions iii and iv in clinical practice. Diazepam (Figure 5 (2)), another benzodiazepine (ii) with the function i, reportedly has the functions v, vi, and vii.[42] Midazolam was reported to attenuate platelet activation (viii) in thrombotic and inflammatory disease (ix).[43] Platelet activation was reported to play significant roles in allergic asthma (x)[44] and rheumatoid arthritis (xi).[45] Sedative/hypnotic benzodiazepine agents, including midazolam, are known to have the functions xii and xiii in general.[46−48] Nevirapine (Figure 5 (3)), a reverse transcriptase inhibitor (xiv) with the function xv, has a tricyclic substructure similar to that of midazolam. Midazolam must not be coadministered with reverse transcriptase inhibitors, because they both inhibit the metabolizing enzyme cytochrome P450 3A4.[49,50] Compounds **1** (CI-1018; Figure 5 (4))[51] and **2** (YM-976; Figure 5 (5))[52] are phosphodiesterase inhibitors (xvi). They are known to have the functions x and xvii, respectively. They share similar ring-system or pharmacophore allocations with midazolam. There is another phosphodiesterase inhibitor, denbufylline (Figure 5 (6)), which has the function xviii. Talampanel (Figure 5 (7)), **3** (E-6123; Figure 5 (8)),[53,54] and resiquimod (Figure 5 (9)) also share similar substructures with midazolam. They have the functions xix, viii, and xiii, respectively.

As in the case of dexmedetomidine, the anxiolytic activity and the anticonvulsant activity of midazolam, which were not annotated in MDDR, were successfully detected. The examples of dexmedetomidine and midazolam demonstrated that our functional profiling is useful to provide candidates for indication discovery.

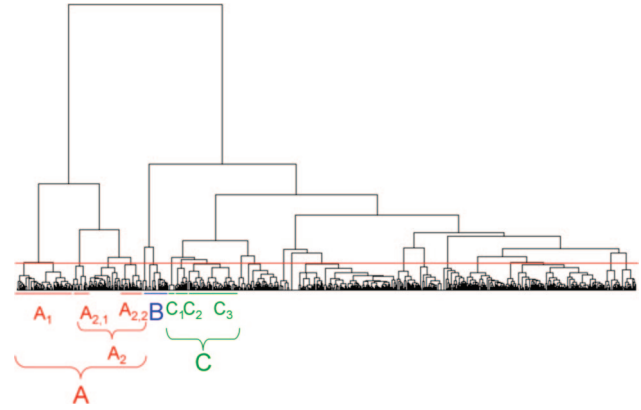


Figure 6. Hierarchical clustering of existing drugs according to their functional profiles. The cluster characterized by CNS action (A) was separated into potential NMDA receptor inhibitors (A_2) and others (A_1), and A_2 included two subclusters characterized by potential interactions with histamine receptors ($A_{2,1}$) and antiischemic effects ($A_{2,2}$). Some of the antiinflammatory drugs were grouped into an isolated cluster (B). A cluster characterized by cardiovascular action (C) was divided into three subclusters, characterized by Ca^{2+} channel blockade (C_1), ACE inhibition (C_2), and other cardiovascular actions (C_3). The thin red horizontal line spanning the entire dendrogram indicates a clustering threshold, which yielded the 23 clusters used in Figure 8.

Clustering of Existing Drugs. It is known that drugs with different molecular mechanisms of action can have a similar therapeutic effect. For example, steroidal and nonsteroidal antiinflammatory drugs act on different proteins to exert a similar effect. In addition, drugs can have a common efficacy by a shared mechanism, while showing different side effects. For example, aspirin and celecoxib are both antiinflammatory drugs that inhibit COX-2, but they show vastly different off-target interactions and side effects.[15,55] Therefore, simply predicting whether or not a drug has a given function provides little information on its mechanisms of action. In general, most drugs have a wide variety of side effects besides their main effects. Their different target specificities against many other off-targets (i.e., functional profiles) are likely to lead to different sets of effects.

In order to investigate the relationship among target specificities, therapeutic effects, and adverse effects, we tried hierarchical clustering of the 871 drugs, based on their functional profiles. The Ward method was used for the clustering, using R.[56] The similarity (or dissimilarity) between two drugs was measured by the Euclidean distance between their functional profiles. The clustering result is shown in Figure 6. Three clusters, A, B, and C, and their subclusters in Figure 6, were found to be populated with drugs that had many positively predicted functions in common in their functional profiles, and therefore, these clusters were likely to reflect the similarities in the mechanisms of action of their member drugs. As representative scaffolds in the clusters A_1, $A_{2,1}$, $A_{2,2}$, B, C_1, C_2, and C_3, the maximal common substructures of their cluster members generated by Pipeline Pilot[57] are shown in Figure 7.

Cluster A contained 176 drugs. Many of them were drugs acting on the central nervous system (CNS). The whole set of 871 drugs contained 44 antidepressants, 40 antipsychotics, and 25 serotonin agonists/antagonists in total, according to the MDDR annotations. Among these, 32 (72.7%), 31 (77.5%), and

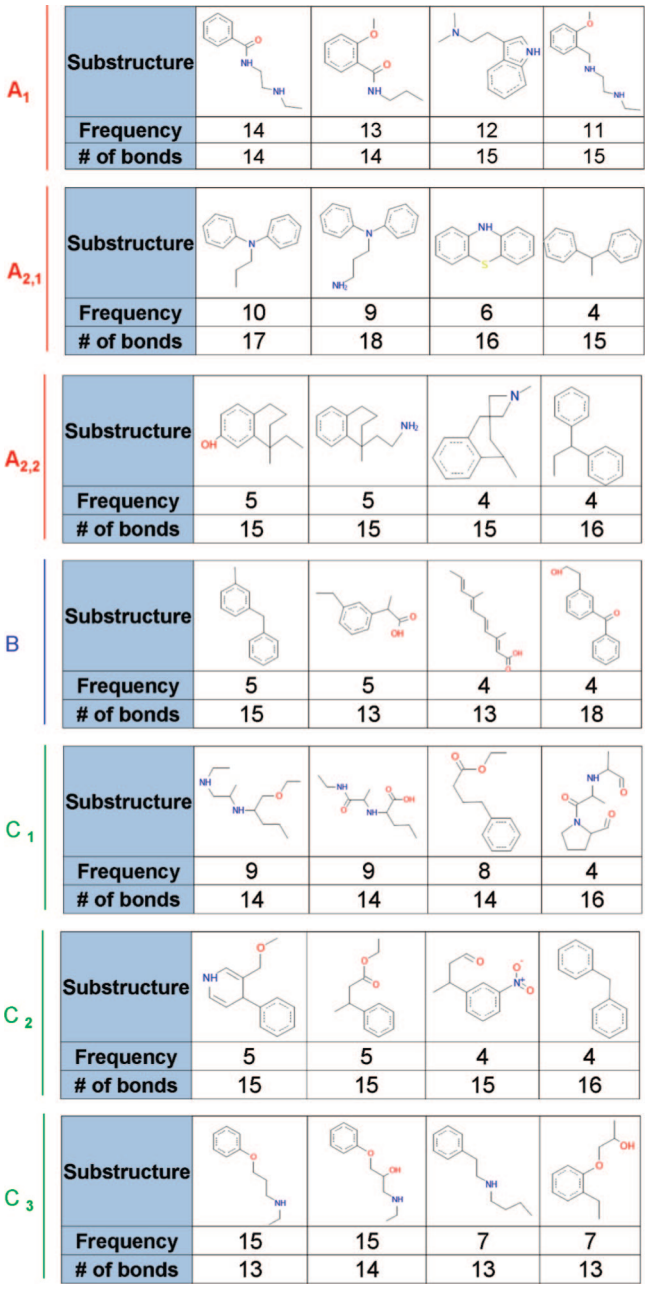


Figure 7. Maximum common substructures of the compounds in the cluster A_1, $A_{2,1}$, $A_{2,2}$, B, C_1, C_2, and C_3 generated by Pipeline Pilot. Each substructure was shared by at least 10% of the compounds in the cluster. The frequency of the appearance of each substructure in the cluster and the number of bonds in each substructure are shown.

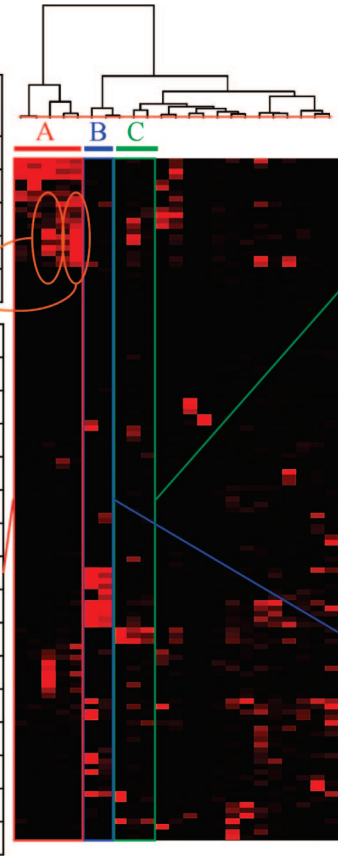
21 (87.5%) drugs, respectively, were included in this cluster. Among the 176 members of cluster A, 150 (85.2%) were positive for "antipsychotic" in their functional profiles, 140 (79.5%) were positive for "serotonin agonist/antagonist", 143 (81.3%) were positive for "anxiolytic", and 131 (74.4%) were positive for "antidepressant". This cluster included several subclusters with different characteristics in their functional profiles. They are described in more detail later.

Cluster B contained 37 drugs. All of them were positive for "antiinflammatory" in their functional profiles, 35 (94.6%) were positive for "antipsoriatic", 33 (89.2%) were positive for "antiacne", 31 (83.8%) were positive for "antiallergic", 30 (83.3%) were positive for "antiarthritic", 28 (77.8%) were positive for "IL-1 inhibitor", 25 (67.6%) were positive for "COX inhibitor", 24 (64.9%) were positive for "leukotriene agonist/antagonist", 24 (64.9%) were positive for "phospholipase inhibitor", and 25 (67.6%) were positive for "platelet antiaggregatory". The whole set of 871 drugs contained 55 antiinflammatory agents in total. Of these, 17 were placed into this cluster.

Cluster C contained 97 drugs. Seventy eight (80.4%) of them were positive for "antihypertensive" in their functional profiles, 49 (50.5%) were positive for "cardiotonic", and 54 (55.7%) were positive for "antianginal". Among the 97 antihypertensive agents in the set of 871 drugs, 54 were placed into this cluster. The cluster consisted of three subclusters, C_1, C_2, and C_3 (Figure 6), containing 9, 18, and 70 drugs, respectively. They were characterized by their notably different positive ratios for the following seven functions: antihypertensive, antianginal, Ca^{2+} channel blocker, antiarrhythmic, bronchodilator, platelet antiaggregatory, and ACE inhibitor (Figure 8). All of the C_1 members were known ACE inhibitors. The subcluster C_2 included 8 of the 18 Ca^{2+} channel blockers in the set of 871 drugs.

As described above, cluster A was characterized by functions in the CNS. This cluster was further divided into two subclusters, A_1 (78 drugs) and A_2 (98 drugs) (Figure 6). The major difference between them was found in their positive ratios for the function of "NMDA receptor antagonist". While no member of A_1 was positive for the function, 65.3% of the A_2 members were positive. A_2 included two subclusters, $A_{2,1}$ (22 drugs) and $A_{2,2}$ (30 drugs), whose members were all positive for "NMDA receptor antagonist". The NMDA receptor is a target of antiischemic agents,[58] and all members of subcluster $A_{2,2}$ were positive for the therapeutic area of "antiischemic". However, $A_{2,1}$ was only 36.4% positive for that. Other significant differences between $A_{2,1}$ and $A_{2,2}$ were observed on "gastric antisecretory" ($A_{2,1}$, 72.7%; $A_{2,2}$, 0.07%) and "antiobesity" (100.0%; 56.7%). The drugs in cluster $A_{2,1}$ that were positive for "NMDA receptor antagonist", "gastric antisecretory", "antiobesity", and negative for "antiischemic" in their functional profiles were all tricyclic antidepressants. Tricyclic antidepressants are known to inhibit NMDA receptors;[59] however, they are not used for the treatment of ischemia, in contrast to the drugs in cluster $A_{2,2}$. A blockade of gastric histamine H2 receptor causes the gastric antisecretory effect. Histamine decreases appetite via the cerebral histamine H1 receptor. Histamine H3 receptor antagonists are promising antiobesity drugs.[60] Histamine release from nerve endings is enhanced in ischemia to contribute to neuroprotection against ischemic damage, and the blockade of cerebral histamine H2 receptors aggravates ischemic injury.[61] Therefore, the $A_{2,1}$ drugs may potentially interact with histamine receptors. In fact, tricyclic antidepressants are also known to inhibit histamine H1 and H2 receptors. They can cause weight gain via cerebral H1 receptor blockade and have been used for the treatment of peptic ulcer, due to their gastric H2 receptor blockade and antisecretory effects.[62] These results suggest that NMDA receptor inhibitors can be safe and better antiischemic drugs by exerting their interactions with histamine receptors.

Relevancies of the Functions to Human Liver Adverse Effects. Table 3 shows the 17 functions whose calculated relevancies ($R_{i,j}$) to at least one of the human liver adverse effects were higher than 2.0. The highest relevancies detected in this study were those of HMG-CoA reductase inhibitor, to increase both liver enzyme composite activity and GGT. In this case, $P_{i,j,\mathrm{active}}$, which represents the ratios of positive prediction of HMG-CoA inhibition for drugs with composite activity increase



	Positive prediction (%)	
	$A_{2,1}$	$A_{2,2}$
NMDA receptor inhibitor	100.0	100.0
antiischemic	41.3	100.0
gastric antisecretory	72.7	6.7
antiobesity	100.0	56.7

	A_1	A_2
anxiolytic	92.3	72.4
antidepressant	66.7	80.6
serotonin agonist/antagonist	100.0	63.3
antipsychotic	82.1	87.8
stimulant peristaltic	51.3	10.2
antiemetic	59.0	12.2
antimigraine	51.3	38.8
dopamine agonist/antagonist	61.5	38.8
analgesic non opioid	14.1	61.2
NMDA receptor antagonist	0.0	65.3
neural injury inhibitor	7.7	83.7
antiischemic	3.8	65.3
sigma antagonist	11.5	80.6
antiparkinsonian	20.5	57.1
agent for cognition disorders	14.1	55.1

	Positive prediction (%)		
	C_1	C_2	C_3
ACE inhibitor	100.0	0.0	1.6
antianginal	77.8	77.8	47.1
antihypertensive	100.0	55.6	84.3
vasodilator	100.0	44.4	35.7
cardiotonic	100.0	22.2	51.4
Ca^{2+} channel blocker	0.0	72.2	8.6
platelet antiaggregatory	100.0	0.0	8.6

	B
antiacne	89.2
COX inhibitor	67.6
thromboxane	97.3
leukotriene agonist/antagonist	64.9
phospholipase inhibitor	64.9
antiinflammatory	100.0
antipsoriatic	94.6
antiallergic	83.8
IL-1 inhibitor	75.7
antiarthritic	81.1
platelet antiaggregatory	67.6

Figure 8. *In silico* functional profiles of existing drugs. The functional profiles of the 871 existing drugs are represented by the matrix of 125 rows and 23 columns. The rows correspond to the 125 functions used in the functional profiling. Each column corresponds to the functional profiles of drugs in one of the 23 clusters, as defined in Figure 6. Each component of the matrix shows the positive ratio of the corresponding cluster for the corresponding function, i.e., the percentage of member drugs in the cluster that was positive for the function in their functional profiles. The positive ratios are indicated in a gray scale, ranging from black (0%) to red (100%). The 23 clusters of the drugs are arranged in the same order as in Figure 6. The clusters characterized by the CNS action (A), the antiinflammatory effect (B), and the cardiovascular action (C) are indicated by bars below the matrix. The positive ratios (%) of the clusters A, B, and C and their subclusters $A_{2,1}$, $A_{2,2}$, C_1, C_2, and C_3 (as defined in Figure 6) are shown for some functions that characterized these clusters.

Table 3. Relevancies of the Functions to the Human Liver Adverse Effects[a]

function	composite activity	alkaline phosphatase increase	SGOT increase	SGPT increase	LDH increase	GGT increase
HMG-CoA reductase inhibitor	27.0	4.756	3.273	3.776	9.364	27.0
acetylcholinesterase inhibitor	4.5	2.378	2.182	0.944	2.341	4.5
TNF inhibitor	3.6	3.805	0.655	1.510	1.873	3.6
K^+ channel activator	3.0	1.585	1.091	1.259	3.121	3.0
carbonic anhydrase inhibitor	2.25	1.189	0.818	0.944	4.682	2.25
tyrosine kinase inhibitor	2.25	1.189	1.636	1.888	0.0	2.25
cAMP phosphodiesterase inhibitor	2.25	3.171	1.636	2.517	4.682	2.25
squalene synthetase inhibitor	2.25	1.585	0.818	0.944	0.669	2.25
Ca^{2+} channel blocker	2.05	1.081	1.206	0.899	0.814	2.045
H^+/K^+ ATPase inhibitor	2.0	1.057	0.727	1.416	0.936	2.0
reverse transcriptase inhibitor	1.929	2.594	2.909	2.643	3.902	1.929
muscarinic agonist/antagonist	1.5	2.378	1.091	1.259	1.561	1.5
dopamine agonist/antagonist	1.421	1.057	2.104	2.014	0.936	1.421
COX inhibitor	0.75	2.378	1.636	1.510	1.702	0.75
adenosine agonist/antagonist	0.0	3.567	1.309	1.510	1.873	0
carbapenem	0.0	0.0	3.273	3.776	9.364	0
leukotriene agonist/antagonist	0.0	1.057	0.727	0.839	2.081	0
topoisomerase inhibitor	0.0	1.359	2.618	1.888	1.338	0

[a] The relevancies to each adverse effect were defined by eq 4. The functions are ordered by the relevancies to liver enzyme composite module activity increase.

or GGT increase, was 0.1304, while the ratio of negative prediction, $P_{i,j,\text{inactive}}$, was 0.004831. Thus, $R_{i,j}$ was 27.0, indicating that HMG-CoA inhibitors are closely related to the occurrence of composite activity increase and GGT increase. Hepatotoxicity was reported as the major complaint during therapy with HMG-CoA reductase inhibitors, along with myotoxicity.[63] Although the cellular mechanisms underlying the liver injury are not fully understood, some hypotheses are that HMG-CoA reductase inhibitors reduce mitochondrial coenzyme Q10 in hepatocyte, and at higher concentrations, they increased DNA oxidative damage and a reduced ATP synthesis and were associated with a moderately higher degree of cell death.[64] The groups of acetylcholinesterase inhibitors, carbapenem, reverse transcriptase inhibitor, tyrosine kinase inhibitor, topoisomerase

inhibitor, and calcium channel blocker also include agents with reported hepatotoxicity. Although the 314 drugs did not include a known topoisomerase inhibitor and a known tyrosine kinase inhibitor, the relevancies of topoisomerase inhibitor and tyrosine kinase inhibitor to the liver adverse effects were detected by the use of functional profiles of the drugs. Topoisomerase inhibition by camptothecin reportedly caused inhibition of mRNA synthesis in hepatocytes and sensitized them against TNF-mediated apoptosis.[65] Thus, we successfully identified high risk functions, such as HMG-CoA reductase inhibitor causing liver-related adverse effects, by assessments based on our functional profile.

Conclusion

In the present work, a method for *in silico* functional profiling of small molecules was developed. The functional profile of each molecule was created from predictions on 125 functions. The application to existing drugs showed that the functional profiling can be useful in capturing the multifunctionality or adverse effects of small molecules. The profiling detected not only the well-known, major functions of the drugs but also other minor or potential functions that were suggested in the literature or consistent with the known functions, as shown by the examples of dexmedetomidine and midazolam. The profiling can thus be useful for indication discovery from existing drugs or, conversely, screening out molecules that have undesirable functions.

In high throughput screening, many compounds are often identified as frequent hitters: they act noncompetitively, show little relationship between structure and activity, and have poor selectivity.[66] Obvious frequent hitters were not found through the functional profiling of the existing drugs. In this study, dexmedetomidine and midazolam were predicted to have the most diverse functions among the 871 existing drugs, and many of these functions were not annotated in MDDR. By investigating the functions of the drugs reported so far, the majority of the functions that appeared in the functional profiles of dexmedetomidine and midazolam were their known functions or the functions of structurally similar molecules. This suggested that dexmedetomidine and midazolam have many functions indeed and are not so-called frequent hitters. Frequent hitters could appear if nondrug-like molecules were profiled. The predicted functions exemplified by dexmedetomidine and midazolam can be promising candidates for indication discovery.

The cluster analysis of the functional profiles of existing drugs showed that the differences in the mechanisms of action can be recognized by comparing the functional profiles. For example, the group of NMDA receptor inhibitors was separated into two groups, "antiischemic" and "nonantiischemic". Analyses of both groups suggested that releasing interactions with histamine receptors are important to develop antiischemic drugs. The consideration of interactions with histamine receptors in addition to the NMDA receptor would contribute to efficiently identify promising chemical series at the early screening stage. Thus, functional profiling would make it possible to search a database for molecules with desirable functional mechanisms. The *in silico* screening using functional profiling can provide a more focused and promising set of drug candidates than the conventional *in silico* screening strategy for only one target, from the viewpoint of target specificity and toxicity.

The comparison of the functional profiles and the liver adverse effect data has led us to the detection of proteins such as HMG-CoA reductase, which is known to cause hepatic injuries. Hepatotoxicity involves many factors and, therefore, is caused by various molecular mechanisms, including unknown ones. The relevancies to hepatotoxicity of topoisomerase inhibitor and tyrosine kinase inhibitor, for which 314 drugs used in the analysis lacked annotations in MDDR, were detected using functional profiling. It suggests that the profiling in terms of various functions could be useful to deal with the functions for which only a limited amount of annotation data is available.

In this work, our functional profiling was based on the set of 125 functions, for which sufficient numbers of drugs were registered in the MDDR database. Obviously, if more diverse functions were taken into account using more comprehensive databases, then the functional profiling would be able to recognize the desirable functional profile of each drug class more comprehensively and accurately.

Acknowledgment. The authors thank Ryoichi Minai and Toshiharu Kawajiri for discussions and Hiroshi Hirota for preparation of the manuscript. This work was supported by the RIKEN Structural Genomics/Proteomics Initiative (RSGI), the National Project on Protein Structural and Functional Analyses, and the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

Supporting Information Available: The structures and CAS numbers of the five most positively predicted compounds, which were not annotated in MDDR for 70 molecular actions and 55 therapeutic areas. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: Definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
(2) Willett, P.; Winterman, V.; Bawden, D. Implementation of nearest-neighbor searching in an online chemical structure search system. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 36–41.
(3) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.
(4) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-hopping" by topological pharmacophore search: A contribution to virtual screening. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 2894–2896.
(5) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391–405.
(6) Vapnik, V. *The nature of statistical learning theory*; Springer-Verlag: New York, 1995.
(7) 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.
(8) Takaoka, Y.; Endo, Y.; Yamanobe, S.; Kakinuma, H.; Okubo, T.; Shimazaki, Y.; Ota, T.; Sumiya, S.; Yoshikawa, K. Development of a method for evaluating drug-likeness and ease of synthesis using a data set in which compounds are assigned scores based on chemists' intuition. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1269–1275.
(9) 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.
(10) Yap, C. W.; Cai, C. Z.; Xue, Y.; Chen, Y. Z. Prediction of torsade-causing potential of drugs by support vector machine approach. *Toxicol. Sci.* **2004**, *79*, 170–177.
(11) Jorissen, R. N.; Gilson, M. K. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **2005**, *45*, 549–561.
(12) Muller, 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.
(13) Yap, C. W.; Chen, Y. Z. Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *J. Chem. Inf. Model.* **2005**, *45*, 982–992.
(14) Byvatov, E.; Schneider, G. SVM-based feature selection for characterization of focused compound collections. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 993–999.
(15) Huang, S. Rational drug discovery: what can we learn from regulatory networks. *Drug Discovery Today* **2002**, *7*, S163–S169.
(16) Ekins, S. Predicting undesirable drug interactions with promiscuous proteins in silico. *Drug Discovery Today* **2004**, *9*, 276–285.

(17) MacDonald, M. L.; Lamerdin, J.; Owens, S.; Keon, B. H.; Bilter, G. K.; Shang, Z.; Huang, Z.; Yu, H.; Dias, J.; Minami, T.; Michnick, S. W.; Westwick, J. K. Identifying off-target effects and hidden phenotypes of drugs in human cells. *Nat. Chem. Biol.* **2006**, *2*, 329–337.
(18) Remick, R. A. Anticholinergic side effects of tricyclic antidepressants and their management. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **1988**, *12*, 225–231.
(19) Fava, M. Weight gain and antidepressants. *J. Clin. Psychiatry* **2000**, *61* (Suppl. 11), 37–41.
(20) Thanacoody, H. K.; Thomas, S. H. Tricyclic antidepressant poisoning: cardiovascular toxicity. *Toxicol. Rev.* **2005**, *24*, 205–214.
(21) Strachan, R. T.; Ferrara, G.; Roth, B. L. Screening the receptorome: an efficient approach for drug discovery and target validation. *Drug Discovery Today* **2006**, *11*, 708–716.
(22) Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Biological spectra analysis: Linking biological activity profiles to molecular structure. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 261–266.
(23) Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Biospectra analysis: model proteome characterizations for linking molecular structure and biological response. *J. Med. Chem.* **2005**, *48*, 6918–6925.
(24) Muller, P.; Lena, G.; Boilard, E.; Bezzine, S.; Lambeau, G.; Guichard, G.; Rognan, D. In silico-guided target identification of a scaffold-focused library: 1,3,5-Triazepan-2,6-diones as novel phospholipase A2 inhibitors. *J. Med. Chem.* **2006**, *49*, 6768–6778.
(25) Chong, C. R.; Sullivan, D. J., Jr. New uses for old drugs. *Nature* **2007**, *448*, 645–646.
(26) O'Connor, K. A.; Roth, B. L. Finding new tricks for old drugs: an efficient route for public-sector drug discovery. *Nat. Rev. Drug Discovery* **2005**, *4*, 1005–1014.
(27) Roth, B. L.; Sheffler, D. J.; Kroeze, W. K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discovery* **2004**, *3*, 353–359.
(28) Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Analysis of drug-induced effect patterns to link structure and side effects of medicines. *Nat. Chem. Biol.* **2005**, *1*, 389–397.
(29) Fliri, A. F.; Loging, W. T.; Volkmann, R. A. Analysis of system structure-function relationships. *ChemMedChem* **2007**, *2*, 1774–1782.
(30) Bender, A.; Scheiber, J.; Glick, M.; Davies, J. W.; Azzaoui, K.; Hamon, J.; Urban, L.; Whitebread, S.; Jenkins, J. L. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem* **2007**, *2*, 861–873.
(31) Kanehisa, M. A database for post-genome analysis. *Trends Genet.* **1997**, *13*, 375–376.
(32) Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K. F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; Hirakawa, M. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **2006**, *34*, D354–D357.
(33) Bush, B. L.; Sheridan, R. P. PATTY: A programmable atom typer and language for automatic classification of atoms in molecular databases. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756–762.
(34) Cristianini, N. and Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press: Cambridge, 2000.
(35) Stokes, M., Davis, C., and Koch, G. Observer agreement. In *Categorical Data Analysis Using the SAS System*, 2nd ed.; Stokes, M. E., Davis, C. S., Koch, G. G., Eds.; SAS Institute, Cary, NC, 1995; pp 98−102.
(36) Joachims, T. Making large-scale SVM learning practical. *Advances in Kernel Methods−Support Vector Learning*; MIT Press: New York, 1999.
(37) Matthews, E. J.; Kruhlak, N. L.; Weaver, J. L.; Benz, R. D.; Contrera, J. F. Assessment of the health effects of chemicals in humans: II. Construction of an adverse effects database for QSAR modeling. *Curr. Drug Discovery Technol.* **2004**, *1*, 243–254.
(38) Heesen, M.; Dietrich, G. V.; Detsch, O.; Drevermann, J.; Boldt, J.; Hempelmann, G. The in vitro effect of alpha-2 agonists on thrombocyte function and density of thrombocyte alpha-2 receptors. *Anaesthesist* **1996**, *45*, 255–258.
(39) Savola, J. M.; Virtanen, R. Central alpha 2-adrenoceptors are highly stereoselective for dexmedetomidine, the dextro enantiomer of medetomidine. *Eur. J. Pharmacol.* **1991**, *195*, 193–199.
(40) Curtis-Prior, P. B.; Tan, S. Application of agents active at the alpha 2-adrenoceptor of fat cells to the treatment of obesity−A critical appraisal. *Int. J. Obes.* **1984**, *8* (Suppl. 1), 201–213.
(41) Taniguchi, T.; Kidani, Y.; Kanakura, H.; Takemoto, Y.; Yamamoto, K. Effects of dexmedetomidine on mortality rate and inflammatory responses to endotoxin-induced shock in rats. *Crit. Care Med.* **2004**, *32*, 1322–1326.
(42) Horvat, A.; Momic, T.; Petrovic, S.; Nikezic, G.; Demajo, M. Selective inhibition of brain Na,K-ATPase by drugs. *Physiol. Res.* **2006**, *55*, 325–338.
(43) Tsai, C. S.; Hsu, P. C.; Huang, G. S.; Lin, T. C.; Hong, G. J.; Shih, C. M.; Li, C. Y. Midazolam attenuates adenosine diphosphate-induced P-selectin expression and platelet-leucocyte aggregation. *Eur. J. Anaesthesiol.* **2004**, *21*, 871–876.
(44) Kowal, K.; Pampuch, A.; Kowal-Bielecka, O.; DuBuske, L. M.; Bodzenta-Lukaszyk, A. Platelet activation in allergic asthma patients during allergen challenge with Dermatophagoides pteronyssinus. *Clin. Exp. Allergy* **2006**, *36*, 426–432.
(45) Wang, F.; Wang, N. S.; Yan, C. G.; Li, J. H.; Tang, L. Q. The significance of platelet activation in rheumatoid arthritis. *Clin. Rheumatol.* **2007**, *26*, 768–771.
(46) Mallmann, P.; Nadstawek, J.; Lauven, P. M.; Koenig, A. Changes in immunological parameters in vitro and in vivo due to midazolam. *Anaesth. Intensivther., Notfallmed.* **1988**, *23*, 141–144.
(47) Helmy, S. A.; Al-Attiyah, R. J. The immunomodulatory effects of prolonged intravenous infusion of propofol versus midazolam in critically ill surgical patients. *Anaesthesia* **2001**, *56*, 4–8.
(48) Massoco, C.; Palermo-Neto, J. Effects of midazolam on equine innate immune response: A flow cytometric study. *Vet. Immunol. Immunopathol.* **2003**, *95*, 11–19.
(49) Antoniou, T.; Tseng, A. L. Interactions between recreational drugs and antiretroviral agents. *Ann. Pharmacother.* **2002**, *36*, 1598–1613.
(50) Zhou, S.; Chan, E.; Lim, L. Y.; Boelsterli, U. A.; Li, S. C.; Wang, J.; Zhang, Q.; Huang, M.; Xu, A. Therapeutic drugs that behave as mechanism-based inhibitors of cytochrome P450 3A4. *Curr. Drug Metab.* **2004**, *5*, 415–442.
(51) Robertson, D. G.; Reily, M. D.; Albassam, M.; Dethloff, L. A. Metabonomic assessment of vasculitis in rats. *Cardiovasc. Toxicol.* **2001**, *1*, 7–19.
(52) Aoki, M.; Kobayashi, M.; Ishikawa, J.; Saita, Y.; Terai, Y.; Takayama, K.; Miyata, K.; Yamada, T. A novel phosphodiesterase type 4 inhibitor, YM976 (4-(3-chlorophenyl)-1,7-diethylpyrido[2,3-d]pyrimidin-2(1H)-one), with little emetogenic activity. *J. Pharmacol. Exp. Ther.* **2000**, *295*, 255–260.
(53) Tsunoda, H.; Sakuma, Y.; Harada, K.; Muramoto, K.; Katayama, S.; Horie, T.; Shimomura, N.; Clark, R.; Miyazawa, S.; Okano, K.; et al. Effects of a novel PAF antagonist, E6123, on PAF-induced biological responses. *Agents Actions Suppl.* **1990**, *31*, 251–254.
(54) Sakuma, Y.; Tsunoda, H.; Katayama, S.; Harada, K.; Obaishi, H.; Shirato, M.; Yamada, K.; Miyazawa, S.; Okano, K.; Machida, Y.; et al. Effects of a novel PAF antagonist, E6123, on passive anaphylaxis. *Agents Actions Suppl.* **1990**, *31*, 255–258.
(55) Weber, A.; Casini, A.; Heine, A.; Kuhn, D.; Supuran, C. T.; Scozzafava, A.; Klebe, G. Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. *J. Med. Chem.* **2004**, *47*, 550–557.
(56) R Development Core Team R: A language and environment for statistical computing; R Foundation for Statistical Computing: Vienna, Austria, 2007 (http://www.R-project.org).
(57) SciTegic Inc. and Accelrys Inc., Pipeline Pilot (TM), San Diego (http://accelrys.com).
(58) Arundine, M.; Tymianski, M. Molecular mechanisms of glutamate-dependent neurodegeneration in ischemia and traumatic brain injury. *Cell. Mol. Life Sci.* **2004**, *61*, 657–668.
(59) Lawson, K. Tricyclic antidepressants and fibromyalgia: What is the mechanism of action. *Expert Opin. Invest. Drugs* **2002**, *11*, 1437–1445.
(60) Leurs, R.; Bakker, R. A.; Timmerman, H.; de Esch, I. J. The histamine H3 receptor: From gene cloning to H3 receptor drugs. *Nat. Rev. Drug Discovery* **2005**, *4*, 107–120.
(61) Adachi, N. Cerebral ischemia and brain histamine. *Brain Res. Brain Res. Rev.* **2005**, *50*, 275–286.
(62) Ries, R. K.; Gilbert, D. A.; Katon, W. Tricyclic antidepressant therapy for peptic ulcer disease. *Arch. Intern. Med.* **1984**, *144*, 566–569.
(63) Kubota, T.; Fujisaki, K.; Itoh, Y.; Yano, T.; Sendo, T.; Oishi, R. Apoptotic injury in cultured human hepatocytes induced by HMG-CoA reductase inhibitors. *Biochem. Pharmacol.* **2004**, *67*, 2175–2186.
(64) Tavintharan, S.; Ong, C. N.; Jeyaseelan, K.; Sivakumar, M.; Lim, S. C.; Sum, C. F. Reduced mitochondrial coenzyme Q10 levels in HepG2 cells treated with high-dose simvastatin: a possible role in statin-induced hepatotoxicity. *Toxicol. Appl. Pharmacol.* **2007**, *223*, 173–179.
(65) Hentze, H.; Latta, M.; Kunstle, G.; Dhakshinamoorthy, S.; Ng, P. Y.; Porter, A. G.; Wendel, A. Topoisomerase inhibitor camptothecin sensitizes mouse hepatocytes in vitro and in vivo to TNF-mediated apoptosis. *Hepatology* **2004**, *39*, 1311–1320.
(66) McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem.* **2002**, *45*, 1712–1722.

JM800504Q