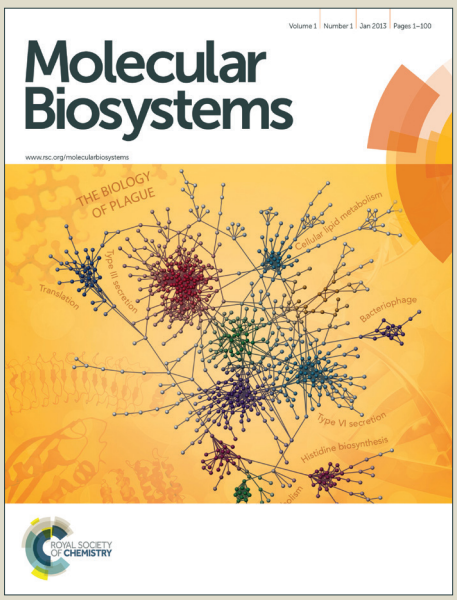
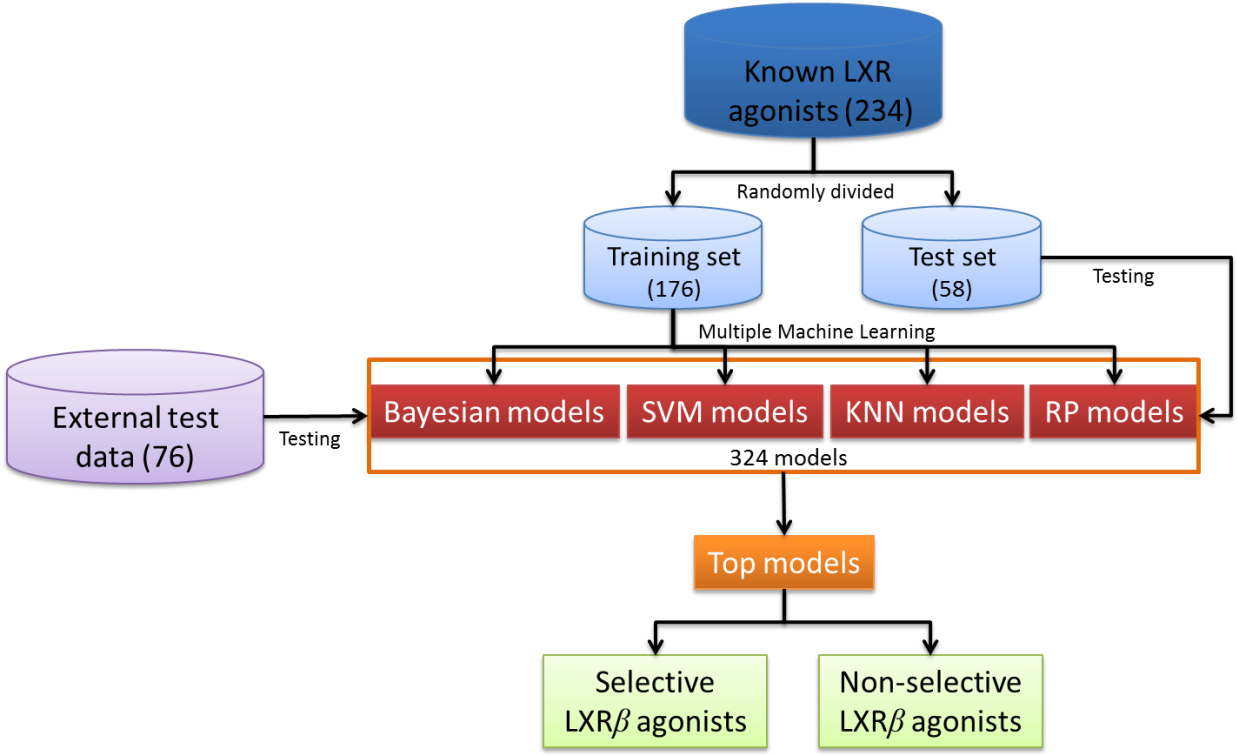
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems



The classification models for predicting selective LXRβ agonists were firstly established using multiple machine learning methods. The top models can predict selective LXRβ agonists with chemical structure diversity.

Predicting Selective Liver X Receptor β Agonists Using Multiple Machine Learning Methods

Yali Li,[a] Ling Wang,[b] Zhihong Liu,[a] Chanjuan Li,[a] Jiake Xu, [c] Qiong Gu[a] and Jun Xu*[a, d]

Abstract

Liver X receptor (LXR) α and β are cholesterol sensors; they respond to excess cholesterol and stimulate reverse cholesterol transport. Activating LXRs represents a promising therapeutic option for dyslipidemia. However, activating LXRα may cause unwanted lipogenicity. A better anti-dyslipidemia strategy would be to develop selective LXRβ agonists that do not activate LXRα. In this paper, a data set of 234 selective and non-selective LXRβ agonists was collected from the literatures. For the first time, we derived the classification models from the data set to predict selective LXRβ agonists using multiple machine learning methods (naïve Bayesian (NB), Recursive Partitioning (RP), Support Vector Machine (SVM), and *k*-Nearest Neighbors (kNN) methods) with optimized property descriptors and structural fingerprints. The models were optimized from 324 multiple machine learning models, and most of the models showed high predictive abilities (overall predictive accuracies > 80%) for both training and test sets. The top 15 models were evaluated using an external test set of 76 compounds (all containing new scaffolds), and 10 of them displayed overall predictive accuracies exceeding 90%. The top models can be used for the virtual screening selective LXRβ agonists. The NB models can identify privileged and unprivileged fragments for selective LXRβ agonists, and the fragments can be used to guide the design of new selective LXRβ agonists.

1 Introduction

Cardiovascular diseases are the leading causes of death worldwide, and one major risk factor associated with these diseases is hyperlipidemia.[1, 2] Hyperlipidemia is characterized by increased plasma cholesterol, triglycerides (TG) and decreased high-density lipoprotein (HDL) level. Lipid-lowering drugs represent the primary treatment strategy for hyperlipidemia. However, the current drugs used to treat dyslipidemia (e.g.,

HMG-CoA reductase inhibitors (stains), fibrates, and bile acid-sequestering resins) simultaneously cause liver steatosis or hypertriglyceridemia.[3, 4] Thus, discovering new anti-lipemic agents without side effects is urgently needed. LXRs are cholesterol sensors that protect cells from cholesterol overload.[5, 6] Activating LXRs can stimulate reverse cholesterol transport and inhibit its absorption, synthesis, uptake, and promote HDL formation.[6] Increasing reports suggest that LXRs are promising therapeutic targets for dyslipidemia.[3, 6, 7]

The LXR nuclear receptor family consists of two subtypes, LXRα (NR1H3) and LXRβ (NR1H2).[7] LXRα is expressed predominately in some tissues, including the liver, kidney, macrophages, and adipose tissue; however, LXRβ is ubiquitously expressed.[8] Activating LXRα (mainly expressed in liver) results in high triglyceride production[3], and growing evidence suggests that selective LXRβ agonists can reduce this side effect.[9, 10] The sequences of LXRα and LXRβ share approximately 78% identity, with little differences in their ligand binding domains (LBD).[11] Therefore, it can be more challenging to design selective LXRβ agonists using structure-based approach.[12]

To date, selective LXRβ agonists are assessed experimentally *in vitro* and *vivo* (e.g., scintillation proximity assay or transactivation assays against both LXRα and LXRβ).[13-19] However, these experimental assays are time-consuming, expensive and laborious. For example, the SPA assay involves handling radioisotopes, which are costly and low throughput.[20] Therefore, the development of computational methods that provide a rapid and efficient screening platform to predict selective LXRβ agonists is vital for the early stages of lead discovery or optimization.

Several computational pharmacophore and QSAR models predicting LXRβ agonists have been reported.[21-23] For examples, Zhao and co-workers described three-dimensional pharmacophore models to predict LXRβ agonists.[21] Salum and coworkers predicted selective LXRβ agonists using a fragment-based QSAR method.[22] Most recently, Temml and coworkers discovered LXRβ agonists using a pharmacophore modeling approach.[24] However, these models were unable to distinguish selective LXRβ agonists from non-selective agonists. Salum's models predicted selective LXRβ agonists for a specific

scaffold. Thus, there is a need for models based on various known LXRβ agonist scaffolds to predict selective LXRβ agonists with broader chemical structure diversity.

To develop models to predict selective LXRβ agonists with new chemical scaffolds, a data set of 234 structurally diverse, selective and non-selective LXRβ agonists was collected from literatures. Then, we employed multiple machine learning methods (naïve Bayesian (NB), Recursive Partitioning (RP), Support Vector Machine (SVM), and *k*-Nearest Neighbors (KNN)) to build classification models for predicting selective LXRβ agonists based on the data set. Finally, we selected the top models to discriminate selective LXRβ agonists from non-selective agonists. The flowchart of the process is depicted in Fig. 1.

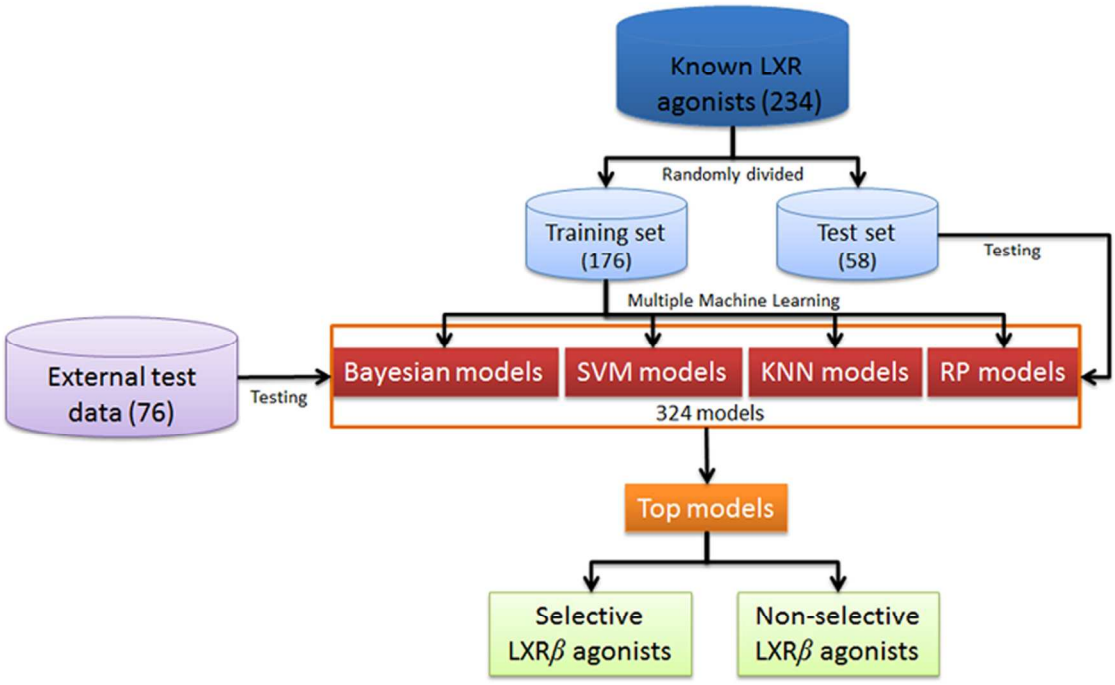


Fig. 1 The flowchart for generating models to predict selective LXRβ agonists using multiple machine learning methods.

2 Materials and methods

2.1 Data preparation

LXRs agonists were collected from the literatures[12-19, 25-38] based on following criteria: (1) the compound should be tested in the LXRs scintillation proximity assay (SPA); (2) the compound should have SPA IC_{50} values for both LXRα and LXRβ subtypes; and (3) duplicate data were removed. The whole LXR data set consisted of 391 structurally diverse compounds. A selective ratio (SR) was calculated using the following equation:

$$SR = \frac{IC_{50}(LXR\alpha)}{IC_{50}(LXR\beta)} \quad (1)$$

A compound was considered to be a selective LXRβ agonist if its SR was equal or greater than 10 (≥10). Compounds with IC_{50} values exceeding 10 μM for both LXR subtypes were removed. Some compounds were also removed due to large IC_{50} variation resulting from different measurement conditions or labs. A compound was considered to be non-selective if its SR was less than 4 (≤3). Ultimately, the data set for building the predictive models for selective LXRβ agonists contained 234 compounds.

Chemical structures of the data set were processed in two steps: (1) removing the counter ions, solvent moieties, and salts in the structures; and (2) optimizing the 2D conformations of the structures through energy minimizations with the MMFF94 force field (MOE version 2013.08, Chemical Computing Group Inc., Canada).

The structural diversity of the data set was analyzed using the S-cluster algorithm.[39] In the data set, selective LXRβ agonists were marked as "1", and non-selective agonists were marked as "0". The data set was divided into a training set (176 compounds) and test set (58 compounds) using randomized algorithm in DS 3.5 (Discovery Studio version 3.5, Accelrys Inc., USA). The ratio of the number of compounds in the training set and the number of compounds in the testing set was 3:1.[40] The data set is available in the Electronic Supplementary Information.

2.2 Molecular descriptor calculation

Molecular descriptors were calculated using MOE and PaDEL-Descriptor software.[41] A total of 192 two-dimensional molecular descriptors were generated from MOE, and 770 one- and two-dimensional descriptors were calculated using the PaDEL-Descriptor program.

2.3 Molecular descriptor selection

Pearson correlation analysis was employed to exclude redundant descriptors and descriptors unrelated to activity.[42-44] In the present work, descriptors whose Pearson correlation coefficients with the SR were less than 0.1, or descriptors whose correlation coefficients with other descriptors were higher than 0.9 were removed. Finally, 12 (derived with MOE) and 14 property descriptors (derived from PaDEL-Descriptors program) were used for the modeling studies (Table 1).

Table 1 Molecular descriptors selected for the modeling studies

Program	No. of descriptors	Descriptor list
MOE	12	BCUT_PEOE_0, BCUT_PEOE_1, BCUT_SLOGP_0, BCUT_SLOGP_1, BCUT_SMR_1, GCUT_PEOE_2, GCUT_SMR_0, GCUT_SMR_1, GCUT_SMR_2, logP(o/w), PEOE_VSA+0, Q_VSA_FNEG
PaDEL	14	VC-4, VC-6, SPC-4, SwHBa, SHCsats, SHother, SssCH2, ETA_Shape_Y, ETA_dBetaP, nAtomP, MDEC-24, MDEC-33, MDEC-34, MLFER_A

2.4 Structural fingerprint calculation

Four types of structural fingerprints[45] (EState, MACCS, Substructure, and Substructure Fingerprint Count) were calculated using the PaDEL-Descriptor program. These structural fingerprints were successfully used to predict toxicity and biodegradability.[46, 47] The ECFC_4 fingerprints were calculated using DS 3.5.

2.5 Modeling methods

The following machine learning methods[48-50] were employed for modeling: naïve Bayesian (NB), Recursive Partitioning (RP), Support Vector Machine (SVM) and *k*-Nearest Neighbors (kNN) methods.

2.5.1 Naïve Bayesian

A NB model is generated using prior evidence that an object belongs to a certain class; for example, an active class or inactive class from a training data set. In present work, we used DS 3.5 to build NB models. Topological descriptors and fingerprints were used to describe the properties of chemical structures. A Bayesian model classifies compounds by confirming the frequency at which an attribute appears.[51-53] The following equation represents the Bayesian law:

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)} \quad (2)$$

where A represents an attribute; B represents a compound's class; $P(B|A)$ refers to the posterior probability of a compound that belongs to a certain class; $P(A|B)$ is the probability that the compound belonging to a certain class (in our case, selective or non-selective) has certain attributes; P(B) is the prior probability resulting from the training set; and P(A) is the marginal probability that the attribute appears in the training set. The three probabilities on the right side of formula (2) can be derived from the training set.[54, 55]

2.5.2 Recursive Partitioning

The RP method recursively splits a data set into smaller subsets, and it generates a hierarchical tree called a decision tree, which represents relationships among data points and independent properties (in our case, molecular descriptors and fingerprints).[56, 57] Our RP models were built using an RP module from DS 3.5. Tree depths ranged from 2 to 10 to acquire the best predictive performance.

2.5.3 Support Vector Machine

The SVM method employs the structural risk minimization principle to reduce generalization error on the training data and avoid over-fitting effects.[58, 59] Non-linear SVM simplifies the classification problem by transforming a data space into a higher dimensional feature space.[58, 60] To determine a hyper-plane to divide a data set into two classes, SVM models were constructed using the LIBSVM 3.18 package developed by Chang and Lin.[61] The SVM models were built using a radial basis function (RBF) kernel.

An auto-searching program, "grid.py", was used to select the parameters of the SVM (c, g), and every SVM model was validated using the 5-fold cross-validation.

2.5.4 *k*-Nearest Neighbor

The kNN models were built using Orange 2.7 (http://www.ailab.si/orange/). The kNN is a non-parametric instance-based learning that classifies objects based on the closest training samples in a feature space. An object is classified by assigning the most frequent class among the *k* training samples nearest to that object.[62, 63] The performance of kNN models largely depends on the original data set. In this study, the parameter *k* was changed from 1 to 10 to determinate the nearest neighbors.

2.6 Model performance validation

To validate the accuracy and robustness of the models, we employed a 5-fold cross-validation scheme. True positives (TP), true negatives (TN), false positives (FP), false negatives (FN), sensitivity (SE), specificity (SP), overall predictive accuracy (Q) and the Matthews correlation coefficient (MCC) were calculated using the following equations:

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{4}$$

$$\mathrm{Q} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{5}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})}} \tag{6}$$

TP and TN represent the numbers of selective LXRβ agonists and non-selective LXRβ agonists that are correctly predicted, respectively; FP represents the number of non-selective agonists that are mistaken for selective LXRβ agonists; FN stands for the number of selective LXRβ agonists that are predicted to be non-selective LXRβ agonists; SE represents the predictive accuracy for selective LXRβ agonists; and SP represents the

predictive accuracy for non-selective LXRβ agonists. The MCC is the foremost indicator for evaluating models.[64]

3 Results and discussion

3.1 Selecting property descriptors

192 descriptors were computed for all the compounds in the LXR agonists' data set using MOE. Based on Pearson correlation analyses, we removed redundant descriptors, leaving 12 descriptors that were correlated with the LXRβ selectivity measurements (SR values). Using the same protocol, 14 property descriptors were selected from 770 descriptors in PaDEL. Detailed results are available in the Electronic Supplementary Information.

3.2 Correlations between property descriptors and LXR agonist binding affinities

The correlation coefficients of R_1(LXRα), and R_2(LXRβ) between the 12 selected descriptors and binding affinities of LXRs agonists (convert to pIC50) were computed as listed in Table 2. The statistical significances (*p*-values) between average values of selective and non-selective LXRβ agonists for the descriptors were computed via Student's *t*-tests (Table 2). The *p*-values indicate that most descriptors (except logP(o/w)) are significantly different between selective and non-selective LXRβ agonists. This result is consistent with the correlation analysis results for both LXRα and LXRβ assay activities. For instance, the BCUT_PEOE_0 averages of non-selective LXRβ agonists and selective LXRβ agonists are -2.636 and -2.465, respectively. The *p*-value for BCUT_PEOE_0 is 5.190×10^{-15}, which means statistically significant difference. This result is consistent with the correlation analysis results for both LXRα and LXRβ assay activities (Table 2). As shown in Table 2, BCUT_PEOE_0 has a better correlation with LXRβ assay activity (R_2 = 0.344), whereas it exhibits a lower correlation with LXRα assay activity (R_1 = 0.067).

The *p*-value for logP(o/w) is 0.022 (Table 2), indicating that the difference of average logP(o/w) values between selective and non-selective LXRβ agonists is not significant. However, logP(o/w), an indicator of a compound's hydrophobicity, is highly correlated with both LXRα and LXRβ assay activities (R_1 = 0.195, R_2 = 0.284, see Table 2). The logP(o/w) averages for LXRα agonists and LXRβ agonists are 6.305 and 6.357,

suggesting that logP(o/w) is almost equally important to both LXRα and LXRβ agonists. The larger average logP(o/w) values for both LXRα and LXRβ agonists indicate the LXRs agonists should be hydrophobic molecules so as to form stronger hydrophobic interactions with the LXRα/β binding pockets. Our analysis results are consistent with the computational and experimental results.[21, 23, 65-67]. Therefore, we cannot build a predictive model for selective LXRβ agonist without logP(o/w).

The capacities to discriminate selective LXRβ agonists from non-selective LXRβ agonists for the 12 descriptors are depicted in Fig. 2. No descriptor could perfectly discriminate the two classes of agonists; thus, all of the selected descriptors must be taken into account for the modeling, and multiple modeling approaches must be used to identify the best combination of descriptors to achieve maximal performance.

Table 2 Correlation coefficients and *p*-values for the binding affinities of LXR agonists and descriptors derived from the LXR agonist data set.

Descriptor	R_1(LXRα)*	R_2(LXRβ)**	*p*-value***
BCUT_PEOE_0	0.067	0.344	5.190×10^{-15}
BCUT_PEOE_1	0.068	-0.202	5.375×10^{-9}
BCUT_SLOGP_0	0.067	0.323	1.896×10^{-11}
BCUT_SLOGP_1	0.080	-0.122	1.844×10^{-6}
BCUT_SMR_1	0.105	0.147	6.075×10^{-9}
GCUT_PEOE_2	-0.120	0.114	2.479×10^{-9}
GCUT_SMR_0	-0.056	0.285	3.749×10^{-13}
GCUT_SMR_1	0.128	-0.061	4.161×10^{-4}

GCUT_SMR_2	-0.058	0.242	9.485×10^{-12}
logP(o/w)	0.195	0.284	0.022
PEOE_VSA+0	0.057	-0.255	9.00×10^{-12}
Q_VSA_FNEG	-0.067	0.301	2.058×10^{-10}

* R_1(LXRα) represents the correlation coefficient between a descriptor and IC_{50}(LXRα).
** R_2(LXRβ) represents the correlation coefficient between a descriptor and IC_{50}(LXRβ).
*** *p*-value represents the statistical significance between average values of selective and non-selective LXRβ agonists for a descriptor.

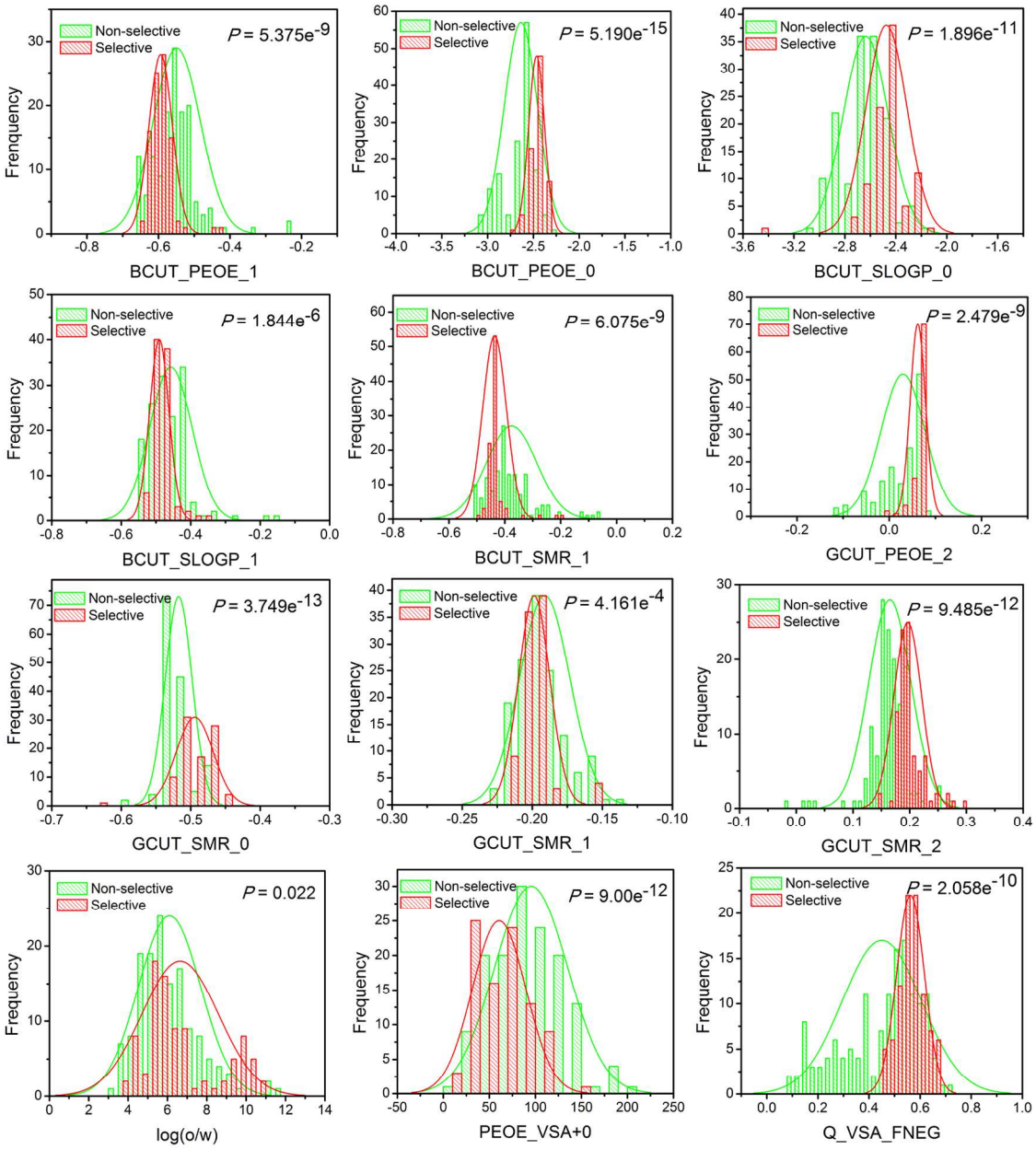


Fig. 2 Bar charts indicate the capacities to discriminate selective LXRβ agonists from non-selective LXRβ for the 12 descriptors.

3.3 Determining the SR threshold to identify selective LXRβ agonists

To determine the best SR threshold to distinguish selective LXRβ agonist from other compounds, a number of SR thresholds (5, 10, 15, and 20) were trailed with a SVM

model based on MOE descriptors (Fig. 3). The best SR was 10, and it was selected based on the maximal MCC value for the test set.

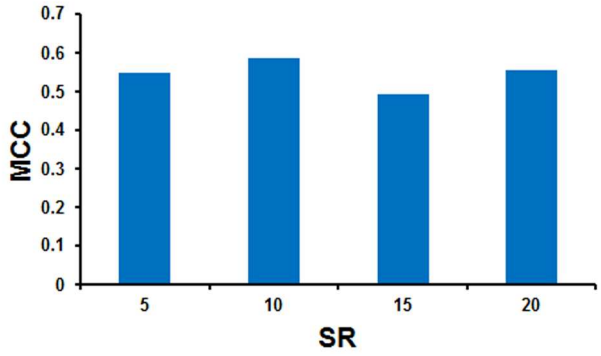


Fig. 3 MCC changes when different SR thresholds were applied to select selective LXRβ agonists.

3.4 Determining the SR threshold to identify non-selective LXRβ agonists

The smaller SR values (i.e., 1~9) associated with some compounds were due to different measuring conditions or system errors from different labs; thus, these results were considered to be suspicious and were removed. One way to remove these data is to identify a SR threshold using a predictive model or directly defining the SR threshold, as reported previously.[68-70] The best SR threshold for defining non-selective LXRβ agonists was determined based on the performance of SVM model using MOE descriptors with a set of SR thresholds (1~9). The results indicated that 3 was the best SR threshold for removing suspicious LXR agonists, and this threshold resulted in maximal predictive performance for test set (Fig. 4).

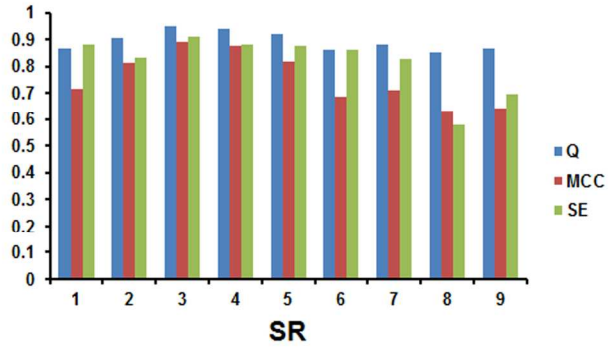


Fig. 4 The SVM model performance changes when different SR thresholds were used to remove non-selective LXR agonists.

3.5 Performance of property descriptor-based models

Four machine learning methods (SVM, NB, RP, and kNN) were employed to build models based on optimized property descriptors (12 descriptors from MOE and 14 descriptors from PaDEL), producing 42 models. 8 models were selected based upon 5-fold cross-validation results (Table 3).

Table 3 Cross-validation results of 8 models derived from property descriptors

Model Name	Training set								Test set							
	TP	TN	FP	FN	SE	SP	Q	MCC	TP	TN	FP	FN	SE	SP	Q	MCC
kNN_MOE	59	97	10	10	0.855	0.907	0.886	0.762	18	28	8	4	0.818	0.778	0.793	0.581
kNN_PaDEL	60	98	9	9	0.870	0.916	0.898	0.785	19	30	6	3	0.864	0.833	0.845	0.683
NB_MOE	57	98	9	12	0.826	0.916	0.881	0.748	19	34	2	3	0.864	0.944	0.914	0.816
NB_PaDEL	65	91	16	4	0.942	0.850	0.886	0.776	17	35	1	5	0.772	0.972	0.897	0.781
RP_MOE	60	95	12	9	0.870	0.888	0.881	0.752	19	33	3	3	0.864	0.917	0.897	0.780
RP_PaDEL	55	102	5	14	0.797	0.953	0.892	0.773	17	36	0	5	0.773	1.000	0.914	0.824
SVM_MOE	**69**	**105**	**2**	**0**	**1.000**	**0.981**	**0.989**	**0.977**	**20**	**35**	**1**	**2**	**0.909**	**0.972**	**0.948**	**0.890**
SVM_PaDEL	68	104	3	1	0.986	0.972	0.977	0.953	20	34	2	2	0.909	0.944	0.931	0.814

NB: naïve Bayesian; RP: Recursive Partitioning; SVM: Support Vector Machine; kNN: *k*-Nearest Neighbors; MOE represents 12 descriptors from MOE software and PaDEL represents 14 descriptors calculated using PaDEL-Descriptors software.

As shown in Table 3, all models (except kNN_MOE) display overall predictive accuracies (Q) above 80% for both the training and test sets. For the test set, NB_MOE, RP_PaDEL, SVM_MOE, and SVM_PaDEL achieve overall predictive accuracies of 0.914, 0.914, 0.948 and 0.931, respectively, and their MCC values exceed 0.8. The SVM_MOE model, which displays the best MCC (0.890), Q (0.948), sensitivity (90.9%), and specificity (97.2%) for test set, is considered to be the best model. The SVM_MOE model show similar results for the training set (Table 3). SVM_PaDEL is the second best

model. Therefore, the SVM method appears to be a better classifier for our data set. Similar results were reported in other studies.[46, 47, 49]

3.6 Performance of structural fingerprint-based models

Four machine learning methods (SVM, NB, RP, and kNN) were employed to build the models based on four types of structural fingerprints (ES: EState Fingerprint; MA: MACCS Fingerprint; S: Substructure Fingerprint; SC: Substructure Fingerprint Count) generated using the PaDEL program. This approach yielded 84 models, 16 of which were selected based on 5-fold cross-validation results (ESI, Table S1).

The SVM and RP models are the best classifiers in this case. Both SVM_MA and RP_S achieve the best performance (Q=0.931, ESI, Table S1). Similar to the property descriptor-based models, four structural fingerprint-based SVM models result in better performance than other classifiers, suggesting that SVM is a better method of building a predictive model for selective LXRβ agonists.

Overall, the models derived from the structural fingerprints and property descriptors show consistent performance (Q are greater than 0.8), demonstrating that both descriptors and fingerprints are properly selected and the classification models are successfully constructed.

3.7 Performance of models based on the combinations of property descriptors and structural fingerprints

Previous reports indicated that models derived from the combinations of property descriptors and structural fingerprints showed enhanced performance.[40, 49, 54] In this work, 32 models were selected from 168 models derived from the combinations of property descriptors (computed in MOE and PaDEL) and structural fingerprints (ES, MA, S, and SC computed in PaDEL) using 5-fold cross-validation.

The performance of these 32 combinatorial models is depicted in Fig. 5.

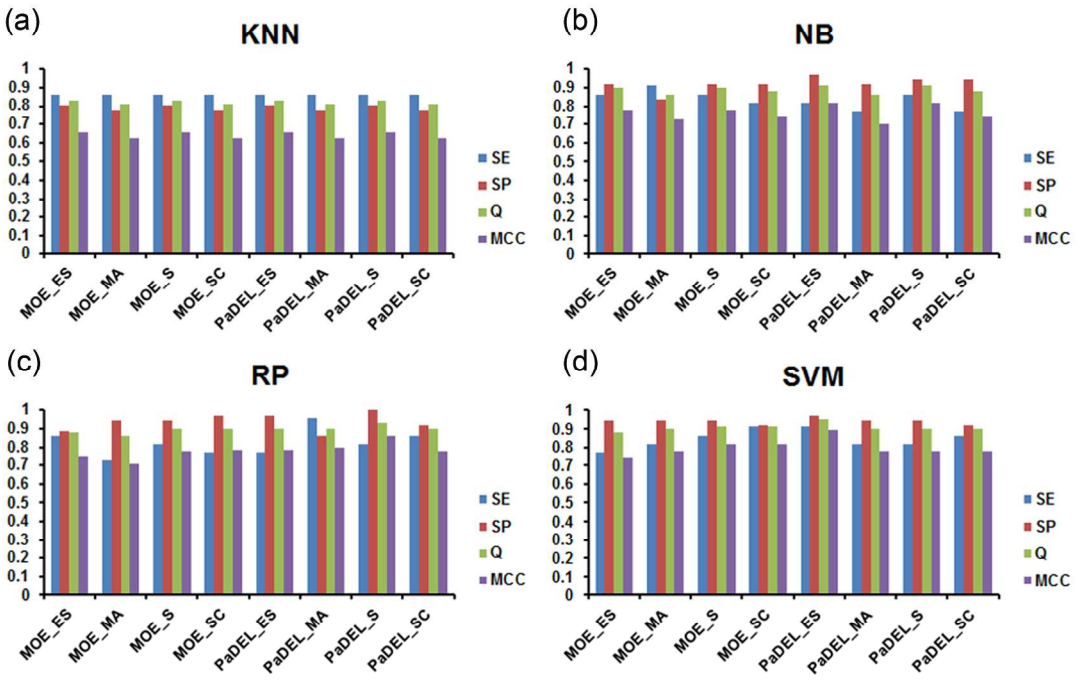


Fig. 5 The performance of 32 combinatorial models based on the combinations of two groups of property descriptors and four groups of structural fingerprints validated with the test set.

All 32 combinatorial models achieve overall predictive accuracies above 80% for both training set and test set. Overall, these models are worse classifiers than those non-combinatorial models (Fig. 5). Again, all combinatorial SVM models exhibit better performance than other combinatorial models, although the SVM models (except PaDEL_ES) exhibit worse performance than models derived from pure property descriptors or structural fingerprints.

For RP models, the performance of PaDEL_S model is slightly improved than non-combinatorial RP models and the combinatorial models (MOE_ES, PaDEL_ES, PaDEL_MA) exhibit better performance than corresponding fingerprint-based RP models. For NB models, the performance of combinatorial models (PaDEL_ES and PaDEL_S) is better than pure descriptor-based models. Most combinatorial NB models (except MOE_ES) exhibit better performance than fingerprint-based models. The performance of

other combinatorial models is not improved than pure property descriptor- or structural fingerprint-based models.

Simply combining property descriptors and structural fingerprints may make the combined descriptors overemphasize particular factors, depressing the model performance. For example, ECFC_4 fingerprints are systemically derived from compounds, whereas empirical fingerprints, such as, MACCS, are biased due to pre-defined structural fragments. When these fingerprints are combined, some structural features are overemphasized or omitted.

As shown in Table 4, there is no significant deference between a combinatorial model (for example, NB_MOE_ECFC_4) and non-combinatorial model (for example, NB_ECFC_4).

Table 4 Performance of RP and NB models based on ECFC_4* and property descriptors.

Models	Training set								Test set							
	TP	TN	FP	FN	SE	SP	Q	MCC	TP	TN	FP	FN	SE	SP	Q	MCC
NB_ECFC_4	65	97	10	4	0.942	0.907	0.920	0.838	19	34	2	3	0.864	0.944	0.914	0.816
NB_MOE_ECFC_4	65	97	10	4	0.942	0.907	0.920	0.838	19	34	2	3	0.864	0.944	0.914	0.816
NB_PaDEL_ECFC_4	65	97	10	4	0.942	0.907	0.920	0.838	19	34	2	3	0.864	0.944	0.914	0.816
RP_ECFC_4	61	97	10	8	0.884	0.907	0.898	0.787	19	34	2	3	0.864	0.944	0.914	0.816
RP_MOE_ECFC_4	58	104	3	11	0.841	0.972	0.920	0.834	15	35	1	7	0.682	0.972	0.862	0.710
RP_PaDEL_ECFC_4	59	102	5	10	0.855	0.953	0.915	0.820	19	33	3	3	0.864	0.917	0.897	0.780

* ECFC_4 represents ECFC_4 fingerprints calculated using DS 3.5.

3.8 Privileged and unprivileged fragments for selective LXR*β* agonists

The NB models allow us to determine the privileged fragments responsible for selective LXR*β* agonist activity. A set of privileged fragments for selective LXR*β* agonists was derived from NB model (ESI, Fig. S1: PS1~20) using ECFC_4 fingerprints. The Bayesian scores of the top-20 privileged fragments are greater than 0.720, suggesting that these fragments significantly improve LXR*β* agonist selectivity. The common features of these fragments are N-hetero aromatic rings or conjugated amines.

Unprivileged fragments for selective LXRβ agonists were also extracted from the NB model (ESI, Fig. S1: NS1~20) using the DS 3.5 program. Unprivileged fragments are mainly bulky groups without N-hetero aromatic groups.

The characteristics of typical privileged fragments and unprivileged fragments are listed in Fig. S1. Most privileged fragments are frequently displayed in selective LXRβ molecules, whereas most unprivileged fragments are frequently displayed in non-selective LXRβ molecules. The privileged fragment PS9 appears in the data set with a frequency of 22 among selective agonists versus a frequency of 1 among non-selective agonists (ESI, Table S2). The privileged fragments PS4 and PS7 are only encoded in the selective LXRβ agonists using the substructure search method. As shown in Fig. S1, 20 unprivileged fragments contain saturated carbon chains, and ten of contain rings. The unprivileged fragments NS3 and NS9 are only encoded in non-selective agonists. These privileged and unprivileged fragments can guide in designing new selective LXRβ agonists.

3.9 Validating the models with external test data

External test data (3 non-selective and 64 selective LXRβ agonists) was collected from Wyeth's patents[71-74]. The compounds in the external test data set contain quinazoline, pyrazolo [1,5-a] pyrimidine, and imidazo [1,2-b] pyridazine scaffolds, which are different from the scaffolds contained in the training and test sets.

Top-15 models were tested using external test data. 10 out of the 15 models have overall predictive accuracies (Q) exceeding 90%. In addition, these models exhibit consistent predictive results for the training, test, and the external test sets (even if it contained different scaffolds), suggesting that these models can be used to identify new selective LXRβ scaffolds (Table 5).

Table 5 Top-15 models validated with external test data, test data, and training data.

Models	External test set		Test set	Training set
	NCP*	Q1	Q2	Q3
SVM_MOE	53	69.74	94.82	98.86

SVM_PaDEL	72	94.74	93.10	97.72
NB_MOE	48	63.16	91.38	88.07
RP_PaDEL	68	89.47	91.38	89.21
SVM_MA	69	90.79	93.10	97.16
SVM_S	73	96.05	91.38	92.05
SVM_SC	73	96.05	91.38	92.61
RP_S	73	96.05	93.10	90.34
NB_ECFC_4	74	97.37	91.38	92.05
NB_MOE_ECFC_4	74	97.37	91.38	92.05
NB_PaDEL_ECFC_4	74	97.37	91.38	92.05
SVM_MOE_S	73	96.05	91.38	91.48
SVM_MOE_SC	73	96.05	91.38	92.61
SVM_PaDEL_ES	66	86.84	94.83	96.02
RP_PaDEL_S	68	89.47	93.10	91.48

* NCP: Number of correct predictions; Q1~3: overall predictive accuracies.

3.10 Comparisons of these classifiers

Our studies demonstrated that SVM approach could produce the best multi-descriptors based models. However, the kernel functions were difficult to select, and the parameters were hard to be optimized. A NB model was a simple probabilistic classifier based on the Bayesian theorem, scalable, and interpretable. Comparing with SVM classifier, the NB classifier was non-parametric, and resulted in confidence intervals. By means of recursive partitioning process, RP approach divided a set of objects into subsets with pre-defined parameter thresholds, and organized the subsets hierarchically. Our studies indicated that RP models and SVM models were comparable. KNN classifies were built by grouping objects (nearest neighbors) with a given similarity threshold. The similarity was calculated based upon descriptor metrics. It could have high computing complexity for a big data set. To conclude, if very significant discriminators were not found in a feature space, one may combine a set of descriptors to improve the predictivity, although each

descriptor was not very significant discriminator. SVM is a proper approach for this situation.

4 Conclusions

Here, we employed multiple machine learning methods with property descriptors and structural fingerprints to develop predictive models for selective LXRβ agonists. Although some descriptors are highly correlated with selectivity, no single descriptor is capable of discriminating selective and non-selective LXRβ agonists. A predictive model must be derived from combined descriptors or fingerprints. However, combining property descriptors and structural fingerprints cannot significantly improve the performance of models for predicting selective LXRβ agonists.

SVM is the best method for generating models for predicting selective LXRβ agonists, although other methods can also produce predictive models with similar performance.

While generating predictive models, the NB method can also produce structure fragments that contribute to the selectivity or non-selectivity of LXRβ agonists. These results may guide the design of new, selective LXRβ scaffolds.

The top-10 models demonstrated the capacity of hopping new scaffolds for selective LXRβ agonists. These models can be used as *in silico* tools for virtual screening or predicting new selective LXRβ agonists.

Funding

This work was supported by the National Science Foundation of China (81173470), National High Technology Research and Development Program of China (863 Program, 2012AA020307), National Supercomputer Center in Guangzhou (2012Y2-00048/2013Y2-00045, 201200000037), the introduction of innovative R&D team program of Guangdong Province (2009010058), Guangdong Provincial Key Laboratory of Construction Foundation (2011A060901014), and the Fundamental Research Funds

for the Central Universities (2013HGCH0015). Jiake Xu thanks for the support from the National Health and Medical Research Council of Australia.

Notes

[a]*Research Center for Drug Discovery & Institute of Human Virology, School of Pharmaceutical Sciences, Sun Yat-Sen University, Guangzhou 510006, China. Email: junxu@biochemomes.com; Fax: +86-20-3994-3023; Tel(PRC): +86-20-3994-3074*

[b]*School of Bioscience and Bioengineering, South China University of Technology, Guangzhou 510006, China*

[c]*Centre for Orthopaedic Research, School of Surgery, The University of Western Australia, Perth, WA 6009, Australia*

[d]*School of Medical Engineering, Hefei University of Technology, Hefei 230009, China*

† Electronic Supplementary Information (ESI) available.

References

1 M. Spreafico, M. Smiesko, O. Peristera, G. Rossato and A. Vedani, *Molecular Informatics*, 2010, 29, 27-36.
2 A. Berkenstam, M. Farnegardh and J. A. Gustafsson, *Mechanisms of ageing and development*, 2004, 125, 707-717.
3 E. G. Lund, J. G. Menke and C. P. Sparrow, *Arteriosclerosis, thrombosis, and vascular biology*, 2003, 23, 1169-1177.
4 J. Beltowski, *Cardiovascular therapeutics*, 2008, 26, 297-316.
5 S. Marie Ulven, K. Tomas Dalen, J. A. Gustafsson and H. Irene Nebb, *Prostaglandins, leukotrienes, and essential fatty acids*, 2005, 73, 59-63.
6 P. Tontonoz and D. J. Mangelsdorf, *Molecular endocrinology*, 2003, 17, 985-993.
7 J. R. T. Schultz, H. Luk, A. Repa, J. J. Medina, J. C. Li, L. Schwendner, S. Wang, S. Thoolen, M. Mangelsdorf, D. J. Lustig, K. J. Shan, B. , *Genes & Development*, 2000, 14, 2831-2838.
8 E. Viennois, K. Mouzat, J. Dufour, L. Morel, J. M. Lobaccaro and S. Baron, *Molecular and cellular endocrinology*, 2012, 351, 129-141.
9 E. M.Quinet, D. A.Savio, A. R.Halpern, L. Chen, G. U.Schuster, J. A. Gustafsson, M. D.Basso and P. Nambi, *Mol. Pharmacol.*, 2006, 70, 1340-1349.
10 E. G. Lund, L. B. Peterson, A. D. Adams, M. N. Lam, C. A. Burton, J. Chin, Q. Guo, S. Huang, M. Latham, J. C. Lopez, J. G. Menke, D. P. Milot, L. J. Mitnaul, S. E.

Rex-Rabe, R. L. Rosa, J. Y. Tian, S. D. Wright and C. P. Sparrow, *Biochemical pharmacology*, 2006, 71, 453-463.
11 M. Baranowski, *Journal Physiol Pharmacol*, 2008, 59, 31-55.
12 B. Hu, R. Unwalla, M. Collini, E. Quinet, I. Feingold, A. Goos-Nilsson, A. Wihelmsson, P. Nambi and J. Wrobel, *Bioorganic & medicinal chemistry*, 2009, 17, 3519-3527.
13 B. Hu, R. Bernotas, R. Unwalla, M. Collini, E. Quinet, I. Feingold, A. Goos-Nilsson, A. Wilhelmsson, P. Nambi, M. Evans and J. Wrobel, *Bioorganic & medicinal chemistry letters*, 2010, 20, 689-693.
14 B. Hu, E. Quinet, R. Unwalla, M. Collini, J. Jetter, R. Dooley, D. Andraka, L. Nogle, D. Savio, A. Halpern, A. Goos-Nilsson, A. Wilhelmsson, P. Nambi and J. Wrobel, *Bioorganic & medicinal chemistry letters*, 2008, 18, 54-59.
15 B. Hu, R. J. Unwalla, I. Goljer, J. W. Jetter, E. M. Quinet, T. J. Berrodin, M. D. Basso, I. B. Feingold, A. G. Nilsson, A. Wilhelmsson, M. J. Evans and J. E. Wrobel, *Journal of medicinal chemistry*, 2010, 53, 3296-3304.
16 H. Ratni, D. Blum-Kaelin, H. Dehmlow, P. Hartman, P. Jablonski, R. Masciadri, C. Maugeais, A. Patiny-Adam, N. Panday and M. Wright, *Bioorganic & medicinal chemistry letters*, 2009, 19, 1654-1657.
17 J. W. Szewczyk, S. Huang, J. Chin, J. Tian, L. Mitnaul, R. L. Rosa, L. Peterson, C. P. Sparrow and A. D. Adams, *Bioorganic & medicinal chemistry letters*, 2006, 16, 3055-3060.
18 V. Molteni, X. Li, J. Nabakka, F. Liang, J. Wityak, A. Koder, L. Vargas, R. Romeo, N. Mitro, P. Mak, H. Seidel, J. Haslam, D. Chow, T. Tuntland, T. Spalding, A. Brock, M. Bradley, A. Castrillo, P. Tontonoz and E. Saez, *Journal of medicinal chemistry*, 2007, 50, 4255-4259.
19 E. Chao, J. Caravella, M. Watson, N. Campobasso, S. Ghisletti, A. Billin, C. Galardi, P. Wang, B. Laffitte, M. Iannone, B. Goodwin, J. Nichols, D. Parks, E. Stewart, R. Wiethe, S. Williams, A. Smallwood, K. Pearce, C. Glass, T. Willson, W. Zuercher and J. Collins, *Journal of medicinal chemistry*, 2008, 51, 5758-5765.
20 J. Cao, Y. Zhou, H. Peng, X. Huang, S. Stahler, V. Suri, A. Qadri, T. Gareski, J. Jones, S. Hahm, M. Perreault, J. McKew, M. Shi, X. Xu, J. F. Tobin and R. E. Gimeno, *The Journal of biological chemistry*, 2011, 286, 41838-41851.
21 W. Zhao, Q. Gu, L. Wang, H. Ge, J. Li and J. Xu, *Journal of chemical information and modeling*, 2011, 51, 2147-2155.
22 L. B. Salum, A. D. Andricopulo and K. M. Honorio, *Journal of molecular graphics & modelling*, 2012, 32, 19-31.
23 S. von Grafenstein, J. Mihaly-Bison, G. Wolber, V. N. Bochkov, K. R. Liedl and D. Schuster, *Journal of chemical information and modeling*, 2012, 52, 1391-1400.
24 V. Temml, C. V.Voss, V. M.Dirsch and D. Schuster, *Journal of chemical information and modeling*, 2014, 54, 367-371.
25 R. C. Bernotas, D. H. Kaufman, R. R. Singhaus, J. Ullrich, R. Unwalla, E. Quinet, P. Nambi, A. Wilhelmsson, A. Goos-Nilsson and J. Wrobel, *Bioorganic & medicinal chemistry*, 2009, 17, 8086-8092.
26 R. C. Bernotas, R. R. Singhaus, D. H. Kaufman, J. M. Travins, J. W. Ullrich, R. Unwalla, E. Quinet, M. Evans, P. Nambi, A. Olland, B. Kauppi, A. Wilhelmsson, A.

Goos-Nilsson and J. Wrobel, *Bioorganic & medicinal chemistry letters*, 2010, 20, 209-212.
27 R. C. Bernotas, R. R. Singhaus, D. H. Kaufman, J. Ullrich, H. Fletcher, 3rd, E. Quinet, P. Nambi, R. Unwalla, A. Wilhelmsson, A. Goos-Nilsson, M. Farnegardh and J. Wrobel, *Bioorganic & medicinal chemistry*, 2009, 17, 1663-1670.
28 B. Hu, M. Collini, R. Unwalla, C. Miller, R. Singhaus, E. Quinet and D. Savio, *Journal of medicinal chemistry*, 2006, 49, 6151-6154.
29 B. Hu, J. Jetter, D. Kaufman, R. Singhaus, R. Bernotas, R. Unwalla, E. Quinet, D. Savio, A. Halpern, M. Basso, J. Keith, V. Clerin, L. Chen, Q. Y. Liu, I. Feingold, C. Huselton, F. Azam, A. Goos-Nilsson, A. Wilhelmsson, P. Nambi and J. Wrobel, *Bioorganic & medicinal chemistry*, 2007, 15, 3321-3333.
30 H. Jayasuriya, K. Herath, J. Ondeyka, Z. Guan, R. P. Borris, S. Tiwari, W. Jong, F. Chavez, J. Moss, D. Stevenson, H. Beck, M. Slattery, N. Zamora, M. Schulman, A. Ali, N. Sharma, K. Macnaul, N. Hayes, J. Menke and S. Singh, *J. Nat. Prod.*, 2005, 68, 1247-1252.
31 D. J. Kopecky, X. Y. Jiao, B. Fisher, S. McKendry, M. Labelle, D. E. Piper, P. Coward, A. K. Shiau, P. Escaron, J. Danao, A. Chai, J. Jaen and F. Kayser, *Bioorganic & medicinal chemistry letters*, 2012, 22, 2407-2410.
32 W. Liu, S. Chen, J. Dropinski, L. Colwell, M. Robins, M. Szymonifka, N. Hayes, N. Sharma, K. MacNaul, M. Hernandez, C. Burton, C. P. Sparrow, J. G. Menke and S. B. Singh, *Bioorganic & medicinal chemistry letters*, 2005, 15, 4574-4578.
33 N. Panday, J. Benz, D. Blum-Kaelin, V. Bourgeaux, H. Dehmlow, P. Hartman, B. Kuhn, H. Ratni, X. Warot and M. B. Wright, *Bioorganic & medicinal chemistry letters*, 2006, 16, 5231-5237.
34 S. B. Singh, J. G. Ondeyka, W. Liu, S. Chen, T. S. Chen, X. Li, A. Bouffard, J. Dropinski, A. B. Jones, S. McCormick, N. Hayes, J. Wang, N. Sharma, K. Macnaul, M. Hernandez, Y. S. Chao, J. Baffic, M. H. Lam, C. Burton, C. P. Sparrow and J. G. Menke, *Bioorganic & medicinal chemistry letters*, 2005, 15, 2824-2828.
35 R. R. Singhaus, R. C. Bernotas, R. Steffan, E. Matelan, E. Quinet, P. Nambi, I. Feingold, C. Huselton, A. Wilhelmsson, A. Goos-Nilsson and J. Wrobel, *Bioorganic & medicinal chemistry letters*, 2010, 20, 521-525.
36 J. M. Travins, R. C. Bernotas, D. H. Kaufman, E. Quinet, P. Nambi, I. Feingold, C. Huselton, A. Wilhelmsson, A. Goos-Nilsson and J. Wrobel, *Bioorganic & medicinal chemistry letters*, 2010, 20, 526-530.
37 J. W. Ullrich, R. Morris, R. C. Bernotas, J. M. Travins, J. Jetter, R. Unwalla, E. Quinet, P. Nambi, I. Feingold, C. Huselton, C. Enroth, A. Wilhelmsson, A. Goos-Nilsson and J. Wrobel, *Bioorganic & medicinal chemistry letters*, 2010, 20, 2903-2907.
38 J. Wrobel, R. Steffan, S. Bowen, R. Magolda, E. Matelan, R. Unwalla, M. Basso, V. Clerin, S. Gardell, P. Nambi, E. Quinet, J. Reminick, G. Vlasuk, S. Wang, I. Feingold, C. Huselton, T. Bonn, M. Farnegardh, M. Hansson, A. G. Nilsson, A. Wilhelmsson, E. Zamaratski and M. Evans, *Journal of medicinal chemistry*, 2008, 51, 7161-7168.
39 J. Xu, *J. Med. Chem.*, 2002, 45, 5311-5320.
40 L. Wang, L. Chen, Z. Liu, M. Zheng, Q. Gu and J. Xu, *Plos One*, 2014, 9.
41 C. W. Yap, *Journal of computational chemistry*, 2011, 32, 1466-1474.

42 Z. Wang, Y. Chen, H. Liang, A. Bender, R. C. Glen and A. Yan, *Journal of chemical information and modeling*, 2011, 51, 1447-1456.
43 L. Wang, M. Wang, A. Yan and B. Dai, *Molecular diversity*, 2013, 17, 85-96.
44 L. Wang, X. Le, L. Li, Y. Ju, Z. Lin, Q. Gu and J. Xu, *Journal of chemical information and modeling*, 2014, DOI: 10.1021/ci500253q.
45 D. Rogers, R. D. Brown and M. Hahn, *Journal of biomolecular screening*, 2005, 10, 682-686.
46 X. Li, L. Chen, F. Cheng, Z. Wu, H. Bian, C. Xu, W. Li, G. Liu, X. Shen and Y. Tang, *Journal of chemical information and modeling*, 2014, 54, 1061-1069.
47 F. Cheng, Y. Ikenaga, Y. Zhou, Y. Yu, W. Li, J. Shen, Z. Du, L. Chen, C. Xu, G. Liu, P. W. Lee and Y. Tang, *Journal of chemical information and modeling*, 2012, 52, 655-669.
48 A. Yan, X. Nie, K. Wang and M. Wang, *European journal of medicinal chemistry*, 2013, 61, 73-83.
49 J. Fang, R. Yang, L. Gao, D. Zhou, S. Yang, A. L. Liu and G. H. Du, *Journal of chemical information and modeling*, 2013, 53, 3009-3020.
50 S. Tian, Y. Li, J. Wang, J. Zhang and T. Hou, *Molecular pharmaceutics*, 2011, 8, 841-851.
51 P. Watson, *J. Chem. Inf. Model.*, 2008, 48, 166-178.
52 Z. Liu, M. Zheng, X. Yan, Q. Gu, J. Gasteiger, J. Tijhuis, P. Maas, J. Li and J. Xu, *J. Comput. Aided Mol. Des.*, 2014, 28, 941-950.
53 M. Zheng, Z. Liu, X. Yan, Q. Ding, Q. Gu and J. Xu, *Molecular diversity*, 2014, 18, 829-840.
54 D. Li, L. Chen, Y. Li, S. Tian, H. Sun and T. Hou, *Molecular pharmaceutics*, 2014, 11, 716-726.
55 P. Geurts, A. Irrthum and L. Wehenkel, *Molecular bioSystems*, 2009, 5, 1593-1605.
56 S. Zhou, G. B. Li, L. Y. Huang, H. Z. Xie, Y. L. Zhao, Y. Z. Chen, L. L. Li and S. Y. Yang, *Computers in biology and medicine*, 2014, 51, 122-127.
57 L. Chen, Y. Li, Q. Zhao, H. Peng and T. Hou, *Molecular pharmaceutics*, 2011, 8, 889-900.
58 X. MA, R. Wang, C. Tan, Y. Jiang, T. Lu, H. Rao, X. Li, M. Go, B. Low and Y. Chen, *Mol. Pharmaceutics,*, 2010, 7, 1545-1560.
59 I. Saha, J. Zubek, T. Klingstrom, S. Forsberg, J. Wikander, M. Kierczak, U. Maulik and D. Plewczynski, *Molecular bioSystems*, 2014, 10, 820-830.
60 L. Y. Han, X. H. Ma, H. H. Lin, J. Jia, F. Zhu, Y. Xue, Z. R. Li, Z. W. Cao, Z. L. Ji and Y. Z. Chen, *Journal of molecular graphics & modelling*, 2008, 26, 1276-1286.
61 C. Chang and C. J. Lin, *http://www.csie.ntu.edu.tw/~cjlin/libsvm*, 2001.
62 P. Vasanthanathan, O. Taboureau, C. Oostenbrink, N. P. Vermeulen, L. Olsen and F. S. Jorgensen, *Drug metabolism and disposition: the biological fate of chemicals*, 2009, 37, 658-664.
63 Z. Li, L. Chen, Y. Lai, Y. Xie, Z. Dai and X. Zou, *Analytical Methods*, 2014, 6, 5281.
64 F. Klepsch, P. Vasanthanathan and G. F. Ecker, *Journal of chemical information and modeling*, 2014, 54, 218-229.
65 M. Farnegardh, T. Bonn, S. Sun, J. Ljunggren, H. Ahola, A. Wilhelmsson, J. A. Gustafsson and M. Carlquist, *The Journal of biological chemistry*, 2003, 278, 38821-38828.

66 S. Hoerer, A. Schmid, A. Heckel, R. M. Budzinski and H. Nar, *J. Mol. Biol.*, 2003, 334, 853-861.
67 X. Fradera, D. Vu, O. Nimz, R. Skene, D. Hosfield, R. Wynands, A. J. Cooke, A. Haunso, A. King, D. J. Bennett, R. McGuire and J. C. Uitdehaag, *J. Mol. Biol.*, 2010, 399, 120-132.
68 J. Zhang, B. Han, X. Wei, C. Tan, Y. Chen and Y. Jiang, *Plos One*, 2012, 7.
69 A. M. Wassermann, H. Geppert and J. Bajorath, *Journal of chemical information and modeling*, 2009, 49, 582-592.
70 D. Stumpfe, H. Geppert and J. Bajorath, *Chemical biology & drug design*, 2008, 71, 518-528.
71 *US Pat.,* WO2009086129, 2009.
72 *US Pat.,* WO2009086130, 2009.
73 *US Pat.,* WO2010059627, 2010.
74 *US Pat.,* WO2009020683, 2009.