MedChemComm

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/medchemcomm

Identifying Farnesoid X Receptor Agonists with Naïve Bayesian and Recursive Partitioning Approaches

Qianzhi Ding[a], Chanjuan Li[a], Ling Wang[b], Yali Li[a], Huihao Zhou[a], Qiong Gu[a*], and Jun Xu[a*]

Abstract

The farnesoid X receptor (FXR), a ligand-modulated transcription factor, is a multiple functional hepatic cell protector. Therefore, FXR agonists represent promising dyslipidemia and anti-diabetes agents. To identify novel FXR agonists, models were created from 144 known FXR agonists with naïve Bayesian (NB) and recursive partitioning (RP) approaches. The predictive and reliable models were selected with Matthews correlation coefficient (MCC) criterion (>0.900 with 117 testing compounds). The top 4 models were validated with the external data (282 compounds having cell-free activities and 500 decoys). Two optimal FXR agonist models (one from the NB method and the other from the RP method) were obtained from the top models by further validations. A virtual screening campaign was conducted against our in-house compound library with the optimal models and produced 15 virtual hits, which were further confirmed with cell-based luciferase assays. Finally, we discovered two new FXR agonists. Molecular docking studies indicated that the two new FXR agonists have similar binding modes to the known FXR agonists. This work demonstrated that a machine learning approach with combined NB and RP methods was able to identify novel FXR agonists and that the approach could be applied in other lead identification processes.

1 Introduction

The farnesoid X receptor (FXR), a ligand-activated transcriptional factor and multiple functional hepatic cell protector, is mainly expressed in liver and small intestine[1]. After activation by bile acids (BAs) or other agonists, FXR binds to specific DNA response elements as a heterodimer with the retinoid X receptor (RXR). Subsequently, with regulating the expression of genes, such as SREPB1c (sterol regulatory element

binding protein 1c), PEPCK1 (Phosphoenolpyruvate carboxykinase 1), BSEP (bile salt export pump), FXR is involved in bile acids, hepatic triglyceride, lipid and glucose homeostasis, liver regeneration/repair, hepatocyte survival, and tumor suppression[2]. FXR also down regulates hepatic inflammation and oncogenes[3, 4]. Hence, FXR is a potential drug target for the therapy of metabolic diseases[5, 6].

Known FXR agonists include steroidal and non-steroidal agents[7]. The steroidal agonists, such as BA derivatives[8] (e.g., chenodeoxycholic acids[9], CDCA, **1**), are endogenous ligands; steroidal agonists, such as 6α-ethyl-CDCA (6-ECDCA, **2**)[10] and MFA-1 (**3**)[11], are non-endogenous FXR agonists. Among these steroidal agnonists, 6-ECDCA (**2**) has entered into the clinical research for treating primary biliary cirrhosis (PBC) or non-alcoholic fatty liver disease (NAFLD)[12-15]. Moreover, a number of non-steroidal FXR agonists have been found, such as GW4064 (**4**)[16], fexaramine (**5**)[17], XL335 (**6**)[18] and the others with new scaffolds[19]. Some of the agonists have been tested in a reporter gene assay that is widely used in discovering FXR agonists and exhibited potent activities[20]. However, many of the known agonists, such as GW4064 and XL335, were not druggable due to problems of intrinsic toxicity, absorption, or metabolism[7, 18, 21, 22]. More new FXR agonists with acceptable pharmaceutical properties are in high demand.

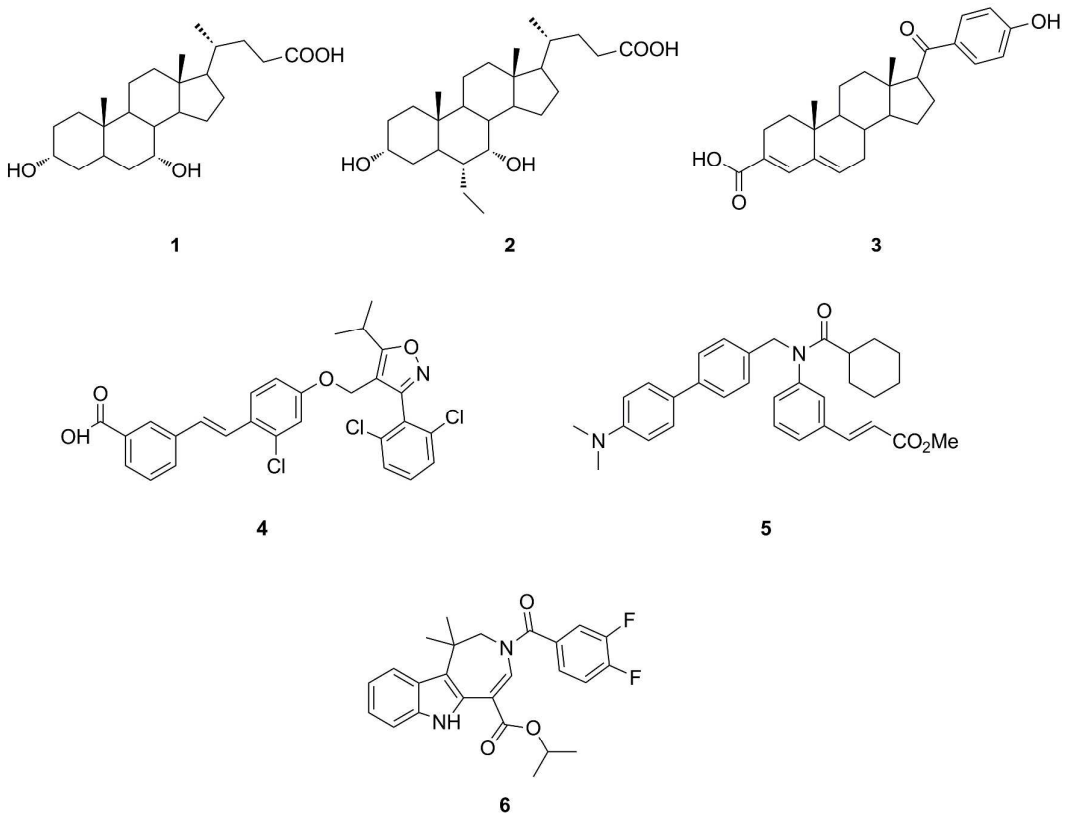

Fig 1. The known steroidal and non-steroidal FXR agonists.

FXR binding pocket is flexible[23] and allows structurally diverse agonists (Fig 1) with a number of different binding modes[11, 17, 24-26]. A number of structure-based pharmacophore models have been created since 2011[21,27,28] to predict FXR agonists. However, these models were ligand-scaffold dependent and referenced with a smaller number of ligands (<10 ligands). Therefore, it would be difficult for them to predict novel scaffold for FXR agonists.

To avoid the limitations of structure-based pharmacophore modeling, ligand-based machine learning approaches have been successfully applied for searching anti-MRSA (anti-methicillin-resistant Staphylococcus aureus) compounds[29] or identifying agents regulating PPAR (peroxisome proliferators-activated receptor)[30], LXR (liver X receptor)[31] and other proteins[32]. Therefore, in this paper, we built FXR agonist models using multiple machine learning approaches (NB and RP) based upon a larger training data set with diverse scaffolds.

To achieve optimal models, we generated many ligand-based models, which were evaluated by cross validations and external validations. The optimal models were used

in a virtual screening campaign against our in-house compound library for FXR agonists. The virtual screening hits were then tested with *in vitro* cell-based luciferase assays.

2 Materials and Methods

2.1 Data set

The human FXR agonist cell-based assay data used in this work were derived from the ChEMBL database (version 19)[33], and the data were selected by the following criteria: (1) the data of the human FXR agonist assay were selected; (2) the data of the cell-based assay were selected; and (3) duplicated compounds were removed. This approach resulted in 170 human FXR agonists with EC_{50} values ranging from 2 to over 100,000 nM (that is, five-order of magnitude). One hundred forty-four of the 170 human FXR agonists were marked as "active" (EC_{50} values were under or equal to 5 μM); the remaining 26 compounds were marked as "inactive" (EC_{50} values were greater than 5 μM). The activity threshold was set at 5 μM (see Fig S1 in the Supplementary information for more details).

2.2 Decoy generation

Decoys data were generated from DUD-E (a database of useful decoys: enhanced)[34] (http://dude.docking.org/), and added to the training data to keep it balanced. Ten diverse structures were selected from the "active" part of the database using the diverse molecules module in Pipeline Pilot 7.5 (Accelrys, Inc., San Diego, CA.). Subsequently, these 300 decoy structures were generated by calculating their molecular properties based upon the 10 reference structures in the DUD-E server. Three hundred decoys marked "inactive" were added into the database, and the whole data set was optimized using MOE 2013.08 (Chemical Computing Group Inc.) based on the MMFF94 force field[35]. All structures were saved as MACCS (Molecular ACCess System) sdf files and SMILES (Simplified molecular input line entry specification) files. Finally, the whole database was divided into two parts, a training set (353) and a test set (117), based on the random algorithm in Discovery Studio 2.5.5 (DS2.5.5, Accelrys, Inc., San Diego, CA.). The number of molecules in the training set was three times as many as that in the test set. This proportion was

employed in reference[36].

2.3 Calculation of molecular properties

The computed molecular properties (MP) were molecular weight (MW), the octanol/water partitioning coefficient (ALogP) based on the Ghose and Crippen's method, the molecular solubility (Molecular_Solubility), the apparent partition coefficient at pH = 7.4 (LogD) based on the Csizmadia's method, the molecular surface area (MSA), the molecular polar surface area (MPSA), the molecular fractional polar surface area (MFPSA), the number of rings (nR), the number of aromatic rings (nAR), the number of hydrogen bond donors (nHBDon), the number of hydrogen bond acceptors (nHBAcc), the count of oxygen and nitrogen (NPlusO), and the number of rotatable bonds (nRB). These values were all calculated with DS 2.5.5.

2.4 Calculation of molecular fingerprints

Two sets of fingerprints, SciTegic extended-connectivity fingerprints (ECFP, FCFP and LCFP) and Daylight-style path-based fingerprints (EPFP, FPFP and LPFP), were calculated using DS 2.5.5. Each type of fingerprint was used in four diameters: 4, 6, 8, and 10. All of these fingerprints are frequently applied in ADME, QSAR (quantitative structure–activity relationship), and QSPR (quantitative structure-property relationship) models[36, 37].

2.5 Naïve Bayesian

Naïve Bayesian is a simple probabilistic classification approach based on Bayes' theorem. Naïve Bayesian is highly scalable and unsupervised in a learning problem. The core function is eq. 1.

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \qquad (1),$$

where, H is the hypothesis or model, E is the observed data, P(H) is the probability of hypothesis H before observing any data, P(E) is the marginal probability of the data, and P(H|E) is the probability that the hypothesis H is correct for the observed data. P(E|H) is the likelihood that the probability of data E if hypothesis H is true. More details can be found in reference[38]. In this work, a Laplacian-corrected Bayesian classifier algorithm (implemented in DS2.5.5) was applied for building Bayesian

models. In our case, the models were trained with both the agonist ("active") and non-agonist ("inactive") data and considered each of the MPs and molecular fingerprint as the features to gain the knowledge to distinguish active from inactive. This building process is unbiased and takes the complexity of the model into consideration, which can avoid the over-fitting problem.

2.6 Recursive partitioning

Recursive partitioning (RP) is a type of accurate and comprehensible classification method that is used to discover the relationship between a dependent property (Y variable) and a number of independent properties (X variables). A decision tree will be created to classify the data points in the training set when RP proceeds. RP is a dichotomous process that divides independent variables (fingerprints and MPs). All RP models were built based on 12 fingerprints and 13 molecular descriptions in this study. Subsequently, 5-fold cross-validation was employed to determine the degree of pruning, which was required for the best predictive model. More details can be found in reference[39].

2.7 Evaluation of the model performance

To evaluate the performance of Bayesian and RP classifiers, 5-fold cross-validation was used in this study. A set of evaluation indexes, including true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), sensitivity (SE), specificity (SP), the prediction accuracy for agonists (Q_a), the prediction accuracy for non-agonists (Q_{na}), overall predictive accuracy (Q), and the Matthews correlation coefficient (MCC), were calculated with the formulas (2) to (7), and the receiver operating characteristic (ROC) curve was plotted. The area under the curves (AUC), which represents the classification ability of a binary classifier, was calculated through iteratively seeking the proper classifier threshold[40].

$$SE = \frac{TP}{TP+FN} \quad (2)$$

$$SP = \frac{TN}{TN+FP} \quad (3)$$

$$Qa = \frac{TP}{TN+FP} \quad (4)$$

$$Qna = \frac{TN}{TN+FN} \quad (5)$$

$$Q = \frac{TP+TN}{TP+FN+TN+FP} \quad (6)$$

$$\mathrm{MCC} = \frac{TP \times TN - FN \times FP}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}} \quad (7)$$

The MCC values are the measures for the classification accuracies of the models.

2.8 Cell culture

HEK293T cells were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 100 U/ml penicillin and 100 μg ml^{-1} streptomycin at 37°C in 5% CO_2 (V/V). The tested compounds were dissolved in DMSO and supplemented at indicated concentrations.

2.9 Transfection and luciferase assay

Activation studies on FXR were performed according to the method of Andrea A. Cronican[41] with a few modifications. To be brief, the HEK-293T cells were seeded into 96-well plates at 3×10^4 cells per well and allowed to attach overnight at 37°C. Plasmids pSG5/hFXR and pSG5/hRXRα, reporter plasmid pGL3/(DR-4)-c-fos-FF-luc, and the internal control plasmid pCMV/Renilla-luc were kindly gifts from Prof. Qing Song (University of Science and Technology, Beijing, China)[42]. These plasmids were co-transfected into cells using Lipofectamine™ 2000 (Invitrogen, USA) in accordance with the manufacturer's instructions. After 10 hours, cells were treated with tested compounds. FXR agonist GW4064 was used as a positive control; 0.1% DMSO was taken as vehicle. Luminescence measurements were processed 20 hours later. The results are expressed as relative firefly luciferase activity normalized to the renilla luciferase activity (fold change compared to vehicle control).

2.10 Molecular docking

Molecular docking was employed to gain an insight into the binding modes of two active compounds and FXR. The FXR-GW4064 crystal complex (PDB code: 3DCT) was downloaded from Protein Data Bank (PDB, http://www.rcsb.org/pdb/home/home.do) and prepared using the protein preparation protocol in the Schrödinger 2013.01. The extra precision (XP) mode in the Glide 5.9[43-45] of the Schrödinger software suite was employed to study the binding modes.

The docking parameters were all validated using re-docking methods. Two active compounds were prepared by Ligprep module in the Schrödinger software suite and docked into the FXR crystal structure in XP mode.

3 Results and discussion

3.1 Chemical space and structural diversity analysis

The structural diversity of the training and testing sets has a significant influence on the reliability and predictive ability of the models. In this study, an S-cluster approach (SCA)[46] (*in-house* software) was employed to measure the structural diversity (Fig 2). The cluster ID (CID), the serial number to each compound cluster, is proportional to the chemical structure complexity. More CIDs indicate higher structural diversity. The cyclicity is the metric of the cyclic degree of a molecule. The higher cyclicity value indicates the molecule has fewer/shorter substituents. Fig 2 demonstrates that the data points from both FXR agonists and non-agonists are widely spread, indicating that the structures of the 470 molecules are diverse.

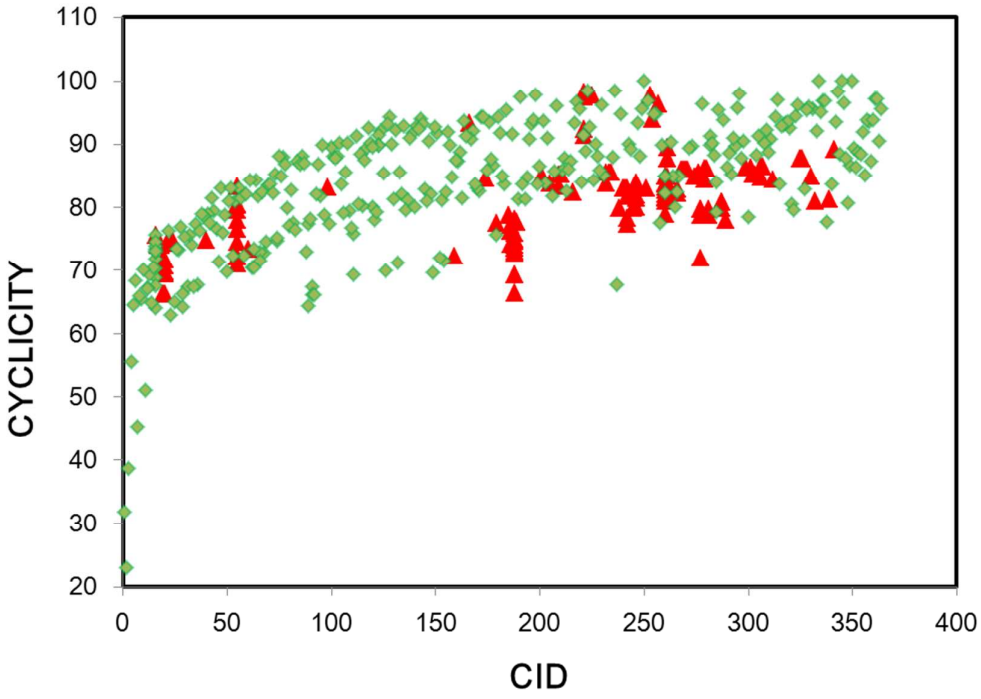

Fig 2. The SCA-plot of 470 compounds for model building. Red: FXR agonists. Green: FXR non-agonists.

3.2 Correlation analyses of molecular properties and FXR agonist activity

Thirteen physicochemical properties, including MW, ALogP, Molecular_Solubility,

LogD, MSA, MPSA, MFPSA, nR, nAR, nHBDon, nHBAcc, NPlusO and nRB, were calculated. The correlations between the 13 properties and FXR agonist activity were measured using correlation coefficients (R), and the significances of the difference between paired samples were evaluated with the student's *t* test (*p*-value) as listed in Table 1. LogD, MW, Molecular_Solubility, nAR, MSA, and MFPSA were significantly different from others. Molecular_Solubility was identified as the best property to discriminate FXR agonists and non-agonists, although it was not a strong predictor. ALogP, LogD, nR, nHBDon and nHBAcc exhibited relatively higher correlations with FXR agonist activity. Because no molecular property had a significantly higher value for FXR agonist activity, multiple molecular properties had to be used in building machine learning models.

Table 1. The MPs and their relationships with FXR agonist activity (R) and *p*-values (significances).

MP	*p*-value[a]	R[b]
ALogP	5.64×10^{-7}	0.154
LogD	7.98×10^{-11}	0.139
Molecular_Solubility	3.22×10^{-20}	0.080
MW	4.75×10^{-11}	0.045
NPlusO	2.36×10^{-3}	0.078
nRB	1.77×10^{-6}	0.076
nR	6.16×10^{-3}	0.100
nAR	9.30×10^{-10}	0.085
nHBAcc	9.22×10^{-3}	0.131
nHBDon	5.39×10^{-2}	0.153
MSA	5.31×10^{-14}	0.074
MPSA	1.08×10^{-3}	0.077
MFPSA	1.31×10^{-14}	0.085

[a]*p*-value: the statistical significance between FXR agonists and non-agonists.

[b]R: the correlation coefficient between a descriptor and EC_{50} (FXR).

3.3 Recursive partitioning models

The decision tree generated with RP is more intuitive compared with those "blind modeling" approaches, such as ANN (Artificial Neural Network) and SVM (Support Vector Machine). A deeper decision tree is more accurate, but it may cause over-fitting problems. A shorter decision tree may increase the possibility of applying the tree to new data sets, but it may reduce the accuracy of the prediction[36]. To optimize the depth of a decision tree for the best prediction performance, a number of experiments with depth thresholds ranging from 3 to 10 were tried. A total of 104 RP models were built and evaluated with evaluation indexes. The 5-fold cross-validation method was used to measure the robustness of those models.

With increasing depth thresholds from 3 to 10, 8 decision trees were built using 13 molecular properties. The MCC values of a test set indicated that the tree with the depth of 6 reached the best performance. The evaluation indexes of the best RP model based on MPs are listed in Table 2. In Table 2, MP represents 13 descriptors calculated by DS 2.5.5; Depth* represents the best tree depth for the corresponding RP model. For the training set, the best model with depth 6 achieves a sensitivity of 81.4%, specificity of 61.4%, MCC value of 0.387, and an AUC value of 0.716. For the test set, the performance was poor according to those evaluation indexes (SE_{test}=68.4%, SP_{test}=51.9%, MCC =0.191, and AUC=0.607, Table 2). All results of the test set suggested that the best RP model based on molecular properties is limited in distinguishing agonist from non-agonist because MPs represent whole molecular structure contributions, not sub-structural contributions. To take sub-structural contributions into account, molecular fingerprints must be taken into consideration. Therefore, 96 RP models were generated using MPs and 12 sets of molecular fingerprints.

As shown in Fig 3, the RP models derived from MPs and molecular fingerprints have much better MCC values than those derived only from MPs. The differences of the MCC values are more significant when the models were validated with the test data set.

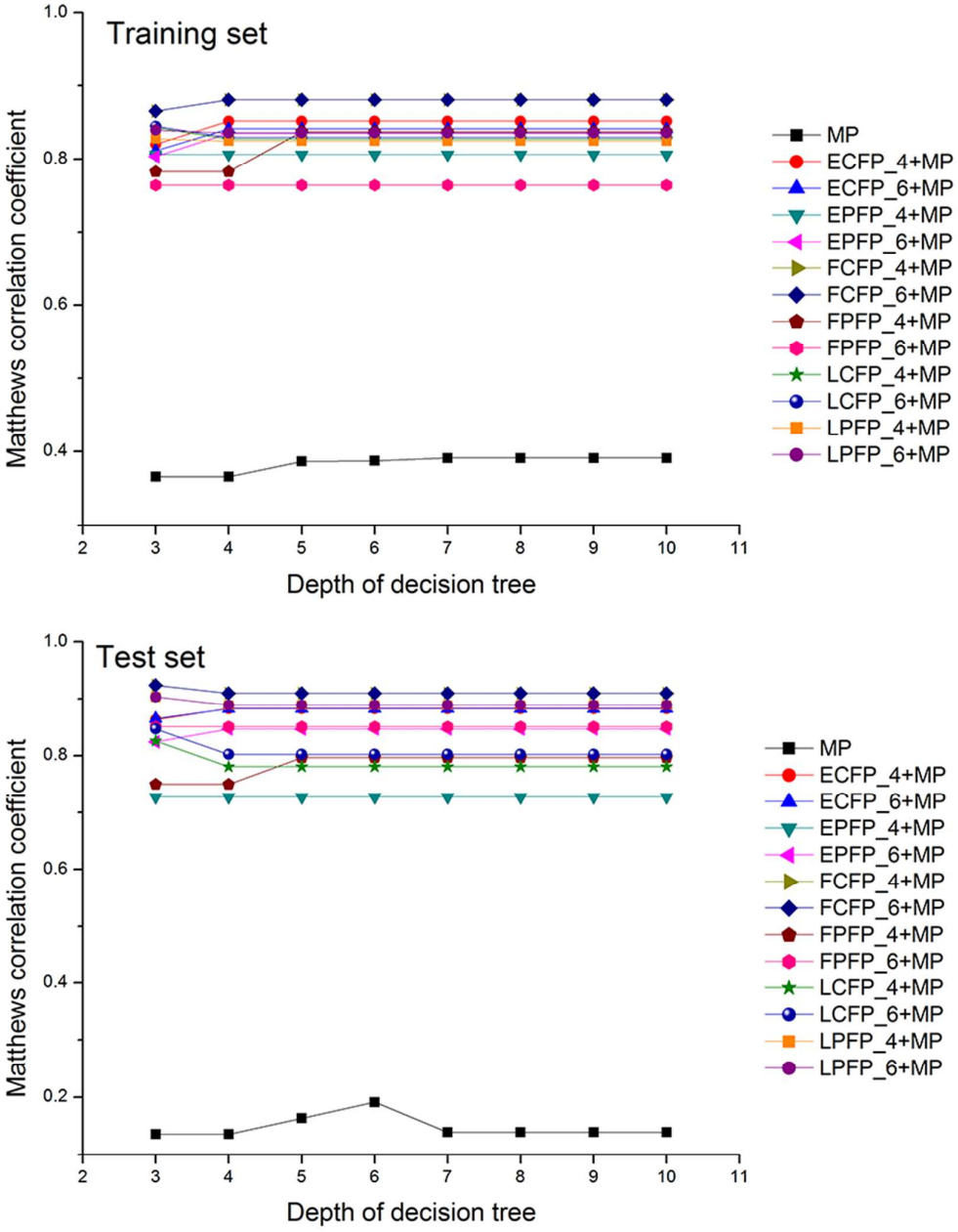

Fig 3. The relationships between the Matthews correlation coefficient values and the decision three depths, and descriptors. MP: the descriptors consist of only MPs (calculated with DS 2.5.5.). *+MP: the descriptors consist of different fingerprints plus MPs.

Table 2 lists the validation parameters for all of the best RP models tested with the test set and training set. In this table, RP models using FCFP_4 and FCFP_6 fingerprints have the highest MCC values (0.924) for the test set. The best decision tree depth of both models is 3. The two models have the same sensitivities (97.4%),

specificities (96.2%) and AUC values (0.975) for the test set.

Table 2. Performance of the best RP models with the combination of different fingerprints and MPs.

Models	Training set								Test set							
	TP	FN	TN	FP	*SE*	*SP*	*MCC*	AUC	TP	FN	TN	FP	*SE*	*SP*	*MCC*	AUC
MP[a]_depth6[b]	83	19	154	97	0.814	0.614	0.387	0.716	26	12	41	38	0.684	0.519	0.191	0.607
ECFP_4_depth4+MP	94	8	237	14	0.922	0.944	0.852	0.939	33	5	78	1	0.868	0.987	0.883	0.956
ECFP_6_depth4+MP	95	7	234	17	0.931	0.932	0.841	0.934	34	4	77	2	0.895	0.975	0.882	0.98
EPFP_4_depth3+MP	97	5	225	26	0.951	0.896	0.806	0.922	34	4	68	11	0.895	0.861	0.727	0.894
EPFP_6_depth4+MP	95	7	233	18	0.931	0.928	0.835	0.947	35	3	74	5	0.921	0.937	0.847	0.968
FCFP_4_depth3+MP	95	7	238	13	0.931	0.948	0.865	0.93	37	1	76	3	0.974	0.962	0.924	0.975
FCFP_6_depth3+MP	95	7	238	13	0.931	0.948	0.865	0.93	37	1	76	3	0.974	0.962	0.924	0.975
FPFP_4_depth5+MP	96	6	232	19	0.941	0.924	0.837	0.921	35	3	71	8	0.921	0.899	0.796	0.945
FPFP_6_depth3+MP	93	9	223	28	0.912	0.888	0.764	0.908	36	2	73	6	0.947	0.924	0.851	0.946
LCFP_4_depth3+MP	96	6	233	18	0.941	0.928	0.843	0.933	34	4	74	5	0.895	0.937	0.826	0.917
LCFP_6_depth3+MP	97	5	232	19	0.951	0.924	0.845	0.941	35	3	74	5	0.921	0.937	0.847	0.93
LPFP_4_depth3+MP	94	8	233	18	0.922	0.928	0.828	0.936	36	2	76	3	0.947	0.962	0.903	0.965
LPFP_6_depth3+MP	94	8	235	16	0.922	0.936	0.840	0.936	36	2	76	3	0.947	0.962	0.903	0.965

[a]MP: the 13 descriptors calculated with DS 2.5.5. [b]Depth*: the best tree depth for the corresponding model.

3.4 Naïve Bayesian models

One NB model was derived from the MPs calculated with DS 2.5.5 software. Twenty-four models were derived from 13 MPs combined with different types of molecular fingerprints (four diameters and six types). The MCC values for all NB models are depicted in Fig 4. The NB MP-only model has much lower MCC value than the MCC values of NB models derived from the combination of MPs and fingerprints. This is consistent with the cases of RP models (Fig 3).

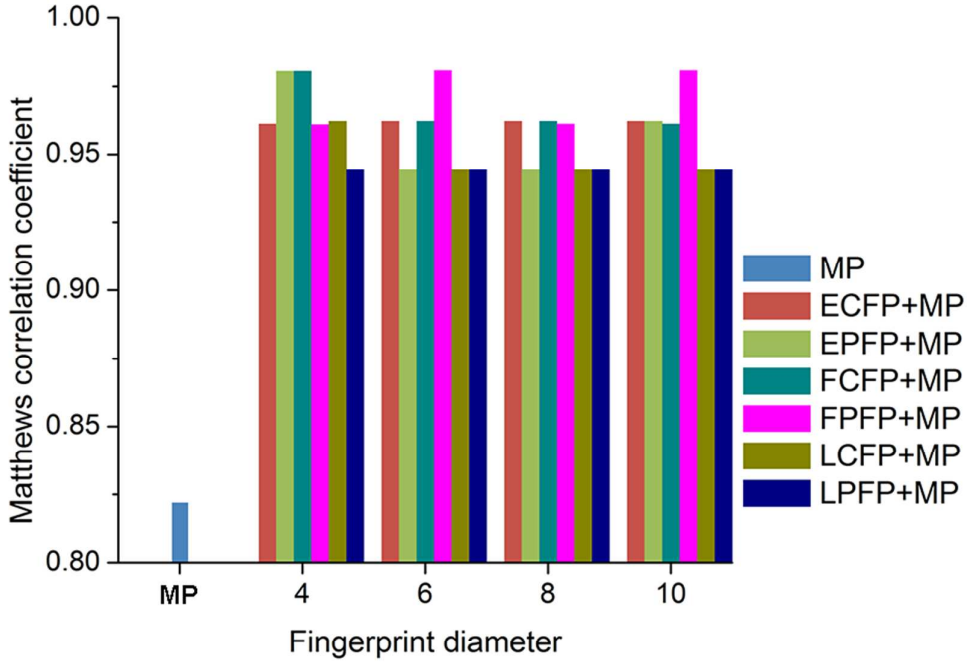

Fig 4. The Matthews correlation coefficient values for the NB MP model and the NB models derived from MP+fingerprints with different diameters.

The performance parameters for the top NB models are listed in Table 3. For the test data set, the NB models achieved the same performance (MCC = 0.981). Because greater diameter fingerprint requires higher computation resource, the best NB model was determined to be NB_FPFP_6+MP, which achieves a sensitivity of 100.0%, specificity of 98.7%, prediction accuracies for FXR agonist class of 97.4% and, an AUC value of 0.999. For the training set, the NB_FPFP_6+MP model achieves $SE_{training}$=98.0%, $SP_{training}$=95.2%, $MCC_{training}$=0.908, and AUC = 0.987.

Table 3. Performance parameters for the best RP and NB models

Models	Training set								Test set							
	TP	FN	TN	FP	*SE*	*SP*	*MCC*	AUC	TP	FN	TN	FP	*SE*	*SP*	*MCC*	AUC
RP_FCFP_4_depth3[b]+MP[a]	95	7	238	13	0.931	0.948	0.865	0.930	37	1	76	3	0.974	0.962	0.924	0.975
RP_FCFP_6_depth3+MP	95	7	238	13	0.931	0.948	0.865	0.930	37	1	76	3	0.974	0.962	0.924	0.975
NB_FPFP_6+MP	100	2	239	12	0.980	0.952	0.908	0.987	38	0	78	1	1.000	0.987	0.981	0.999
NB_FPFP_10+MP	100	2	236	15	0.980	0.940	0.890	0.984	38	0	78	1	1.000	0.987	0.981	1.000

[a]MP: the 13 descriptors calculated with DS 2.5.5. [b]Depth*: the best tree depth for the corresponding model.

3.5 Interpreting fingerprint modeling results

Structural fragments that make positive contributions to FXR agonists can be derived from the best NB model. These fragments (or privileged fragments) were exported from the top-*n* (*n*>0) fragments that have P(H|E) values greater than zero. The top-20 privileged fragments are listed in Fig 5 and represent the guidelines for FXR agonist design, virtual screening or lead optimizations. Some of the privileged fragments can be merged, such as, G2, G8 and G10 belong to the same fragment family; G3 and G20 belong to another family; etc. Many privileged fragments are alkaloids with conjugated double bonds system. Unsaturated seven-membered ring alkaloids are privileged scaffolds for FXR agonists.

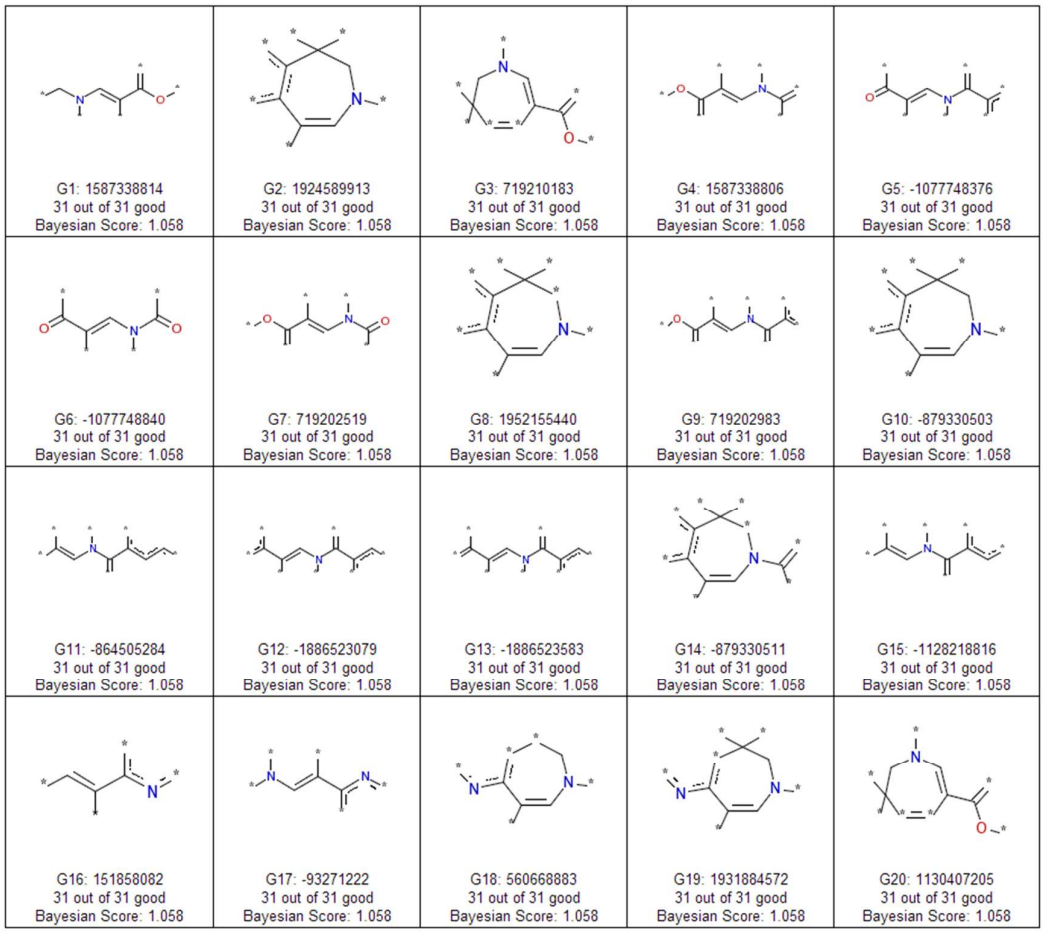

Fig 5. Privileged fragments exported from the best NB model, which was created from FPFP_6 fingerprints in DS 2.5.5.

3.6 Validating the best models with external cell-free data

In FXR agonist assay experiments, cell-free assays are more confirming than the

cell-based assays. Therefore, we use the FXR cell-free assay data to validate the models and determine the final FXR agonist predictive model.

The top-4 models, two RP models (RP_FCFP_4+MP and RP_FCFP_6+MP) and two NB models (NB_FPFP_6+MP and NB_FPFP_10+MP), were validated with an external dataset including 282 cell-free activity data and 500 decoy compounds. The 5-fold cross-validation was employed in this test. The external dataset was divided into five sub-datasets with five activity thresholds (1μM, 5μM, 10μM, 15μM, and 20μM). Twenty MCC values for the twenty test cases (four models by five sub-datasets) are depicted in Fig 6. One micromolar was determined to be the best activity threshold to define active compounds. This is also consistent with the active threshold definition (5 μM) we used when we built the models. Hence, the NB model, NB_FPFP_6+MP, is the best FXR agonist predictive model. The performance data of NB_FPFP_6+MP can be found in Table 3.

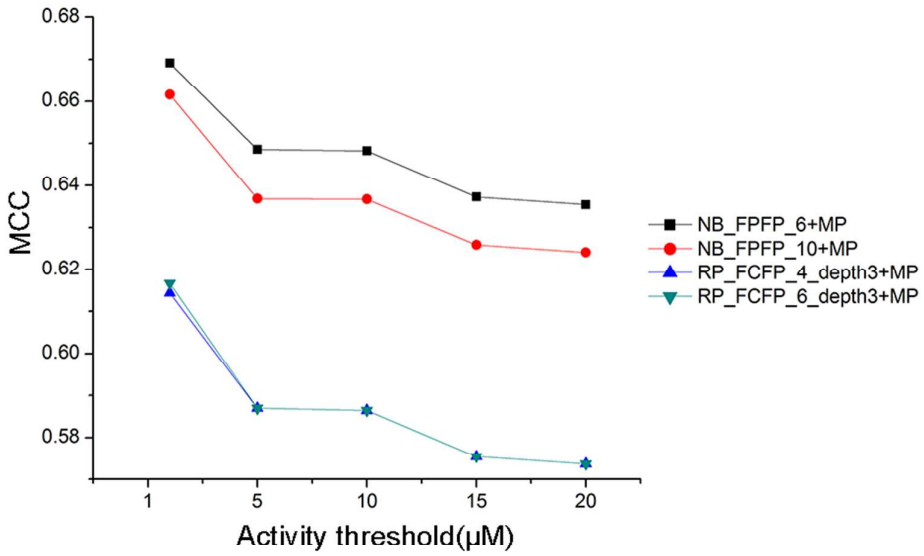

Fig 6. External validation: The relationship between MCC and activity threshold. The top line is for the best predictive model.

3.7 Virtual screening FXR agonists with the best models

Although NB_FPFP_6+MP was the most recommended FXR agonist predictive model, the virtual screening campaign still combined the results from the best model

of the RP approach (RP_FCFP_4_depth3+MP) to avoid potential false negatives. The virtual library is our in-house library, the Guangdong small molecule tangible library (GSMTL)[47], which has more than 7,500 chemical compounds with average purities >95%. The virtual screening resulted in 195 virtual hits (162 hits from NB, and 33 hits from RP). According to a previously study[29], 57 compounds with simple scaffolds and low molecular weight (<200) were abandoned. Finally, 15 compounds (Fig 7) were picked for an *in vitro* cell-based luciferase assay considered the diversity of the scaffolds and their availability.

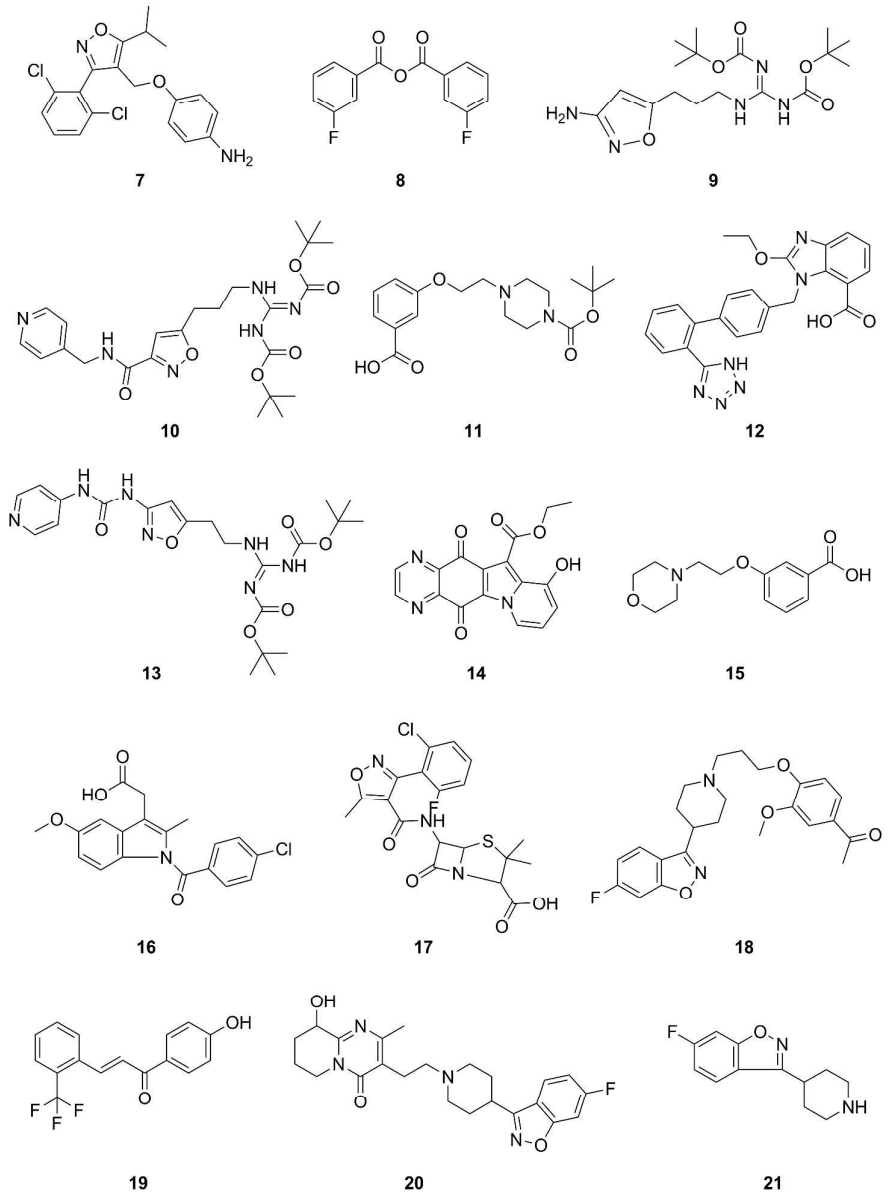

Fig 7. The 15 compounds confirmed with cell-based luciferase assays.

3.8 *In vitro* cell-based bioassay results

GW4064 was used as a positive control[16] for assaying the 15 compounds. The results are represented in relative firefly luciferase activities normalized to the renilla luciferase activities. The computational formula of the fold-activation is (firefly luciferase activities / renilla luciferase activities of test compound) / (firefly luciferase activities / renilla luciferase activities of control). The agonist activities of the 15 compounds are depicted in Fig 8 as a bar chart. Compounds **10** and **13** significantly activated FXR relative to the blank control. The EC_{50} values of compounds **10** and **13** are 15.39 and 29.94 μM. The activation curves of compounds **10** and **13** are shown in Fig S3, which exhibits a clear dose-dependent effect. The EC_{50} values are not very strong. This may be due to the poor bioavailability, which can be improved in the lead optimization process.

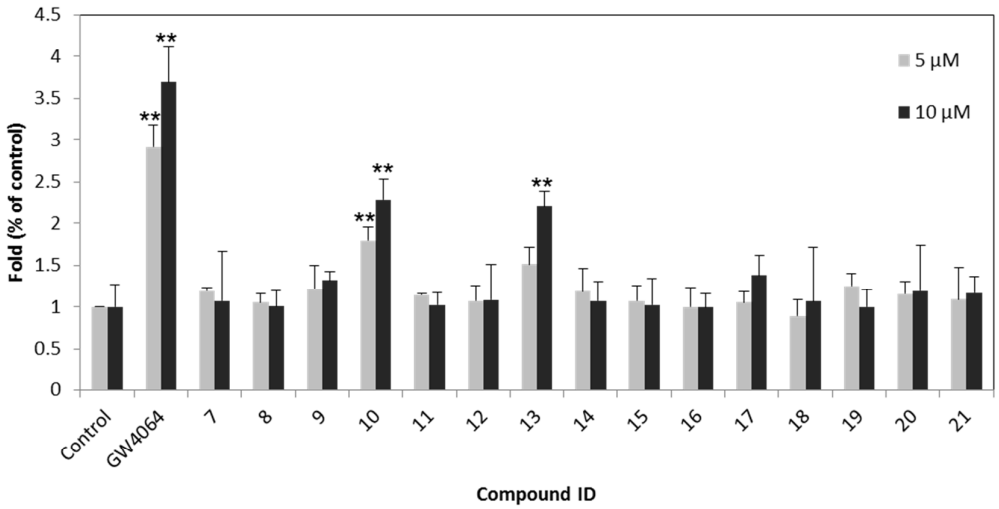

Fig 8. Bar chart for cell-based assay against FXR. The data are presented as the mean ± SE. Fold = (firefly luciferase activities / renilla luciferase activities of test compound) / (firefly luciferase activities / renilla luciferase activities of control). Compared with the control: ** $p < 0.01$; (n = 3).

3.9 Scaffold analyses

By inspecting the structures of compounds **10** and **13**, we suggest new scaffolds A and B (Fig 9) for FXR agonists. It is worth noting that scaffolds A and B are topologically

different, but structurally similar to each other. Both scaffolds A and B can be traced back to a known scaffold C that was derived from the training data set by means of the SCA[46] method. The old and new scaffolds all have the isoxazole ring and pyridine ring. However, these rings are connected by different links in different substitute positions. Furthermore, the side chains of compounds **10** and **13** have a similar structure to the link-like privileged fragments derived from the best NB models. These results demonstrate that established models are powerful in discovering FXR agonists with new scaffolds, and that the privileged fragments can guide the virtual screening of FXR agonists.

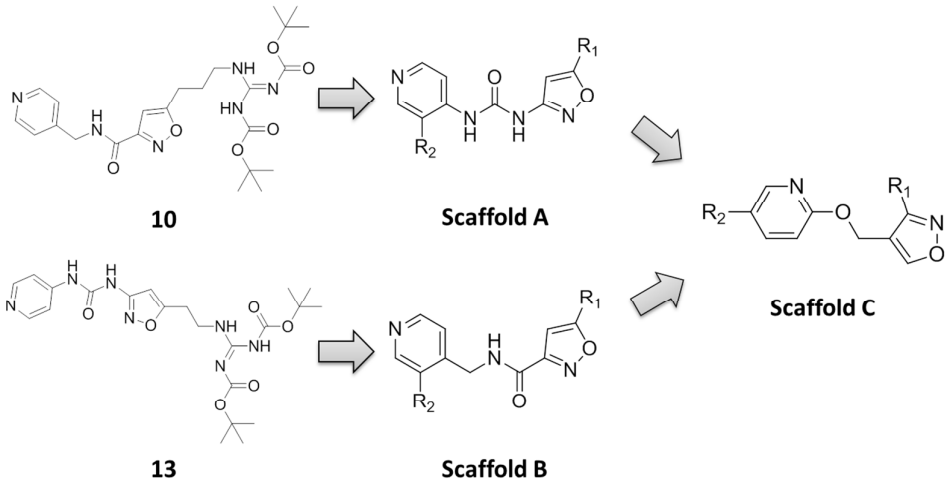

Fig 9. The scaffold analyses for compounds **10** and **13**.

3.9 Binding analyses for compounds 10 and 13

To ensure the XP mode of Glide 5.9 (Schrödinger, Inc.) is good for docking a ligand to FXR, the crystallized ligand was extracted from an experimental FXR co-crystal structure (PDB code: 3DCT), and the ligand was docked back to the FXR structure. This resulted in a number of ligand-FXR complexes. The average RMSD of the top-10 docking poses was 1.06Å, indicating that the XP mode of Glide 5.9 is suitable for FXR system.

Compounds **10** and **13** were prepared with Ligprep module, and docked to the FXR structure by means of the XP mode of Glide 5.9. The proposed binding modes are depicted in Fig 10. Both compounds interact with His447, which is consistent with the FXR-GW4064 complex[25] (PDB code: 3DCT) and FXR-MFA-1 complex[11] (PDB

code: 3BEJ). This interaction stabilizes the activation conformation of helix 12. Furthermore, the interaction of compound **10** with Met265 is consistent with the interaction of the existing FXR-GSK-8062 complex[25] (PDB code: 3DCU). The binding interactions further support the experimental data.

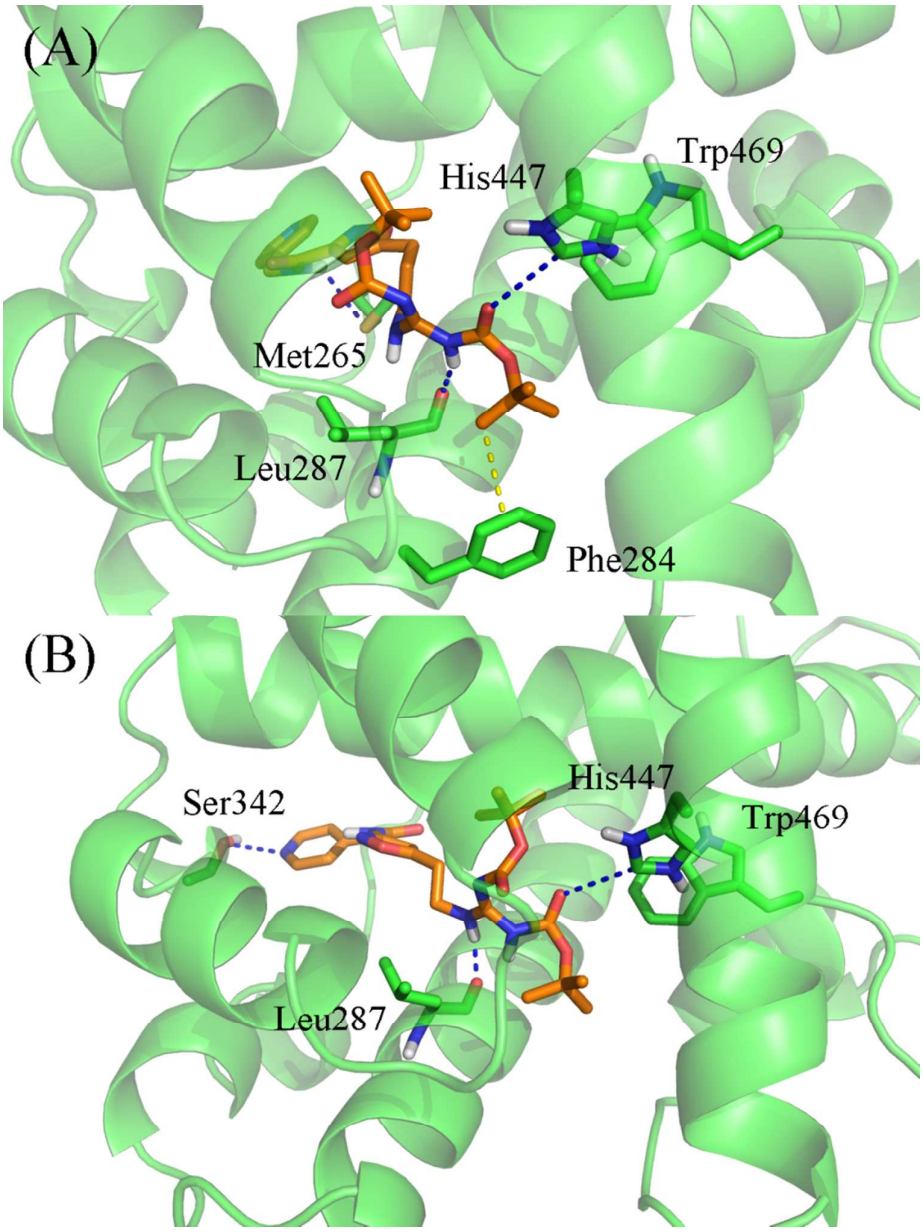

Fig 10. Binding analyses. A: Binding mode for compound **10**; B: Binding mode for compound **13**. The blue dashed line represents the H-bond interaction; the yellow dashed line represents the CH-π interaction.

4 Conclusions

It is a significant challenge to predict new FXR agonists because the FXR binding site is highly flexible. RP and NB approaches can be employed in building FXR agonist predictive models to avoid the problem caused by the flexibility. The keys for these ligand-based approaches are to identify proper descriptors. This study demonstrates the following: (1) the descriptors composed from the combinations of MPs and fingerprints are better than MPs alone; (2) privileged structural fragments can be derived from the best models using structural fingerprints that can serve as guidelines for FXR agonist design; and (3) the naïve Bayesian approach seems capable of producing better models. However, to avoid potential false negatives, we suggest that the best models from both NB and RP approaches are used, as they may generate similar hits with very different topological scaffolds.

Corresponding Author

*Phone: +86-20-39943077. Fax: +86-20-39943077. E-mail: guqiong@mail.sysu.edu.cn (Q.G.).

*Phone: +86-20-39943023. Fax: +86-20-39943023. Email: junxu@biochemomes.com. (J.X.).

Acknowledgement

This work was supported by the National Science Foundation of China (81173470, 81473138), 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). We thank Prof. Qing Song (University of Science and Technology, Beijing, China) for kindly providing plasmids.

Notes

[a]*School of Pharmaceutical Sciences & Institute of Human Virology, Sun Yat-Sen*

University, 132 East Circle Road at University City, Guangzhou, 510006, China.

[b]*Pre-Incubator for Innovative Drugs & Medicine, School of Bioscience and Bioengineering, South China University of Technology, Guangzhou 510006, China.*

† Electronic Supplementary Information (ESI) available.

References

1. B. M. Forman, E. Goode, J. Chen, A. E. Oro, D. J. Bradley, T. Perlmann, D. J. Noonan, L. T. Burka, T. McMorris, W. W. Lamph, R. M. Evans and C. Weinberger, *Cell*, 1995, 81, 687-693.
2. Y. Jiao, Y. Lu and X. Y. Li, *Acta Pharmacol Sin*, 2015, 36, 44-50.
3. M. Makishima, A. Y. Okamoto, J. J. Repa, H. Tu, R. M. Learned, A. Luk, M. V. Hull, K. D. Lustig, D. J. Mangelsdorf and B. Shan, *Science*, 1999, 284, 1362-1365.
4. X. F. Huang, W. Y. Zhao and W. D. Huang, *Acta Pharmacol Sin*, 2015, 36, 37-43.
5. S. Fiorucci, A. Mencarelli, E. Distrutti, G. Palladino and S. Cipriani, *Current medicinal chemistry*, 2010, 17, 139-159.
6. S. Fiorucci, A. Mencarelli, E. Distrutti and A. Zampella, *Future medicinal chemistry*, 2012, 4, 877-891.
7. D. Merk, D. Steinhilber and M. Schubert-Zsilavecz, *Future medicinal chemistry*, 2012, 4, 1015-1036.
8. D. J. Parks, S. G. Blanchard, R. K. Bledsoe, G. Chandra, T. G. Consler, S. A. Kliewer, J. B. Stimmel, T. M. Willson, A. M. Zavacki, D. D. Moore and J. M. Lehmann, *Science*, 1999, 284, 1365-1368.
9. D. W. Russell, *Annual review of biochemistry*, 2003, 72, 137-174.
10. R. Pellicciari, S. Fiorucci, E. Camaioni, C. Clerici, G. Costantino, P. R. Maloney, A. Morelli, D. J. Parks and T. M. Willson, *Journal of medicinal chemistry*, 2002, 45, 3569-3572.
11. S. M. Soisson, G. Parthasarathy, A. D. Adams, S. Sahoo, A. Sitlani, C. Sparrow, J. Cui and J. W. Becker, *Proceedings of the National Academy of Sciences of the United States of America*, 2008, 105, 5337-5342.
12. A. H. Ali, E. J. Carey and K. D. Lindor, *Annals of translational medicine*, 2015, 3, 5.
13. L. Adorini, M. Pruzanski and D. Shapiro, *Drug Discov Today*, 2012, 17, 988-997.
14. M. G. Silveira and K. D. Lindor, *Expert opinion on pharmacotherapy*, 2014, 15, 365-372.
15. S. Mudaliar, R. R. Henry, A. J. Sanyal, L. Morrow, H. U. Marschall, M. Kipnes, L. Adorini, C. I. Sciacca, P. Clopton, E. Castelloe, P. Dillon, M. Pruzanski and D. Shapiro, *Gastroenterology*, 2013, 145, 574-582 e571.
16. P. R. Maloney, D. J. Parks, C. D. Haffner, A. M. Fivush, G. Chandra, K. D. Plunket, K. L. Creech, L. B. Moore, J. G. Wilson, M. C. Lewis, S. A. Jones and T. M. Willson, *Journal of medicinal chemistry*, 2000, 43, 2971-2974.
17. M. Downes, M. A. Verdecia, A. J. Roecker, R. Hughes, J. B. Hogenesch, H. R. Kast-Woelbern, M. E. Bowman, J. L. Ferrer, A. M. Anisfeld, P. A. Edwards, J. M. Rosenfeld, J. G. Alvarez, J. P. Noel, K. C. Nicolaou and R. M. Evans, *Molecular cell*, 2003, 11, 1079-1092.
18. B. Flatt, R. Martin, T. L. Wang, P. Mahaney, B. Murphy, X. H. Gu, P. Foster, J. Li, P. Pircher, M. Petrowski, I. Schulman, S. Westin, J. Wrobel, G. Yan, E. Bischoff, C. Daige and R. Mohan, *Journal of medicinal chemistry*, 2009, 52, 904-907.
19. D. Merk, C. Lamers, K. Ahmad, R. Carrasco Gomez, G. Schneider, D. Steinhilber and M.

Schubert-Zsilavecz, *Journal of medicinal chemistry*, 2014, 57, 8035-8055.
20. D. Merk, D. Steinhilber and M. Schubert-Zsilavecz, *Expert Opin Drug Discov*, 2014, 9, 27-37.
21. J. Fu, P. Si, M. Zheng, L. Chen, X. Shen, Y. Tang and W. Li, *Bioorganic & medicinal chemistry letters*, 2012, 22, 6848-6853.
22. D. L. Howarth, S. H. W. Law, J. M. Law, J. A. Mondon, S. W. Kullman and D. E. Hinton, *Toxicology and applied pharmacology*, 2010, 243, 111-121.
23. D. Schuster, P. Markt, U. Grienke, J. Mihaly-Bison, M. Binder, S. M. Noha, J. M. Rollinger, H. Stuppner, V. N. Bochkov and G. Wolber, *Bioorganic & medicinal chemistry*, 2011, 19, 7168-7180.
24. L. Z. Mi, S. Devarakonda, J. M. Harp, Q. Han, R. Pellicciari, T. M. Willson, S. Khorasanizadeh and F. Rastinejad, *Molecular cell*, 2003, 11, 1093-1100.
25. A. Akwabi-Ameyaw, J. Y. Bass, R. D. Caldwell, J. A. Caravella, L. Chen, K. L. Creech, D. N. Deaton, S. A. Jones, I. Kaldor, Y. Liu, K. P. Madauss, H. B. Marr, R. B. McFadyen, A. B. Miller, F. N. Iii, D. J. Parks, P. K. Spearing, D. Todd, S. P. Williams and G. B. Wisely, *Bioorganic & medicinal chemistry letters*, 2008, 18, 4339-4343.
26. J. Achenbach, M. Gabler, R. Steri, M. Schubert-Zsilavecz and E. Proschak, *Medchemcomm*, 2013, 4, 920-924.
27. D. Schuster, P. Markt, U. Grienke, J. Mihaly-Bison, M. Binder, S. M. Noha, J. M. Rollinger, H. Stuppner, V. N. Bochkov and G. Wolber, *Bioorganic & medicinal chemistry*, 2011, 19, 7168-7180.
28. M. Marinozzi, A. Carotti, E. Sansone, A. Macchiarulo, E. Rosatelli, R. Sardella, B. Natalini, G. Rizzo, L. Adorini, D. Passeri, F. De Franco, M. Pruzanski and R. Pellicciari, *Bioorganic & medicinal chemistry*, 2012, 20, 3429-3445.
29. L. Wang, X. Le, L. Li, Y. Ju, Z. Lin, Q. Gu and J. Xu, *Journal of chemical information and modeling*, 2014, 54, 3186-3197.
30. S. Derksen, O. Rau, P. Schneider, M. Schubert-Zsilavecz and G. Schneider, *Chemmedchem*, 2006, 1, 1346-+.
31. Y. Li, L. Wang, Z. Liu, C. Li, J. Xu and Q. Gu, *Molecular bioSystems*, 2015, DOI: 10.1039/c4mb00718b.
32. N. Ai, M. D. Krasowski, W. J. Welsh and S. Ekins, *Drug Discov Today*, 2009, 14, 486-494.
33. A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, *Nucleic acids research*, 2012, 40, D1100-1107.
34. M. M. Mysinger, M. Carchia, J. J. Irwin and B. K. Shoichet, *Journal of medicinal chemistry*, 2012, 55, 6582-6594.
35. T. A. Halgren, *Journal of computational chemistry*, 1996, 17, 490-519.
36. L. Chen, Y. Y. Li, Q. Zhao, H. Peng and T. J. Hou, *Mol Pharmaceut*, 2011, 8, 889-900.
37. S. Tian, Y. Y. Li, J. M. Wang, J. Zhang and T. J. Hou, *Mol Pharmaceut*, 2011, 8, 841-851.
38. P. Watson, *Journal of chemical information and modeling*, 2008, 48, 166-178.
39. G. De'ath and K. E. Fabricius, *Ecology*, 2000, 81, 3178-3192.
40. P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen and H. Nielsen, *Bioinformatics*, 2000, 16, 412-424.
41. A. A. Cronican, N. F. Fitz, T. Pham, A. Fogg, B. Kifer, R. Koldamova and I. Lefterov, *Biochem Pharmacol*, 2010, 79, 1310-1316.
42. J. Fukuchi, C. Song, Q. Dai, R. A. Hiipakka and S. Liao, *J Steroid Biochem*, 2005, 94, 311-318.
43. R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis and P. S. Shenkin, *Journal of medicinal*

chemistry, 2004, 47, 1739-1749.

44. R. A. Friesner, R. B. Murphy, M. P. Repasky, L. L. Frye, J. R. Greenwood, T. A. Halgren, P. C. Sanschagrin and D. T. Mainz, *Journal of medicinal chemistry*, 2006, 49, 6177-6196.
45. T. A. Halgren, R. B. Murphy, R. A. Friesner, H. S. Beard, L. L. Frye, W. T. Pollard and J. L. Banks, *Journal of medicinal chemistry*, 2004, 47, 1750-1759.
46. J. Xu, *Journal of medicinal chemistry*, 2002, 45, 5311-5320.
47. Q. O. Gu, J. Xu and L. Q. Gu, *Molecules*, 2010, 15, 5031-5044.