Published in final edited form as:

Methods Mol Biol. 2015 ; 1260: 149–164. doi:10.1007/978-1-4939-2239-0_9.

Ligand Biological Activity Predictions Using Fingerprint-Based Artificial Neural Networks (FANN-QSAR)

Kyaw Z. Myint and Xiang-Qun Xie

Abstract

This chapter focuses on the fingerprint-based artificial neural networks QSAR (FANN-QSAR) approach to predict biological activities of structurally diverse compounds. Three types of fingerprints, namely ECFP6, FP2, and MACCS, were used as inputs to train the FANN-QSAR models. The results were benchmarked against known 2D and 3D QSAR methods, and the derived models were used to predict cannabinoid (CB) ligand binding activities as a case study. In addition, the FANN-QSAR model was used as a virtual screening tool to search a large NCI compound database for lead cannabinoid compounds. We discovered several compounds with good CB2 binding affinities ranging from 6.70 nM to 3.75 μ M. The studies proved that the FANN-QSAR method is a useful approach to predict bioactivities or properties of ligands and to find novel lead compounds for drug discovery research.

Keywords

Quantitative structure-activity relationship (QSAR); Fingerprint; Artificial neural networks (ANN); Biological activity; Cannabinoid

1 Introduction

Quantitative structure-activity relationship (QSAR) studies play essential roles in pharmaceutical research to identify and generate high-quality leads in the early stages of drug discovery [1–3]. QSAR studies help reduce the costly failures of drug candidates by identifying promising lead compounds and reducing the number of costly experiments. Such studies correlate chemical or structural features of compounds with their bioactivities. Descriptors are used to encode molecular features, and a mathematical relationship among a set of descriptors and biological endpoints of existing compounds is derived to construct a QSAR model which is then used to predict activities of new structures. Hence it is important to have a robust model which can learn important molecular features from a set of training compounds and then recognize such features in new structures to effectively predict their biological endpoints or activities.

The artificial neural networks (ANN) method is a machine learning algorithm with adaptive learning behavior in which the algorithm learns from previous examples and adapts to changes in input parameters. In comparison with receptor docking search [4], ligand similarity search [5, 6], machine-learning-based search [7, 8], and QSAR approaches [9–11], the ANN method possesses good generalization and pattern recognition ability for unseen data. Such features make it effective and robust for nonlinear regression problems with

multiple inputs. In fact, several studies have used the ANN algorithm to predict molecular properties or biological endpoints of chemical analogs in several case studies such as anti-diabetes, anticancer, anti-HIV, and allergenicity prediction [12–16].

In this chapter, we focus on the fingerprint-based ANN-QSAR (FANN-QSAR) research work [17] in predicting biological activities of structurally diverse cannabinoid (CB) ligands using ANN. To the best of our knowledge, there have been no previous studies which have used molecular fingerprints as descriptors to predict biological activities (such as pIC_{50} or pK_i), although a few studies have been reported to predict ligand classes [18, 19]. Three types of molecular fingerprints were used as network inputs to train ANN-QSAR models, and the results were compared to well-known 2D and 3D QSAR methods [1] using five data sets. As a case study, we used the FANN-QSAR method to predict binding affinities of cannabinoid ligands using a large and structurally diverse CB ligand data set [20]. In addition, we applied the method as a virtual screening tool to find new cannabinoid ligands from a large NCI database containing over 200,000 compounds, and we found three compounds with good cannabinoid receptor binding affinities which were experimentally validated. The results demonstrated that combination of molecular fingerprints and ANN can lead to a reliable and robust high-throughput virtual screening method which can be a useful tool in chemogenomics and computer-aided drug discovery research.

2 Methods

2.1 Data Sets

As shown in Table 1, a total of six data sets were used in this study. Five of them were compiled by Sutherland et al. [21] and were downloaded from their supplemental data. The sixth data set, cannabinoid receptor 2 (CB2) data set, was curated by the Xie lab [5, 20]. For this data set, if there were more than one reported CB2 activity for a ligand, an average activity was used.

Once the structure-activity relationship (SAR) data were collected, each data set was then divided into training, validation, and testing sets. For the ACE, AchE, BZR, COX2, and DHFR data sets, the same training and testing data sets provided by Sutherland et al. were used in order to allow the direct comparison of FANN- QSAR models to 3D and 2D QSAR models reported [21]. For each data set, 10 % of randomly selected compounds from the training set were used as a validation set. For the cannabinoid data set, the training and test sets were randomly divided. The training set contained 80 % of the compounds while the test set contained 10 %; the other 10 % were used as a validation set. The training set was used to train the model while the validation set was used to prevent overfitting of the model. The test set was used as an external set to evaluate the generalization ability of the trained FANN- QSAR models. The numbers of compounds found in each training, validation, and test sets for each data set are summarized in Table 2. For statistical modeling, the process was repeated five times for each data set, resulting in five different pairs of randomly divided training and test sets.

2.2 Implementation of Fingerprint-Based Artificial Neural Network QSAR (FANN-QSAR)

A feed-forward back-propagation neural network method was implemented using MATLAB® R2007b Neural Network Toolbox [22, 23]. As shown in Fig. 1, there are three layers in the network: an input layer, a hidden layer, and an output layer. Three different types of molecular fingerprints, namely FP2 [24], MACCS [25], and Extended-Connectivity Fingerprint (ECFP6) [26], were used in this study. The generation of fingerprints is described in Subheading 3. The number of input layer neurons was equal to the size of the fingerprint. For example, FP2 and ECFP6 fingerprints have 1,024 bits and therefore, the number of input neurons is equal to 1,024. Similarly, there are 256 input neurons for the MACCS fingerprint. The number of hidden layer neurons was varied between 100 and 1,000. The networks were trained using gradient descent with momentum training function (*traingdm*) to update weights and biases, the tangent sigmoid transfer function (*tansig*) for the hidden layer and the linear transfer function (*purelin*) for the output layer. 10 % randomly selected compounds from the training data were used as a validation set to decide when to stop training. The model training was stopped after 4,000 epochs (iterations) or if the mean square error (MSE) of prediction on the training set had reached the minimum value of 0.1. In addition, early stopping was enabled when the prediction error on the validation set increased for 300 epochs and the weights and biases at the minimum of the validation error were returned. The optimal number of hidden neurons was selected via cross-validation experiments in which the model was trained using different numbers of hidden neurons, and an average of training set and validation set mean squared errors (MSE) was calculated. The number of hidden neurons which gave the lowest average MSE was used as the optimal number for subsequent model testing on the test set. The mean squared error (MSE) is defined as

$$\text{MSE} = \frac{1}{T} \sum_{i=1}^T (t(i) - p(i))^2$$

where T is the total number of training samples, $t(i)$ is the target value of the i th sample, and $p(i)$ is the predicted value of the i th sample.

2.3 Comparison of the FANN- QSAR Model with Other Methods

The performance of the FANN-QSAR was compared to those of other reported 3D- and 2D-QSAR methods, including CoMFA [27], CoMSIA [28], Hologram QSAR (HQSAR) [29, 30], QSAR by eigenvalue analysis (EVA) [31], back-propagation feed-forward neural network implemented in Cerius2 using 2.5D descriptors (NN 2.5D), and ensemble neural network [32] (NN-ens) using 2.5D descriptors which were implemented and tested by Sutherland et al. [21]. Three different fingerprints, namely FP2, ECFP6, and MACCS, were used as inputs for FANN-QSAR models, and each model was trained separately for each fingerprint type. During each training process, a cross-validation experiment was performed to decide the optimal number of hidden neurons which was used subsequently on the test set prediction. To compare objectively, the FANN-QSAR models were trained and tested on the same training and test data sets provided by Sutherland et al. [21]. Three FANN-QSAR models were named as follows based on which molecular fingerprint was used as an input:

ECFP6-ANN-QSAR, FP2-ANN-QSAR, and MACCS-ANN-QSAR. Results of CoMFA, CoMSIA basic, HQSAR, EVA, NN (2.5D), and NN-ens (2.5D) methods were taken from the work of Sutherland et al. [21].

Final correlation coefficient (r^2 test) values of each data set are listed in Table 3. Comparisons of r^2 (test) values across all data sets show that ECFP6 fingerprint-based ANN-QSAR model (ECFP6-ANN-QSAR) performed better than FP2 and MACCS fingerprint-based models for all data sets. For ACE, AchE, and COX2 data sets, the CoMFA model performed better than ECFP6-ANN-QSAR model but by a small margin. The ECFP6-ANN-QSAR model performed better for the DHFR and BZR data sets. Performance of the CoMSIA model was similar to that of the ECFP6-ANN-QSAR model. It is important to note that CoMFA and CoMSIA are field-based 3D QSAR methods which require similar scaffolds and high-quality molecular alignments to make effective predictions [11]. On the other hand, ECFP6-ANN-QSAR is a fingerprint-based method which works on structurally diverse data sets and requires no alignment during the model training process, which makes it more robust and high throughput in virtual screening. However, different fingerprints can produce different results and, in our work, ECFP6 produced an overall better result across different data sets compared to FP2 and MACCS fingerprints. In addition to 3D QSAR methods, the FANN-QSAR models were compared to another 2D QSAR method known as hologram QSAR (HQSAR), which is based on molecular holograms containing counts of molecular fragments similar to fingerprints. It can be observed that ECFP6-ANN-QSAR performed consistently better than HQSAR in all data sets except for DHFR data set resulting in the same r^2 test value (0.63). The FANN-QSAR models were also compared to other neural network approaches that used 2.5D descriptors as reported by Sutherland et al. The ECFP6-ANN-QSAR model performed better than the NN (2.5D) method in three out of five data sets and an ensemble of ten neural networks (NN-ens) approach using 2.5D descriptors performed slightly better than ECFP6-ANN-QSAR model in three out of five data sets. It is important to note that all QSAR models failed for COX2 and BZR data sets (r^2 test < 0.34) and had moderate performances (r^2 test < 0.64) for the other three data sets. Overall, the ECFP6-ANN-QSAR model performed consistently across all data sets and its performance was comparable to other 3D, 2D, and neural networks QSAR methods previously reported.

2.4 Cannabinoid Receptor Binding Activity Prediction for Known Cannabinoid Ligands

We extended the application of FANN-QSAR by predicting the binding activities of cannabinoid ligands. A total of 1,699 structurally diverse cannabinoid ligands with reported CB2 binding affinities were used. The CB2 binding activity data were downloaded from the CBID data set compiled by the Xie lab (<http://www.cbligand.org/cbid/index.php>). The ligands were randomly divided into training and test sets. FANN-QSAR models using different fingerprints were trained on training sets and the optimal numbers of hidden neurons were selected via cross-validation. Figure 2 contains a summary of cross-validation results for all three FANN-QSAR models. It can be observed that different training and test sets as well as different types of fingerprints resulted in different optimal numbers of hidden neurons, which suggested that cross-validation experiments are necessary to train neural networks for the best results.

After such training and parameter tuning, the predictive accuracy of the final model on the test set was evaluated. The process was repeated five times and a summary of r^2 values from each round of experiments can be seen in Table 4. Within each round, the same training and test compounds were used across all three FANN-QSAR models. For example, the same training and test compounds in Round 1 of the ECFP6-ANN-QSAR model were used in the Round 1 of the FP2-ANN-QSAR and MACCS-ANN-QSAR models. As shown in the table, the ECFP6-ANN-QSAR model consistently outperformed the FP2- and MACCS-ANN-QSAR models in all five rounds of experiments. The ECFP6-ANN-QSAR model achieved an average r^2 test value of 0.56 ($r = 0.75$) across all repeat experiments compared with 0.48 ($r = 0.69$) and 0.45 ($r = 0.67$) for the FP2- and MACCS-ANN-QSAR models, respectively. Results showed that the ECFP6 fingerprint was better than FP2 and MACCS fingerprints for the cannabinoid data set as well as the other five data sets. In fact, it has been also reported that circular fingerprints such as ECFP6 fingerprints are found to be more useful in virtual screening and ADMET properties' prediction studies [33, 34]. Our results suggested that an ECFP6 fingerprint- based ANN-QSAR model can be used in virtual screening of chemical ligands in a high-throughput manner since it only requires 2D fingerprints as inputs instead of 3D molecular alignments and bioactive conformations, as required by other 3D QSAR methods.

2.5 Cannabinoid Receptor Binding Activity Prediction on Newly Reported Cannabinoid Ligands

To more rigorously test the predictive ability of the FANN-QSAR method on new cannabinoid compounds which are not in the Xie group's cannabinoid ligand training data set, the most recently reported cannabinoid ligands and associated CB2 binding affinity data were downloaded from ChEMBL database [35]. These compounds were not found in the training (CBID) data set and were collected to be used as a new test set in order to evaluate the FANN-QSAR performance. The new test data set consisted of 295 compounds with reported CB2 K_i values which were then converted to pK_i values. 41.55 % of new CB2 ligands were less than 80 % similar (2D Tanimoto similarity) and 25.34 % were less than 70 % similar to the training compounds. This similarity analysis indicated that the newly reported CB2 compounds contained a good mixture of similar and dissimilar compounds to the training database. The ECFP6-ANN-QSAR model was trained using the 1,699 CB2 ligand (CBID) data set. Twenty independent rounds of training and testing were performed. For each round, a randomly selected 90 % of the database was used for training and the remaining 10 % was used for validation. As a result, 20 independent trained models were derived. After 20 rounds of predictions, an average predicted value for each test compound was calculated. The average residual value was 0.046 and the standard deviation was 1.03. Seventeen outlier compounds with residuals more than two standard deviations away from the average residual were removed. Figure 3 shows a scatter plot of experimental and predicted pK_i values of 278 test compounds after such outlier removal. The linear regression of these 278 data points provided an r of 0.75, slope of 0.686, and y intercept of 2.249. This result indicated that there was a good correlation between experimental and predicted values, given the fact that many of these test compounds have novel structures and were not included in the model training and validation process. The result suggested that the FANN-QSAR possessed good generalization ability for newly reported cannabinoid ligands.

2.6 Virtual Screening of the NCI Compound Database for Lead Cannabinoid Ligands

In order to illustrate how the FANN-QSAR model could be used in drug discovery research, we applied it as a virtual screening tool to search for CB2 lead ligands from the NCI compound database [36]. For consistency, the same 20 trained models in the previous section were used. The NCI database, containing 329,089 compounds, was filtered to remove duplicate compounds, isotopes, metals, and mixtures using the Tripos Selector program [30, 37]. This filtering reduced it to 211,782 compounds that were used as a test set for each round of prediction. For each compound the ECFP6 fingerprint was generated and used as the network input to predict the CB2 receptor binding activity. After 20 rounds of predictions, an average predicted value for each compound was calculated. The top ranked 50 compounds were selected, but only 10 compounds were physically available from the NCI via material transfer agreement (MTA). These ten compounds were experimentally tested for CB2 activities using a [³H]CP-55940 competition binding assay experiment. The experimental protocol for this validation assay is described in Subheading 3.

Among the ten tested NCI compounds, four (NSC49888, NSC174122, NSC369049, and NSC76301) had CB2 K_i between 6.70 nM ($pK_i = 8.17$) and 3.80 μ M ($pK_i = 5.42$). One compound, which has a similar chemical scaffold to the well-known cannabinoid ligand, delta-9-tetrahydrocannabinol, was found to be a high-affinity compound with an average CB2 K_i value of 6.70 nM ($pK_i = 8.17$). These four compounds and other similar compounds (70 % 2D Tanimoto similarity threshold was used) [38] were not found in the training database. Among the top 50 ligands, there was one NCI compound (NSC768843) which was more than 90 % similar (Tanimoto coefficient = 0.9) to a known classical cannabinoid ligand (CAS ID: 112830-95-2 or HU210), an analog of delta-9-tetrahydrocannabinol, reported in the literature [39]. These findings proved that the FANN-QSAR method can find not only novel compounds with good CB2 binding affinities but also compounds similar to known ligands from a testing database containing thousands of compounds with diverse scaffolds. Hit ligands with novel scaffolds can be used as lead compounds for further medicinal chemistry optimization and SAR studies, while hits similar to known ligands provide additional information for scaffold hopping and R-group variations which may be useful for medicinal chemists. Table 5 contains the structures of NCI hit compounds and their experimental pK_i as well as predicted values. Apart for one compound (NSC746843) that was not available from NCI, the other four compounds were experimentally tested in our lab and competition binding curves are shown in Fig. 4.

It should be noted that the predicted pK_i correlated well with experimental pK_i for two of the five hit ligands but not for the other three ligands. This finding could be attributed to the experimental variability of the reported CB2 binding activities of training compounds among different research labs, or to a possible limitation of 2D fingerprint descriptors which considers individual fragment contributions but sometimes may not be as effective as other 3D descriptors when considering the overall structure of a ligand. Fingerprints such as ECFP6 have, however, been found to be useful in this study as well as in other several cheminformatics studies [5, 33, 34], and they are known to be robust and time efficient for high-throughput virtual screening applications where hundreds of thousands of chemicals are involved, as in this study. To conclude, results from the virtual screening exercise that

was validated experimentally demonstrated that the derived FANN-QSAR model is capable of successfully identifying lead CB2 compounds with good binding affinities as well as compounds similar to known cannabinoid ligands, and providing additional insights for R-group and scaffold hopping of known ligands.

3 Notes

We used three different types of molecular fingerprints, namely FP2 [24], MACCS [25], and Extended-Connectivity Fingerprint (ECFP6) [26]. FP2 is a path-based fingerprint which indexes molecular fragments and MACCS is a key-based fingerprint which uses 166 predefined keys, whereas ECFP6 is a circular topological fingerprint which is derived using a variant of the *Morgan* algorithm [40]. FP2 and MACCS fingerprints were generated using the “*babel*” command from the OpenBabel program [24] while ECFP6 fingerprints were generated using the “*generatemd*” command from the ChemAxon program (<http://www.chemaxon.com>). Ligand chemical structures stored in SDF format were used as inputs to generate fingerprints. For each ligand, polar hydrogens were added using the OpenBabel program [24] before fingerprint generation. All fingerprints were fixed-length binary representations with 1,024 bits for both ECFP6 and FP2, and 256 bits for MACCS fingerprint. Fingerprints were generated for each ligand in the data sets and used as inputs to train the FANN-QSAR models.

In order to evaluate CB₂ binding activity of virtually screened ligands, competition binding assays were performed by displacing radioactive [³H]CP-55940 radioligand. The experimental protocol has been established based on previously reported procedures [41–44] and is described briefly below.

A Perkin Elmer 96-well TopCounter is used in our laboratory to measure the CB receptor binding affinity (K_i) of the *in silico*-screened ligands by displacing [³H]CP-55940. In competition binding experiments, ligands were diluted in dilution buffer (50 mM Tris, 5 mM MgCl₂, 2.5 mM EGTA) containing 0.1 % (w/v) fatty acid-free bovine serum albumin (BSA), 10 % dimethyl sulfoxide, and 0.4 % methyl cellulose. Various concentrations of ligands/samples are added in the same volume to 2.5 nM [³H] CP-55940. Incubation buffer (50 mM Tris, 2.5 mM EGTA, 5 mM MgCl₂, 0.1 % (w/v) fatty acid-free BSA) and cell membrane preparations from CHO cells expressing CB2 receptors (5 µg per well) are added to a final volume of 200 µL. For the saturation binding experiments, varying concentrations of [³H] CP-55940 (0.05–4 nM) with or without 5 µM of an unlabeled known ligand (CP-55940) were incubated with the receptor membrane preparations to determine K_d and nonspecific binding. After the binding suspensions are incubated at 30 °C for 1 h, the reaction is terminated by rapid filtration through microfiltration plates (Unifilter GF/B filterplate, Perkin Elmer) followed by five washes with ice-cold TME buffer containing 0.1 % BSA on a Packard Filtermate Harvester (Perkin Elmer). The plates are then dried overnight and 30 µl MicroScint 0 scintillation liquid is added to each well of the dried filter plates. Then the bound radioactivity is counted using a Perkin Elmer 96-well TopCounter. The K_i is calculated by using nonlinear regression analysis (Prism 5; GraphPad Software Inc., La Jolla, CA), with the K_d values for [³H]CP-55940 determined from saturation

binding experiments. This assay is used for determining binding affinity parameters (K_i) of ligand-receptor interactions between the CB2 receptor and ligands.

References

1. Myint KZ, Xie X-Q. Recent advances in fragment-based QSAR and multi-dimensional QSAR methods. *Int J Mol Sci*. 2010; 11(10):3846–3866. [PubMed: 21152304]
2. Perkins R, Fang H, Tong W, et al. Quantitative structure-activity relationship methods: perspectives on drug discovery and toxicology. *Environ Toxicol Chem*. 2003; 22(8):1666–1679. [PubMed: 12924569]
3. Salum L, Andricopulo A. Fragment-based QSAR: perspectives in drug design. *Mol Divers*. 2009; 13(3):277. [PubMed: 19184499]
4. Chen JZ, Wang J, Xie XQ. GPCR structure-based virtual screening approach for CB2 antagonist search. *J Chem Inf Model*. 2007; 47(4):1626–1637.10.1021/ci7000814 [PubMed: 17580929]
5. Wang L, Ma C, Wipf P, et al. TargetHunter: an in silico target identification tool for predicting therapeutic potential of small organic molecules based on chemogenomic database. *AAPS J*. 2013; 15:395–406. [PubMed: 23292636]
6. Tandon M, Wang L, Xu Q, et al. A targeted library screen reveals a new inhibitor scaffold for protein kinase D. *PLoS One*. 2012; 7(9):e44653.10.1371/journal.pone.0044653 [PubMed: 23028574]
7. Ma C, Wang L, Yang P, et al. LiCABEDS II. Modeling of ligand selectivity for G-protein coupled cannabinoid receptors. *J Chem Inf Model*. 2013; 53(1):11–26.10.1021/ci3003914 [PubMed: 23278450]
8. Wang L, Ma C, Wipf P, et al. Linear and nonlinear support vector machine for the classification of human 5-HT1A ligand functionality. *Mol Inf*. 2012; 31(1):85–95.10.1002/minf.201100126
9. Myint K, Xie X-Q. Fragment-based QSAR algorithm development for compound bioactivity prediction. *SAR QSAR Environ Res*. 2011; 22(3):385–410. [PubMed: 21598200]
10. Chen JZ, Myint KZ, Xie X-Q. New QSAR prediction models derived from GPCR CB2-antagonistic triaryl bis-sulfone analogues by a combined molecular morphological and pharmacophoric approach. *SAR QSAR Environ Res*. 2011; 22(5–6):525–544.10.1080/1062936x.2011.569948 [PubMed: 21714749]
11. Chen J-Z, Han X-W, Liu Q, et al. 3D-QSAR studies of arylpyrazole antagonists of cannabinoid receptor subtypes CB1 and CB2. A combined NMR and CoMFA approach. *J Med Chem*. 2006; 49(2):625–636. [PubMed: 16420048]
12. Vilar S, Santana L, Uriarte E. Probabilistic neural network model for the in silico evaluation of anti-HIV activity and mechanism of action. *J Med Chem*. 2006; 49(3):1118–1124.10.1021/jm050932j [PubMed: 16451076]
13. González-Díaz H, Bonet I, Terán C, et al. ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds. *Eur J Med Chem*. 2007; 42(5):580–585.10.1016/j.ejmech.2006.11.016 [PubMed: 17207560]
14. Patra JC, Chua BH. Artificial neural network-based drug design for diabetes mellitus using flavonoids. *J Comput Chem*. 2011; 32(4):555–567.10.1002/jcc.21641 [PubMed: 20806262]
15. Dimitrov I, Naneva L, Bangov I, et al. Allergenicity prediction by artificial neural networks. *J Chemometrics*. 2014; 1002/cem.2597
16. Vanyúr R, Héberger K, Kövesdi I, et al. Prediction of tumoricidal activity and accumulation of photosensitizers in photodynamic therapy using multiple linear regression and artificial neural networks. *Photochem Photobiol*. 2002; 75(5):471–478.10.1562/0031-8655(2002)0750471potaaa2.0.co2 [PubMed: 12017472]
17. Myint K-Z, Wang L, Tong Q, et al. Molecular fingerprint-based artificial neural networks QSAR for ligand biological activity predictions. *Mol Pharm*. 2012; 9(10):2912–2923.10.1021/mp300237z [PubMed: 22937990]

18. Molnár L, Keser GM. A neural network based virtual screening of cytochrome P450 3A4 inhibitors. *Bioorg Med Chem Lett*. 2002; 12(3):419–421.10.1016/s0960-894x(01)00771-5 [PubMed: 11814811]
19. Muresan S, Sadowski J. “In-House Likeness”: comparison of large compound collections using artificial neural networks. *J Chem Inf Model*. 2005; 45(4):888–893.10.1021/ci049702o [PubMed: 16045282]
20. Wang L, Xie XQ. Cannabinoid Ligand Database. 2012 Accessed Nov 2011.
21. Sutherland JJ, O’Brien LA, Weaver DF. A comparison of methods for modeling quantitative structure–activity relationships. *J Med Chem*. 2004; 47(22):5541–5554.10.1021/jm0497141 [PubMed: 15481990]
22. Greenidge PA, Carlsson B, Bladh L-G, et al. Pharmacophores incorporating numerous excluded volumes defined by x-ray crystallographic structure in three-dimensional database searching: application to the thyroid hormone receptor. *J Med Chem*. 1998; 41(14):2503–2512. [PubMed: 9651155]
23. Mathworks. MATLAB. 7.5.0.342 (R2007b). Natick, MA: 2007. edn
24. O’Boyle N, Banck M, James C, et al. Open Babel: an open chemical toolbox. *J Cheminform*. 2011; 3:33. [PubMed: 21982300]
25. Durant JL, Leland BA, Henry DR, et al. Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci*. 2002; 42(6):1273–1280.10.1021/ci010132r [PubMed: 12444722]
26. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model*. 2010; 50(5):742–754.10.1021/ci100050t [PubMed: 20426451]
27. Cramer R, Patterson D, Bunce J. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J Am Chem Soc*. 1988; 110:5959–5967. [PubMed: 22148765]
28. Klebe G. Comparative molecular similarity indices analysis: CoMSIA. 3D QSAR in Drug Design. Three-Dimensional Quantitative Structure Activity Relationships. 1998; 3:87–104.
29. Lowis D. HQSAR: a new, highly predictive QSAR technique. Tripos Technical Notes. 1997; 1(5): 17.
30. Jain AN. Ligand-based structural hypotheses for virtual screening. *J Med Chem*. 2004; 47(4):947–961. [PubMed: 14761196]
31. Ferguson AM, Heritage T, Jonathon P, et al. EVA: a new theoretically based molecular descriptor for use in QSAR/QSPR analysis. *J Comput Aided Mol Des*. 1997; 11(2):143–152.10.1023/a: 1008026308790 [PubMed: 9089432]
32. Agrafiotis DK, Cedeño W, Lobanov VS. On the use of neural network ensembles in QSAR and QSPR. *J Chem Inf Comput Sci*. 2002; 42(4):903–911.10.1021/ci0203702 [PubMed: 12132892]
33. Bender A, Jenkins JL, Scheiber J, et al. How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J Chem Inf Model*. 2009; 49(1):108–119.10.1021/ci800249s [PubMed: 19123924]
34. Glem R, Bender A, Arnby C, et al. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs*. 2006; 9(3):199–204. [PubMed: 16523386]
35. Bellis LJ, Akhtar R, Al-Lazikani B, et al. Collation and data-mining of literature bioactivity data for drug discovery. *Biochem Soc Trans*. 2011; 39(5):1365–1370. doi:10.1042/BST0391365, BST0391365 [pii]. [PubMed: 21936816]
36. Collins, J.; Crowell, J. Drug Synthesis and Chemistry Branch, Developmental Therapeutics Program (DTP), Division of Cancer Treatment and Diagnosis, National Cancer Institute. 2011. <http://dtp.nci.nih.gov/>
37. Tripos. SYBYL-X 1.2. 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA: 2012.
38. Xie XQ, Chen JZ. Data mining a small molecule drug screening representative subset from NIH PubChem. *J Chem Inf Model*. 2008; 48(3):465–475.10.1021/ci700193u [PubMed: 18302356]
39. Huffman JW, Yu S, Showalter V, et al. Synthesis and pharmacology of a very potent cannabinoid lacking a phenolic hydroxyl with high affinity for the CB2 receptor. *J Med Chem*. 1996; 39(20): 3875–3877.10.1021/jm960394y [PubMed: 8831752]

40. Morgan HL. The generation of a unique machine description for chemical structures—a technique developed at Chemical Abstracts Service. *J Chem Doc.* 1965; 5(2):107–113.10.1021/c160017a018
41. Gertsch J, Leonti M, Raduner S, et al. Beta-caryophyllene is a dietary cannabinoid. *Proc Natl Acad Sci.* 2008; 105(26):9099–9104.10.1073/pnas.0803601105 [PubMed: 18574142]
42. Raduner S, Majewska A, Chen J-Z, et al. Alkylamides from Echinacea are a new class of cannabinomimetics: cannabinoid type 2 receptor-dependent and -independent immunomodulatory effects. *J Biol Chem.* 2006; 281(20):14192–14206. [PubMed: 16547349]
43. Zhang Y, Xie Z, Wang L, et al. Mutagenesis and computer modeling studies of a GPCR conserved residue W5.43(194) in ligand recognition and signal transduction for CB2 receptor. *Int Immunopharmacol.* 2011; 11(9):1303–1310.10.1016/j.intimp.2011.04.013 [PubMed: 21539938]
44. Yang P, Wang L, Feng R, et al. Novel triaryl sulfonamide derivatives as selective cannabinoid receptor 2 inverse agonists and osteoclast inhibitors: discovery, optimization, and biological evaluation. *J Med Chem.* 2013; 56(5):2045–2058.10.1021/jm3017464 [PubMed: 23406429]
45. DePriest SA, Mayer D, Naylor CB, et al. 3D-QSAR of angiotensin-converting enzyme and thermolysin inhibitors: a comparison of CoMFA models based on deduced and experimentally determined active site geometries. *J Am Chem Soc.* 1993; 115(13):5372–5384.10.1021/ja00066a004
46. Sugimoto H, Tsuchiya Y, Sugumi H, et al. Synthesis and structure-activity relationships of acetylcholinesterase inhibitors: 1-benzyl-4-(2-phthalimidoethyl)piperidine, and related derivatives. *J Med Chem.* 1992; 35(24):4542–4548.10.1021/jm00102a005 [PubMed: 1469686]
47. Sugimoto H, Tsuchiya Y, Sugumi H, et al. Novel piperidine derivatives. Synthesis and anti-acetylcholinesterase activity of 1-benzyl-4-[2-(N-benzoylamino)ethyl]piperidine derivatives. *J Med Chem.* 1990; 33(7):1880–1887.10.1021/jm00169a008 [PubMed: 2362265]
48. Haefely W, Kyburz E, Gerecke M, et al. Recent advances in the molecular pharmacology of benzodiazepine receptors and in the structure-activity relationships of their agonists and antagonists. *Adv Drug Res.* 1985; 14:165–322.
49. Chavatte P, Yous S, Marot C, et al. Three-dimensional quantitative structure–activity relationships of cyclo-oxygenase-2 (COX-2) inhibitors: a comparative molecular field analysis. *J Med Chem.* 2001; 44(20):3223–3230.10.1021/jm0101343 [PubMed: 11563921]
50. Talley JJ, Brown DL, Carter JS, et al. 4-[5-Methyl-3-phenylisoxazol-4-yl]-benzenesulfonamide, Valdecixib: a potent and selective inhibitor of COX-2. *J Med Chem.* 2000; 43(5):775–777.10.1021/jm990577v [PubMed: 10715145]
51. Huang H-C, Li JJ, Garland DJ, et al. Diarylspiro[2.4]heptenes as orally active, highly selective cyclooxygenase-2 inhibitors: synthesis and structure–activity relationships. *J Med Chem.* 1996; 39(1):253–266.10.1021/jm950664x [PubMed: 8568815]
52. Penning TD, Talley JJ, Bertenshaw SR, et al. Synthesis and biological evaluation of the 1,5-diarylpyrazole class of cyclooxygenase-2 inhibitors: identification of 4-[5-(4-Methylphenyl)-3-(trifluoromethyl)-1H-pyrazol-1-yl]benzenesulfonamide (SC-58635, Celecoxib). *J Med Chem.* 1997; 40(9):1347–1365.10.1021/jm960803q [PubMed: 9135032]
53. Li JJ, Norton MB, Reinhard EJ, et al. Novel terphenyls as selective cyclooxygenase-2 inhibitors and orally active anti-inflammatory agents. *J Med Chem.* 1996; 39(9):1846–1856.10.1021/jm950878e [PubMed: 8627608]
54. Li JJ, Anderson GD, Burton EG, et al. 1,2-Diarylcyclopentenes as selective cyclooxygenase-2 inhibitors and orally active anti-inflammatory agents. *J Med Chem.* 1995; 38(22):4570–4578.10.1021/jm00022a023 [PubMed: 7473585]
55. Reitz DB, Li JJ, Norton MB, et al. Selective cyclooxygenase inhibitors: novel 1,2-diarylcyclopentenes are potent and orally active COX-2 inhibitors. *J Med Chem.* 1994; 37(23):3878–3881.10.1021/jm00049a005 [PubMed: 7966148]
56. Khanna IK, Yu Y, Huff RM, et al. Selective cyclooxygenase-2 inhibitors: heteroaryl modified 1,2-diarylimidazoles are potent, orally active antiinflammatory agents. *J Med Chem.* 2000; 43(16):3168–3185.10.1021/jm0000719 [PubMed: 10956225]
57. Khanna IK, Weier RM, Yu Y, et al. 1,2-diarylimidazoles as potent, cyclooxygenase-2 selective, and orally active antiinflammatory agents. *J Med Chem.* 1997; 40(11):1634–1647.10.1021/jm9700225 [PubMed: 9171873]

58. Khanna IK, Weier RM, Yu Y, et al. 1,2-Diarylpyrroles as potent and selective inhibitors of cyclooxygenase-2. *J Med Chem.* 1997; 40(11):1619–1633.10.1021/jm970036a [PubMed: 9171872]
59. Gangjee A, Vidwans AP, Vasudevan A, et al. Structure-based design and synthesis of lipophilic 2,4-diamino-6-substituted quinazolines and their evaluation as inhibitors of dihydrofolate reductases and potential antitumor agents. *J Med Chem.* 1998; 41(18):3426–3434.10.1021/jm980081y [PubMed: 9719595]
60. Rosowsky A, Mota CE, Wright JE, et al. 2,4-Diamino-5-chloroquinazoline analogs of trimetrexate and piritrexim: synthesis and antifolate activity. *J Med Chem.* 1994; 37(26):4522–4528.10.1021/jm00052a011 [PubMed: 7799402]
61. Rosowsky A, Cody V, Galitsky N, et al. Structure-based design of selective inhibitors of dihydrofolate reductase: synthesis and anti-parasitic activity of 2,4-diaminopteridine analogues with a bridged diarylamine side chain. *J Med Chem.* 1999; 42(23):4853–4860.10.1021/jm990331q [PubMed: 10579848]
62. Graffner-Nordberg M, Kolmodin K, Åqvist J, et al. Design, synthesis, computational prediction, and biological evaluation of ester soft drugs as inhibitors of dihydrofolate reductase from *Pneumocystis carinii*. *J Med Chem.* 2001; 44(15):2391–2402.10.1021/jm010856u [PubMed: 11448221]
63. Gangjee A, Elzein E, Queener SF, et al. Synthesis and biological activities of tricyclic conformationally restricted tetrahydropyrido annulated furo[2,3-d]pyrimidines as inhibitors of dihydrofolate reductases. *J Med Chem.* 1998; 41(9):1409–1416.10.1021/jm9705420 [PubMed: 9554874]

Fig. 1.
Graphical representation of the fingerprint-based ANN-QSAR (FANN-QSAR) model.
(Reprinted with permission from ref. 17. Copyright (2012) American Chemical Society)

Fig. 2. Cross-validation results of each FANN-QSAR method on CB2 ligand data set. (Reprinted with permission from ref. 17. Copyright (2012) American Chemical Society)

Fig. 3. Scatter plot between experimental pK_i and predicted pK_i values of 278 test cannabinoid ligands after the removal of 17 outliers. (Reprinted with permission from ref. 17. Copyright (2012) American Chemical Society)

Fig. 4. CB2 receptor binding affinity K_i values of four NCI hit compounds measured by [3H]CP-55940 radioligand competition binding assay using human CB2 receptors harvested from transfected CHO-CB2 cells. (Reprinted with permission from ref. 17. Copyright (2012) American Chemical Society)

Table 1

Summary of six data sets used in the FANN-QSAR model development

Target/receptor name	Number of inhibitors	pIC ₅₀ range	Reference
Angiotensin-converting enzyme (ACE)	114	2.1–9.9	[45]
Acetylcholinesterase (AChE)	111	4.3–9.5	[46, 47]
Benzodiazepine receptor (BZR)	147	5.5–8.9	[48]
Cyclooxygenase-2 (COX2)	282	4.1–9.0	[49–58]
Dihydrofolate reductase inhibitors (DHFR)	361	3.3–9.8	[59–63]
Cannabinoid receptor subtype-2 (CB2)	1,699	3.9–10.8	[20]

Table 2

Number of training, validation, and testing set compounds in each data set

	ACE	AchE	BZR	COX2	DHFR	CB2
Training set	69	67	89	170	214	1,361
Validation set	7	7	9	18	23	169
Test set	38	37	49	94	124	169
Total	114	111	147	282	361	1,699

Table 3

Comparison of results from different QSAR methods

	ECFP6-ANN-QSAR	FP2-ANN-QSAR	MACCS-ANN-QSAR	CoMFA	CoMSTIA basic	HQSAR	EVA	NN (2.5D)	NN-ens (2.5D)
ACE	0.41	0.2	0.08	0.49	0.52	0.3	0.36	0.39	0.51
AchE	0.43	0.13	0.04	0.47	0.44	0.37	0.28	-0.04	0.21
BZR	0.31	0.08	0.06	0	0.08	0.17	0.16	0.39	0.34
COX2	0.28	0.22	0.23	0.29	0.03	0.27	0.17	0.31	0.32
DHFR	0.63	0.43	0.48	0.59	0.52	0.63	0.57	0.42	0.54

Table 4

A summary of the performance of each FANN-QSAR model on CB2 ligand data set

Round	r^2 Training	r^2 Test
ECFP6-ANN-QSAR		
1	0.86	0.55
2	0.81	0.63
3	0.87	0.53
4	0.84	0.56
5	0.89	0.54
FP2-ANN-QSAR		
1	0.78	0.55
2	0.74	0.60
3	0.74	0.38
4	0.77	0.46
5	0.79	0.40
MACCS-ANN-QSAR		
1	0.74	0.48
2	0.72	0.53
3	0.74	0.37
4	0.74	0.47
5	0.75	0.41

Table 5

Identified NCI hit compounds with CB2 binding activities

Structure	NSC ID	MW	ClogP	Experimental p <i>K_i</i>	Predicted p <i>K_i</i>
746843	400.55	6.61	8.81 ^a	8.66
49888	330.46	5.59	8.17 ^b	8.28
174122	463.52	4.76	5.59 ^c	8.41
369049	488.66	4.00	5.51 ^c	8.48

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Structure	NSC ID	MW	ClogP	Experimental p <i>K</i> _i	Predicted p <i>K</i> _i
76301	354.44	3.99	5.42 ^c	8.21

^a An average literature reported *K*₁ value of a known cannabinoid compound (HU210) which is more than 90 % similar to 746843

^b An average *K*₁ value of two independent experiments performed in duplicate

^c An experimental *K*₁ value of one experiment performed in duplicate