Artificial Intelligence in the Life Sciences 2 (2022) 100039

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Research Article

Classification of JAK1 Inhibitors and SAR Research by Machine Learning Methods

Zhenwu Yang[1], Yujia Tian[1], Yue Kong[3], Yushan Zhu[2], Aixia Yan[1,*]

[1] State Key Laboratory of Chemical Resource Engineering, Department of Pharmaceutical Engineering, P.O. Box 53, Beijing University of Chemical Technology, 15 BeiSanHuan East Road, Beijing 100029, P. R. China
[2] National Energy R&D Center for Biorefinery, College of Life Science and Technology, Beijing University of Chemical Technology, Beijing 100029, P. R China
[3] Hyper-Dimension Insight Pharmaceuticals Ltd. Room 511, Block A, No. 2 C, DongSanHuan North Road, ChaoYang District, Beijing, P. R. China

ARTICLE INFO

Keywords:
Deep neural networks (DNN)
Janus kinase 1 (JAK1) inhibitor
Molecular modeling
Structure-activity relationship
Substructure analysis

ABSTRACT

Janus kinase 1 (JAK1) is a key regulator of gene transcription, inhibition of JAK1 is an intervention for many diseases including rheumatoid arthritis and Crohn's disease. In this study, we collected a dataset containing 2982 JAK1 inhibitors, characterized molecules by MACCS fingerprints and Morgan fingerprints. We used support vector machine (SVM), decision tree (DT), random forest (RF) and extreme gradient boosting tree (XGBoost) algorithms to build 16 traditional machine learning classification models. Additionally, we utilized deep neural networks (DNN) to develop four deep learning models. The best model (Model 3B) built by RF and Morgan fingerprints achieved the accuracy (ACC) of 93.6% and Mathews correlation coefficient (MCC) of 0.87 on the test set. Furthermore, we made structure–activity relationship (SAR) analyses for JAK1 inhibitors, based on the output from the random forest models. After analyzing the important keys of two types of fingerprints, it was observed that some substructures such as pyrazole, pyrrolotriazolopyrimidine and pyrazolopyrimidine appeared frequently in highly active JAK1 inhibitors.

Introduction

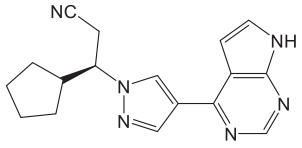
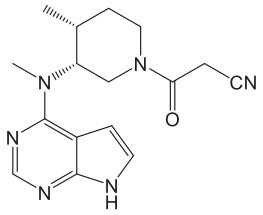
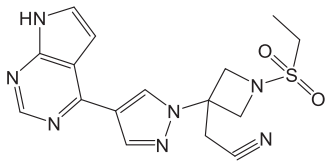
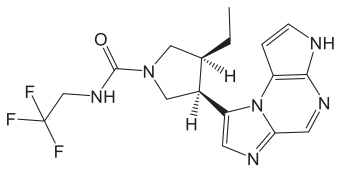
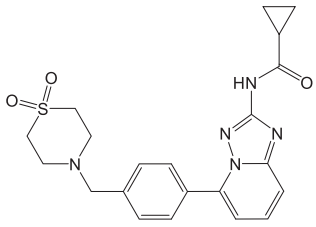
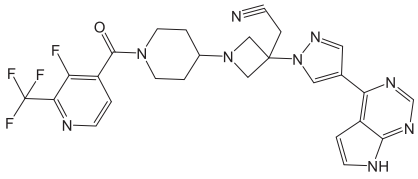
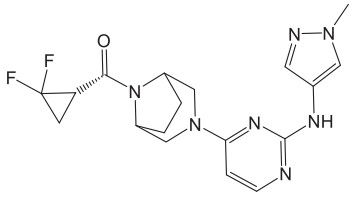
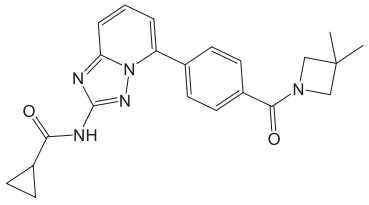
The JAK family proteins, as non-receptor protein tyrosine kinases, play a vital role in both immune cells and hematopoietic cells. They are also involved in cell growth, survival, development and differentiation [1]. To date, four members of this family have been identified (JAK1, JAK2, JAK3 and TYK2) [2,3], which transduce signaling from cytokine receptors by phosphorylation and subsequent activation of signal transducers and activators of transcription (STATs) [1]. The JAK/STAT signal pathway regulated by JAKs is a key regulator of gene transcription and it is also known to be activated by more than fifty different cytokines receptors, receiving signals from pro-inflammatory cytokines, inflammatory cytokines, hematopoietic cell growth factors and metabolic cytokines. Depending on the cytokines receptor that get activated, different JAK/STAT pathways get stimulated [4]. Consequently, JAK inhibition has been proposed as a potential therapeutic intervention for various myeloproliferative and inflammatory diseases, including myeloproliferative neoplasms (MPNs), rheumatoid arthritis (RA), psoriasis and inflammatory bowel disease (IBD) [5]. In addition, some researchers have suggested that persistent activation of JAK1, JAK2 and STAT3 causes the proliferation of cancer cell lines [6]. Biochemical and genetic studies have shown that JAK1 is the most broadly used JAK. In vitro studies using JAK1-deficient cells demonstrated its critical role in the signal transduction mediated by type I and type II IFN (e.g.IFN-α/β and IFN-γ) as well as IL-10. Moreover, JAK1 is involved in signaling downstream of cytokines that utilize the γc receptor subunit including IL-2, IL-4, IL-7, IL-9, IL-15 and IL-21 and regulates a key group of pro-inflammatory cytokines, including IL-6 and others that utilize the gp130 subunit such as IL-11, leukemia inhibitory factor (LIF), oncostatin M (OSM), ciliary neurotrophic factor and G-CSF [7,8]. Therefore, it is important to develop new effective JAK1 inhibitors for the treatment of certain inflammations, autoimmune diseases and cancer.

As shown in Table 1, the US Food and Drug Administration (FDA) approved Ruxolitinib in 2011, 2014 and 2019 for the treatment of intermediate/high-risk myelofibrosis (MF) [9], hydroxyurea (HU) insufficiency or intolerable polycythemia vera (PV) [10] and steroid-refractory (SR) acute graft-versus-host disease (GVHD) [11], EMA approved it for the treatment of PV in 2015. Phase III trials based on Ruxolitinib in the treatment of atopic dermatitis (AD) [12] have also been carried out in North America and Europe. Tofacitinib is a pan-JAK inhibitor that was first used to treat rheumatoid arthritis (RA) [13]. It has been approved successively in the United States and Japan, and is currently recognized all over the world. Recently, Tofacitinib has been approved for the treatment of Psoriatic arthritis (PsA) [14] by various international regulatory authorities, including the FDA, EMA,

* Corresponding author.
E-mail address: yanax@mail.buct.edu.cn (A. Yan).

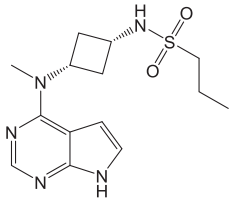
https://doi.org/10.1016/j.ailsci.2022.100039
Received 11 April 2022; Received in revised form 2 June 2022; Accepted 6 June 2022
Available online 8 June 2022

Table 1
Approved drugs of JAK1 and promising JAK1 inhibitors under clinical trials.

Drug[a]	Structure[b]	Status[c]	Diseases[d]
Ruxolitinib (INC424)		FDA Approved FDA Approved, EU Approved FDA Approved Phase III	MF [9] PV [10] GVDH [11] AD [12]
Tofacitinib (CP690,550)		FDA Approved, Japan Approved FDA Approved, EMA Approved, NICE Approved Phase III Phase III	RA [13] PsA [14] UC [15] AS [16]
Baricitinib (INCB28050)		FDA Approved, EMA Approved Phase II	RA [17,18] SLE [26]
Upadacitinib (ABT494)		FDA Approved, EMA Approved	RA [19,20]
Filgotinib (GLPG0634)		EU Approved, Japan Approved Phases III	RA [21] UC [22]
Itacitinib (INCB039110)		Phase I Phase II	aGVHD [23] cHL [24]
PF-06700841		Phase II	Psoriasis [25]
Solcitinib (GSK2586184)*		Phase II	SLE [27]

(*continued on next page*)

Table 1 (*continued*)

Drug[a]	Structure[b]	Status[c]	Diseases[d]
PF-04965842		Phase III	AD [28]

aThe name of the drug, * means that the drug has been stopped for further development.
bStructural formula of the drug.
cThe development stage of the drug. FDA: US Food and Drug Administration; EMA: European Medicines Agency; EU: European Union; NICE: UK National Institute of Health and Clinical Excellence.
dGVHD: graft-versus-host disease; aGVHD: acute graft-versus-host disease; RA: rheumatoid arthritis; AA: alopecia areata; UC: ulcerative colitis; AD: atopic dermatitis; SLE: systemic lupus erythematosus; cHL: classical Hodgkin lymphoma; MF: Myelofibrosis; PV: polycythemia vera; PsA: Psoriatic arthritis; AS: ankylosing spondylitis.

and UK National Institute of Health and Clinical Excellence (NICE). At the same time, Phase III trials for ulcerative colitis (UC) [15] and ankylosing spondylitis (AS) [16] are in progress. In addition to Tofacitinib, Baricitinib, Upadacitinib and Filgotinib have been approved for the treatment of RA [17–21] in the United States, the European Union and Japan. Moreover, Baricitinib, Filgotinib, Itacitinib, Solcitinib, PF-06700841 and PF-04965842 are undergoing clinical Phase I to Phase III trials, and their indications include UC [22], acute graft-versus-host disease (aGVHD) [23], classic Hodgkin's lymphoma (cHL) [24], Psoriasis [25], SLE [26,27] and AD [28]. However, the failure of a trial of Solcitinib on SLE led to a halt in clinical trials.

The exploration of structure-activity relationship (SAR) is one of the most important tasks in medicinal chemistry. SAR analysis provides a basis for key structural features to determine molecular activity, and also provides a basis for screening clinically-oriented candidate molecules [29]. Quantitative structure-activity relationship (QSAR) is a chemical data analysis method that predicts molecular properties by establishing a linear or non-linear relationship between the descriptors calculated from molecular structures and the biological activity values of these molecules [30]. Some machine learning (ML) methods were applied to construct SAR and QSAR models, such as decision tree (DT) [31], random forest (RF) [32], support vector machine (SVM) [33], extreme gradient boosting (XGBoost) [34], and artificial neural networks (ANNs) [35]. Molecular descriptors were used to characterize the physical and chemical properties or structural characteristics of compounds. In modeling, different descriptors needed to be tried to achieve the best effect of the model.

Several published computational works on JAK1 inhibitors provided analysis of the structural requirements of the selected series for JAK1 inhibitory activity. Sarithamol et al. [36]. collected 100 pyrazole derivatives, established a 3D-QSAR comparison model of JAK1 and JAK2, Q^2 of 0.8243 and 0.6917 were obtained on the test set. Itteboina et al. [37] established a 3D-QSAR model of 30 imidazopyrrolopyridine derivatives, and evaluated the performance of the model with 13 molecules as a test set.

The ${q_{loo}}^2$ of 0.504, ${r_{ncv}}^2$ of 0.948, and the ${r_{pred}}^2$ of 0.52 were obtained for the CoMFA model; The ${q_{loo}}^2$ of 0.518, ${r_{ncv}}^2$ of 0.951, and the ${r_{pred}}^2$ of 0.53 obtained for the CoMSIA model. Subsequently, the team established a 3D-QSAR model of 60 JAK1 inhibitor molecules [7], and verified the model with 25 molecules. In the end, Q^2 of 0.525 and 0.534, ${r_{pred}}^2$ of 0.52 and 0.54 were obtained. Keretsu et al. [4] conducted a 3D-QSAR study on 51 pyrrolopyrimidin-4-amine derivatives and established a ligand-based CoMFA model ($Q^2 = 0.5$, $R^2 = 0.96$) and a receptor-based CoMFA model ($Q^2 = 0.78$, $R^2 = 0.98$). All of the above works are based on small data set to build a model. In this work, we build our models on a data set fo 2982 inhibitors, the largest data set to date.

The purpose of this study is to construct JAK1 inhibitor highly/weakly activity classification models, then analyze its structure-activity relationship. Five machine learning algorithms including support vector machine (SVM), decision tree (DT), random forest (RF), extreme gradient boosting (XGBoost) and deep learning algorithm (DNN) [38] were used to establish classification models of JAK1 inhibitors. MACCS molecular fingerprints [39] and Morgan molecular fingerprints [40] were used to characterize the molecular structures during modeling. Additionally, a SAR analysis of important fingerprints was performed to identify important sub-structural features of highly active JAK1 inhibitors.

Materials and methods

Dataset

We collected information of JAK1 inhibitors from three databases of CHEMBL [41], Reaxys [42], scifinder [43] and 65 pieces of literature [44–108]. The biological activity of each inhibitor is characterized by IC_{50}.

We checked the inhibitors (enzymatic method to measure activity, and the receptor was homo sapiens single protein) downloaded from the database one by one, and then we removed duplicate data to construct the JAK1 inhibitor dataset.

For the above dataset, the IC_{50} values of these inhibitors ranged from 0.07 nM to 50000 nM. We referenced three studies to define compounds [6,109,110]. We deleted the compounds with IC_{50} values between 50 nM and 100 nM, then defined compounds with IC_{50} less than 50 nM as highly active, and compounds with IC_{50} greater than 100 nM as weakly active. As a result, the whole dataset was composed of 2982 JAK1 inhibitors, including 1712 highly active and 1270 weakly active inhibitors. The whole dataset is shown in JAK1_dataset.csv in the Supplementary materials.

Splitting the dataset into a training set and a test set

We used two methods to divide the dataset into a training set and a test set. (1) random splitting: we utilized the random split function in scikit-learn [111] module of Python to split the dataset into a training set contained 2236 inhibitors (1279 highly active and 957 weakly active inhibitors) and a test set with 746 inhibitors (433 highly active and 313 weakly active inhibitors). (2) self-organizing map (SOM) [112]: we used SOM to split 2982 inhibitors by the SONNIA software [113], the detailed process according to the following steps:

(I) The inhibitors in each neuron were split separately.
(II) For highly active neurons and weakly active neurons, if there is only one inhibitor in the neuron, then this inhibitor was assigned

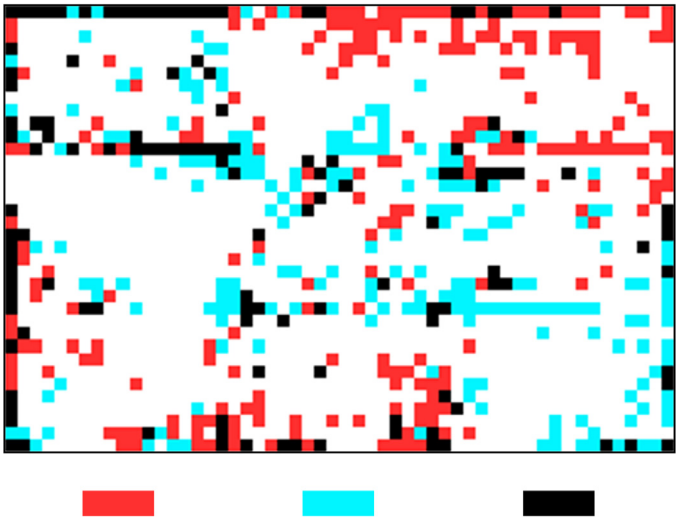


Figure 1. SOM clustering results (36*54), the input are 166-bit MACCS descriptors. Highly active inhibitors, which are represented with red; weakly active inhibitors, which are represented with blue; conflict neurons accommodate both highly active inhibitors and weakly active inhibitors, and represented with black.

into the training set; if there were two inhibitors in the neuron, then one of them was randomly split into the training set, the other was split into the test set; if there were three inhibitors in the neuron, then two inhibitors were randomly split into the training set and one was split into the test set; when the number of inhibitors in the neuron was greater than or equal to four, the part which can be divisible by four was split at a ratio of 3:1, the remaining part was processed according to the above steps.

(III) For conflicting neurons, we individually split the highly active inhibitors and weakly active inhibitors according to the second step.

Through the above process of SOM splitting, we obtained a training set with 2219 inhibitors (highly: weakly = 1277: 942), and a test set containing 763 inhibitors (highly: weakly = 435: 328). The distribution of SOM result is shown in Figure 1.

Molecular fingerprint

Molecular fingerprinting is a molecular characterization method that transforms a molecular structure into a bitstream by judging whether the molecule has a specific substructure. In this study, we used two types of molecular fingerprints, MACCS (166 bits) and Morgan fingerprint (4096 bits) to characterize molecular structures, both of which were calculated by the RDKit [114] toolkit in Python.

Extended connection fingerprints (ECFPs) were a kind of Morgan fingerprints, which were ring fingerprints based on the ECFP algorithm (a variant of the Morgan algorithm) [40]. According to different needs, by changing the radius and bits, Morgan fingerprints can theoretically characterize molecules of any size and any number of molecular features. As the radius and bits increase, the effective information contained in Morgan fingerprints will gradually increase. However, as the number of bits increase, Morgan fingerprints will become very redundant sparse matrices (most of the information is 0). Therefore, it is necessary to select the appropriate radius and bits according to molecules with different characteristics. This feature of Morgan fingerprints provides a high degree of freedom for fingerprint analysis. In this study, we calculated 4096-bit Morgan fingerprints with a radius of four. For

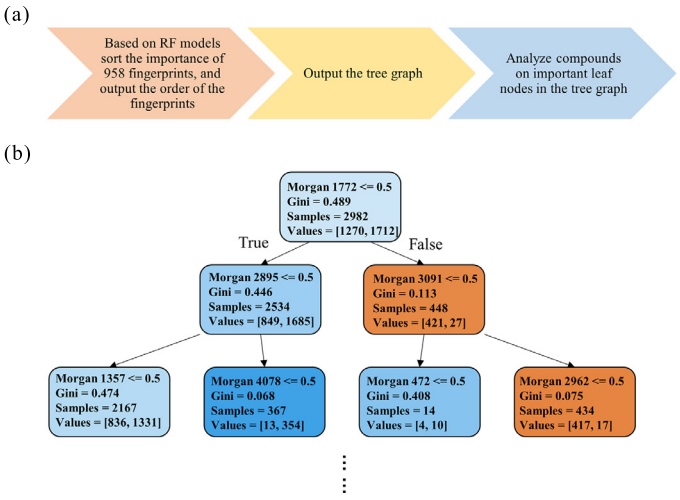


Figure 2. (a) Fingerprint analysis process; (b) Tree graph, take the root node as an example: 'Morgan 1772 <= 0.5' means to determine whether molecules in the next node contain the Morgan 1772 key, 'True' means molecules in the next node do not contain the Morgan 1772 key and 'False' means molecules in the next node contain the Morgan 1772 key; 'Gini' represents Gini index, which determines the split of the tree; 'Samples' represents the number of data that meet the judgment conditions; The left number in 'Values' represents the number of weakly activity molecules, the right number in 'Values' represents the number of high activity molecules.

the two molecular fingerprints, we calculated the variance on each bit, then removed bits with variance less than the average variance. The purpose was to avoid invalid information due to too consistent information on a certain position. The final molecular fingerprints used for modeling were 80 bits (MACCS) and 958 bits (Morgan fingerprints), respectively.

Method of fingerprint analysis

In order to analyze the descriptors more accurately, we used all the compounds to train the model, calculated the information entropy [31] of each descriptor according to the RF model and ranked the importance of each descriptor. Additionally, we visualized the decision tree model using sklearn.tree.export_graphviz [111], and analyzed the fingerprints on important leaf nodes in the tree graph. The process and the tree graph are shown in Figure 2. The color of the nodes in the Figure 2 represents the purity of the compounds. The more weakly active compounds inside a node, the closer the color is to a cool tone; Conversely, the more highly active compounds inside a node, the closer the color is to a warmer tone We performed SAR analysis on each descriptor based on the ranking results as we did in our previous work [115]. We only focus on some fingerprints with the highest importance. For a leaf node at an end, we first determined the path between it and the root node, and then determined the presence or absence of fingerprints represented by each passing node along the path. According to these information, the compounds in the leaf node can be obtained from the dataset. Based on the fingerprints of these compounds, we summarized their scaffolds and found representative compounds among them. These scaffolds were composed of different combined form of the same fingerprints may have different effects on the biological activity of the compounds. Descriptor importance ranking results of the top 80 MACCS fingerprints and top the 100 Morgan fingerprints are displayed in Table S1 and Table S2 in the Supplementary materials.

Chemical Diversity

To evaluate the chemical diversity of the dataset, We calculated the SlogP and molecular weight of all compounds, The SlogP distribu-

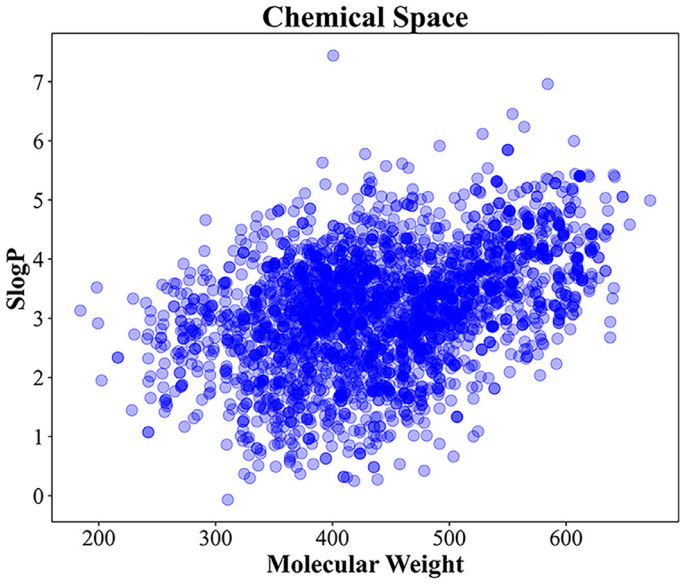


Figure 3. Visualization of SlogP and molecular weight for all compounds. The darker the color, the denser the compound distribution in the area.

tion of these molecules was between -0.07 and 7.44, and the molecular weight distribution was between 184.24 and 671.80. The visualization results are shown in Figure 3. The wide distribution of these two properties also indicated the expansive chemical space of our dataset.

We measured the diversity by Murcko Scaffold of our dataset. Murcko Scaffold is an effect way of gathering together common scaffolds and can be really useful to check for diversity and for clustering. We obtained the main scaffold of each molecule in our dataset by removing its side chains (any nonring, nonlinker atoms are defined as side chain atoms) [116]. Then, we obtained the Murcko scaffold by using the MurckoScaffold module in RDKit [114]. As a result, there were 1057 main scaffolds and 485 Generic Murcko scaffolds in our dataset (listed in files named "Main_Scaffold.csv", and "Generic_Murcko_Scaffold.csv" in supplementary materials). These observation can illustrate the chemical diversity of our dataset.

In addition, we calculated the Tanimoto index [117] between the compounds in the training and test sets divided by each method based on ECFPs of length 1024. The Tanimoto index is shown in formula (1), and the histogram of the calculation results is shown in Figure 4. where V_i and V_j are the descriptors describing the two molecular features, respectively.

$$Tani(V_i, V_j) = \frac{\left|V_i \cap V_j\right|}{\left|V_i\right| + \left|V_j\right| - \left|V_i \cap V_j\right|} \tag{1}$$

The Tanimoto index represents the similarity of two sets, It can be seen from Figure 4 that the Tanimoto similarity of the compounds in the training set and the test set divided by the random method or the SOM method are all concentrated in the range of less than 0.6, accounting for 98.25% and 98.19, respectively. The similarity of compounds in the training set and test set divided by this method is very low. Therefore, it is objective to evaluate the model performance according to the prediction on test set.

Machine learning algorithms

The algorithms involved in this research include: support vector machine (SVM), decision tree (DT), random forest (RF), extreme gradient boosting tree (XGBoost) and deep neural networks (DNN) were used to build classification models.

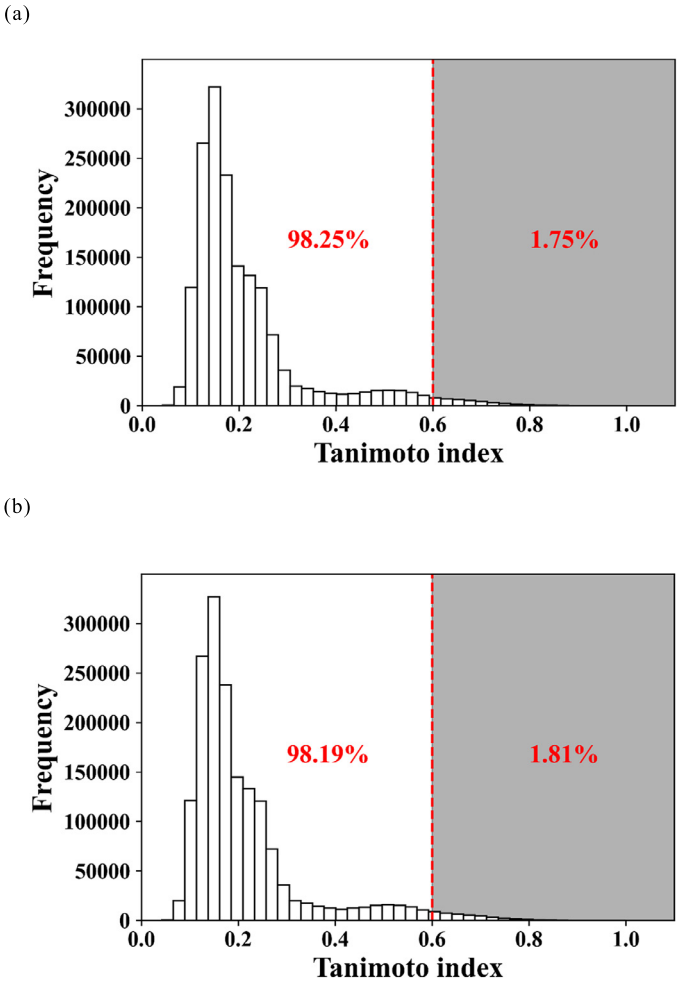


Figure 4. Frequency histogram of Tanimoto index of compounds between training set and test set based on ECFPs fingerprints. (a) Frequency histogram of Tanimoto index of compounds between training set and test set under random method, (b) Frequency histogram of Tanimoto index of compounds between training set and test set under SOM method. The percentages represent the frequency of Tanimoto index with Tanimoto index greater than 0.6 and less than 0.6, respectively.

Decision tree

A decision tree (DT) is a tree structure in which each internal node represents a test on an attribute, each branch represented a test output, each leaf node represented a category. Training a decision tree includes two processes: the growth phase and the pruning phase. The set splitting criteria for the growth phase are shown in formulas (2) and (3)[31]:

$$Gini = \sum_{k=1}^{K} \hat{p}_k \left(1 - \hat{p}_k\right) \tag{2}$$

$$Cross\ entropy = \sum_{k=1}^{K} \hat{p}_k log \hat{p}_k \tag{3}$$

Formula (2) is the Gini index, and formula (3) is the cross-entropy index, where $\hat{p}_k$ is the fraction of samples of each class k associated to a given subset S_i.

$$\alpha = \frac{err\left(T_{i+1}, X'\right) - err\left(T_i, X'\right)}{leaves\left(T_i\right) - leaves\left(T_{i+1}\right)} \tag{4}$$

Formula (4) [31] is the pruning criterion, where $err(T, X')$ is the error rate of tree T on the set of observations X', $leaves(T)$ is the number

of leaves of tree T. After the evaluation of all the collapsed sub-trees of T_i, the sub-tree that gets the lowest value of α is taken as T_{i+1}. α is the complexity parameter. In this study, the parameters that the DT algorithm needs to be optimized are *criterion, max_features, max_depth*, and *max_leaf_nodes*. We used the grid optimization method to find the optimal parameter combination in the four-dimensional space.

Random forest

Random forest (RF) is an ensemble method based on bagging [118]. It randomly builds and integrates multiple decision tree to create a forest structure. In order to classify the dataset, each decision tree is given a classification, the score of each tree is calculated to make the final decision. The process was repeated many times and the final prediction is a function of each prediction derived from different constituent decision tree [32]. In this study, the parameters that need to be optimized for the RF algorithm are *criterion, max_features, max_depth, max_leaf_nodes* and *n_estimators*. We used the grid optimization method to find the optimal parameter combination in the five-dimensional space.

Support vector machine

Support vector machine (SVM) used sum function to transform data into high-dimensional space and established the optimal separation hyperplane [33].

$$K(x_1, x_2) = \phi(x_1)^T \phi(x_2) \tag{5}$$

$$L_p = \frac{1}{2}\|\vec{w}\| - \sum_{i=1}^{t} \alpha_i y_i (\vec{w} \cdot \vec{x}_i + b) + \sum_{i=1}^{t} \alpha_i \tag{6}$$

The kernel function K was expressed as Formula (5), which accepts two parameters, used a specific mapping for the parameters, and then returns their dot product value. Suppose x_1 and x_2 be two data points, ϕ is a mapping [119]. Formula (6) represents the hyperplane defined by $\vec{w}$ and the constant b, where t is the number of training examples, α_i, $i = 1,\ldots, t$, are non-negative numbers, so that the derivative of L_p to α_i is zero, α_i is the Lagrange multiplier and L_p is called the Lagrangian [120]. In this research, the hyperparameters *C* and *gamma* of SVM were determined by the grid optimization method.

Extreme gradient boosting

XGBoost is an algorithm optimization of the gradient boosting decision tree, which can greatly improve the computing performance. These optimizations include new tree learning algorithms for processing sparse data; a theoretically justified weighted quantile sketch procedure enables handling instance weights in approximate tree learning. In addition, parallel and distributed computing greatly improved the learning speed. More importantly, XGBoost utilizes out-of-core calculations to greatly increase the number of data calculations [34]. However, too many adjustable parameters caused XGBoost to require more computing power during the optimization process. In this study, the parameters that needed to be optimized were *gamma, n_estimators, max_depth, min_child_weight, subsample, colsample_bytree, reg_alpha*, and *learning_rate*. We adopted a step-by-step optimization strategy, first optimized the first four parameters, and then optimized the last four parameters.

Deep neural network

Artificial Neural Networks (ANN) deal with complex problems by imitating the neural structure of the brain. They are systems that can modify their internal structure according to functional goals, and are particularly suitable for solving non-linear problems [35]. A deep neural network (DNN) is a neural network with multiple hidden layers. It continuously updates its internal parameters through the backpropagation algorithm [38]. Due to the more complex network structure, DNN has a better fitting effect than ANN [38]. Therefore, it has been widely used in many fields in recent years.

In our research, DNN was built using the pytorch [121] toolkit in python, model evaluation was performed using the scikit-learn [111] toolkit, and the matplotlib [122] toolkit was used for drawing.

We built four DNN models based on two split methods and two types of descriptors. For the two types of descriptors, we used different network structures due to different inputs. We added a regularization layer and Relu activation function between every two hidden layers. The regularization layer can effectively prevent overfitting. The Relu function can prevent the gradient from disappearing and speed up the training process. When outputting, used sigmoid as the activation function. The sigmoid function has a relatively good effect in binary classification problems. In each model training process, we trained in two stages, each stage used a different optimizer and loss function, used the batch training method, and always keeps each batch containing 25 compounds. The first stage used the Adam optimizer, BCE loss function and relatively large learning rate. The Adam optimizer was computationally efficient, suitable for large-scale data and parameter scenarios. The optimal model was found from the first stage, it was reloaded for the second stage training. We can reduce the learning rate or keep it unchanged, depending on the degree of convergence of the model. In the second stage, we used the SGD optimizer to speed up the training process. Finally, we selected the best model from the second stage of training. The network structures of all models are shown in Fig. S1 to S4 in the Supplementary materials.

Model evaluation index

The model evaluation indicators used in this study were: accuracy (ACC), Matthews correlation coefficient (MCC), sensitivity (SE), specificity (SP), their calculation methods were as formula (7-10).

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$\text{MCC} = \frac{TP*TN - FP*FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{8}$$

$$\text{SE} = \frac{TP}{TP + FN} \tag{9}$$

$$\text{SP} = \frac{TN}{TN + FP} \tag{10}$$

Here *TP* is true positive, *TN* is true negative, *FP* is false positive, *FN* is false negative. triple cross-validation (3-CV), five-fold cross-validation (5-CV), ten-fold cross-validation (10-CV), leave-one-out validation (LOO) and ROC curve are also used as evaluation indicators in the training process.

Results and discussion

Results and performances of machine learning models

Based on the two types of molecular fingerprints, two training/test set splitting methods, and four traditional machine learning algorithm, 16 classification models were built: four SVM models (models 1A, 1B, 1C, 1D), four DT models (models 2A, 2B, 2C, 2D), four RF models (models 3A, 3B, 3C, 3D), four XGB models (models 4A, 4B, 4C, 4D). The optimal parameters for all models are displayed in Table S3 to S6 in the Supplementary materials.

As shown in Table 2, all models had ACC and MCC values above 84.6% and 0.68 on the test set, respectively. Except for DT models, SVM, RF, and XGB algorithms performed comparable well on the bioactivity prediction. Comparing with those three algorithms, DT models showed relatively poorer performances. The slightly weak fitting and generalization ability of the DT algorithm is due to the following two limitations: the decision boundary of DT can only be parallel to the coordinate axis; DT would be particularly sensitive to individual data. The ROC curves

Table 2
Performances of classification models which were developed by traditional machine learning algorithms (SVM, DT, RF, and XGB).

				Training set						Test set			
Model ID	Algorithm	Training/Test set[a]	Descriptor	ACC(%)	3-CV(%)[b]	5-CV(%)[b]	10-CV(%)[b]	LOO(%)[c]	MCC	ACC(%)	SE(%)[d]	SP(%)[d]	MCC
Model 1A	SVM	2219/763	MACCS	96.7	90.5	91.3	91.4	91.5	0.93	92.7	94.7	89.9	0.85
Model 1B			Morgan	97.8	92.3	92.8	93.1	92.8	0.95	93.6	95.9	90.5	0.87
Model 1C		2236/746	MACCS	97.0	91.1	91.6	92.1	92.2	0.94	91.8	92.8	90.4	0.83
Model 1D			Morgan	96.7	92.8	93.2	93.5	93.6	0.93	92.1	93.1	90.7	0.84
Model 2A	DT	2219/763	MACCS	90.9	86.8	87.0	87.9	87.6	0.82	87.8	87.8	87.8	0.75
Model 2B			Morgan	94.4	90.0	90.0	91.6	91.0	0.88	91.5	92.9	89.6	0.83
Model 2C		2236/746	MACCS	91.8	87.3	87.3	88.6	88.6	0.83	84.6	87.3	80.8	0.68
Model 2D			Morgan	94.9	91.0	91.8	92.7	92.3	0.89	89.8	91.2	87.9	0.79
Model 3A	RF	2219/763	MACCS	93.4	89.0	89.3	90.3	90.0	0.87	91.2	92.2	89.9	0.82
Model 3B			Morgan	94.9	92.2	92.5	92.7	91.9	0.89	93.6	95.6	90.9	0.87
Model 3C		2236/746	MACCS	94.6	90.4	90.4	90.7	90.7	0.89	89.1	90.8	86.9	0.78
Model 3D			Morgan	94.7	92.3	92.4	93.2	92.6	0.89	91.2	92.6	89.1	0.82
Model 4A	XGB	2219/763	MACCS	97.5	90.3	90.8	91.4	90.8	0.95	92.4	94.9	89.0	0.84
Model 4B			Morgan	95.4	91.7	92.2	92.7	92.2	0.95	93.3	95.4	90.5	0.86
Model 4C		2236/746	MACCS	97.4	90.4	91.1	91.8	92.0	0.95	90.2	92.6	87.0	0.80
Model 4D			Morgan	97.3	92.5	92.6	93.2	92.4	0.94	92.0	93.8	89.5	0.83

[a] The number of JAK1 inhibitors in the training set or test set. "2219/763" represents the highly active/ weakly active inhibitor sets obtained by SOM splitting method; "2236/746" represents the highly active/ weakly active inhibitor sets obtained by random splitting method.
[b] k-fold cross validation, k = 3, 5, 10.
[c] leave-one-out verification.
[d] SE: sensitivity; SP: specificity.

Table 3
Performances of DNN models.

			Training set				Test set			
Model ID	Training/Test set[a]	Descriptor	ACC(%)	SE(%)[b]	SP(%)[b]	MCC	ACC(%)	SE(%)[b]	SP(%)[b]	MCC
Model 5A	2219/763	MACCS	97.61	96.6	98.4	0.95	93.1	90.2	95.2	0.86
Model 5B		Morgan	98.56	99	98.2	0.97	94.5	92.7	95.9	0.89
Model 5C	2236/746	MACCS	97.72	96.8	98.4	0.95	91.6	90.1	92.6	0.83
Model 5D		Morgan	97.45	96.3	98.3	0.95	92.8	89.5	95.2	0.85

[a] The number of JAK1 inhibitors in the training set or test set. "2219/763" represents the highly active/ weakly active inhibitor sets obtained by SOM splitting method; "2236/746" represents the highly active/ weakly active inhibitor sets obtained by random splitting method.
[b] SE: sensitivity; SP: specificity.

of all models are displayed in Fig. S5 to S8 in the Supplementary materials. The AUC values of all traditional machine learning models are above 0.88.

From the perspective of molecular characterization, all models built with Morgan fingerprints performed better than those built with MACCS on the test set, indicating that Morgan fingerprints were more suitable for our dataset. For the strategy of splitting dataset, all models based on training/test set splitting by SOM method performed better than those by the random splitting method on the test set, which verified the rationality of the dataset constructed by the SOM splitting method.

Table 3 shows the performance of the four models constructed using DNN algorithm. Among them, Model 5A constructed by MACCS fingerprints and SOM splitting method; Model 5B constructed by Morgan fingerprints and SOM splitting method; Model 5C constructed by MACCS fingerprints and random splitting method; Model 5D constructed by Morgan fingerprints and random splitting method. The ROC curves of all DNN models are shown in Fig. S9 and the learning curves of them are shown in Fig. S10 to S13 in the Supplementary materials. The AUC values of all DNN models are above 0.96.

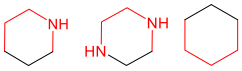
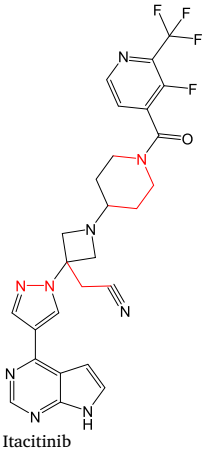
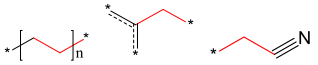
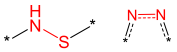
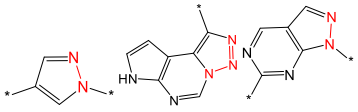
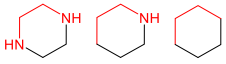
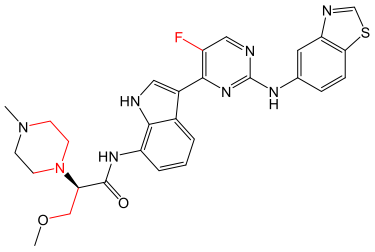
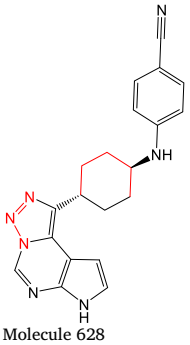
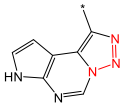
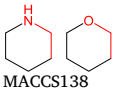
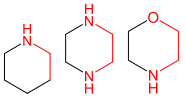
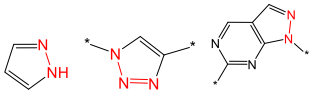
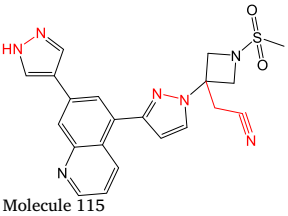
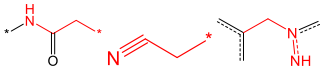
In order to measure the excellence of the models, we used statistical tests to quantify the performance differences between the models. Delong test [123] was used to compare the differences between two models' AUC values. To control the family-wise error rate (FWER) [124] due to the multiple pairwise comparisons, we used the false discovery rate (FDR) correction [125] to adjust p values for multiple pairwise com-

Table 4
The p values between Model 3B and other 19 models by using multiple pairwise statistical test.

Model 1 (I)	Model 2 (J)	AUC (I)[a]	AUC (J)	p value[b]	p value (adj)[c]
Model 3B	Model 2C	0.9845	0.8824	>0.0001	>0.0001
Model 3B	Model 2A	0.9845	0.924	>0.0001	>0.0001
Model 3B	Model 2B	0.9845	0.9615	>0.0001	>0.0001
Model 3B	Model 2D	0.9845	0.927	>0.0001	>0.0001
Model 3B	Model 3C	0.9845	0.9562	0.0003	0.0012
Model 3B	Model 3A	0.9845	0.9723	0.0007	0.0023
Model 3B	Model 5C	0.9845	0.9619	0.0026	0.0070
Model 3B	Model 5A	0.9845	0.9723	0.0032	0.0075
Model 3B	Model 4A	0.9845	0.9751	0.0048	0.0102
Model 3B	Model 1A	0.9845	0.9742	0.0054	0.0102
Model 3B	Model 1C	0.9845	0.9652	0.0078	0.0135
Model 3B	Model 4C	0.9845	0.967	0.0097	0.0153
Model 3B	Model 5D	0.9845	0.9688	0.0232	0.0339
Model 3B	Model 3D	0.9845	0.9694	0.0301	0.0408
Model 3B	Model 1D	0.9845	0.9716	0.0532	0.0674
Model 3B	Model 4D	0.9845	0.9718	0.0564	0.0669
Model 3B	Model 1B	0.9845	0.9824	0.4548	0.5083
Model 3B	Model 5B	0.9845	0.9829	0.5681	0.5997
Model 3B	Model 4B	0.9845	0.9852	0.6585	0.6585

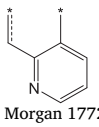
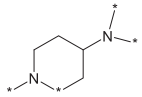
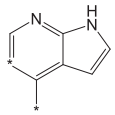
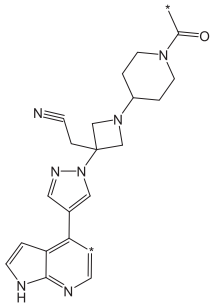
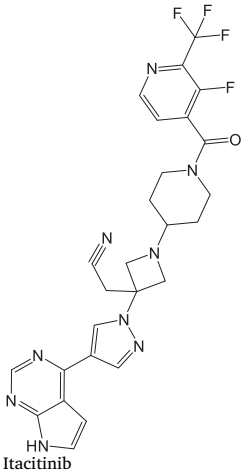
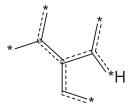
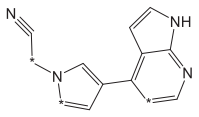
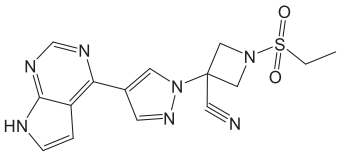
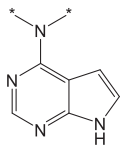
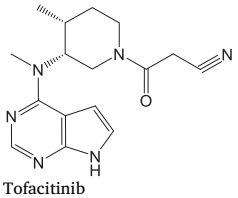
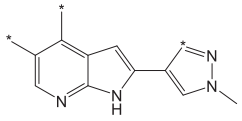
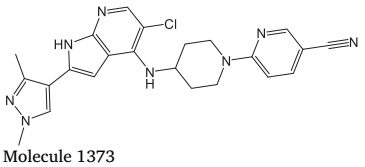
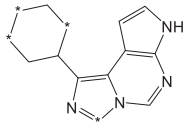
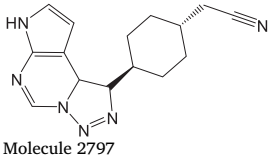
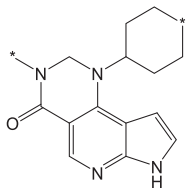
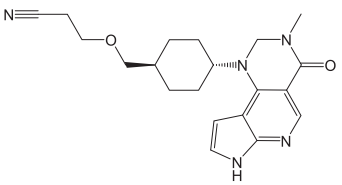
[a] *AUC (I)* is the AUC value of *Model 1 (I), AUC (J)* is the AUC value of *Model 1 (J)*;
[b] *p* value based on Delong test;
[c] Adjusted *p* value based on FDR correction.

Table 5
MACCS fingerprint key and sub-structure analysis.

Classes	MACCS key[a]	Highly/ Weakly[b]	Examples of MACCS key[c]	Representative substructures[d]	Representative compounds[e]
Class 1	MACCS129(1), MACCS155(1), MACCS94(1), MACCS118(1)	1144/117	 MACCS129 ACH2AACH2A		 Itacitinib IC_{50} <= 5 nM
			 MACCS155 A!CH2!A		
			 MACCS94 QN		
			 MACCS118 ACH2CH2A > 1		
Class 2	MACCS129(1), MACCS155(1), MACCS94(0), MACCS42(1)	147/46	 MACCS42 F		 Molecule 2960 IC_{50} = 0.07 nM
Class 3	MACCS129(1), MACCS155(0), MACCS70(1), MACCS138(0)	33/2	 MACCS129 ACH2AACH2A		 Molecule 628 IC_{50} = 2.2 nM
			 MACCS70 QNQ		
			 MACCS138 QCH2A > 1		
Class 4	MACCS129(0), MACCS52(1), MACCS111(1),	20/517	 MACCS52 NN		 Molecule 115 IC_{50} = 6833 nM
			 MACCS111 NACH2A		

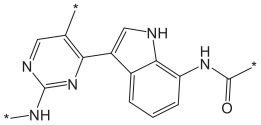
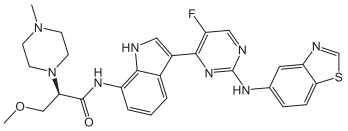
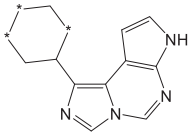
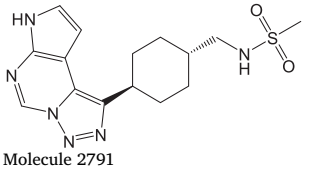
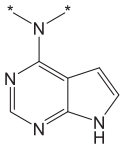
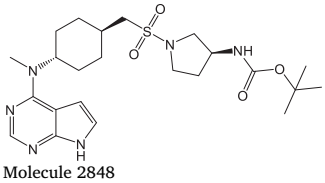
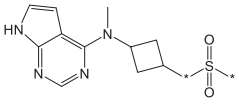
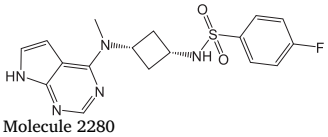
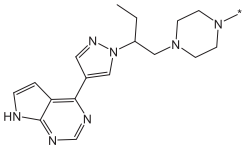
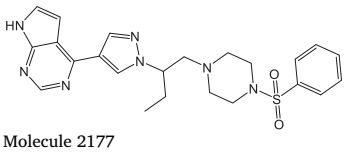
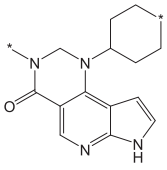
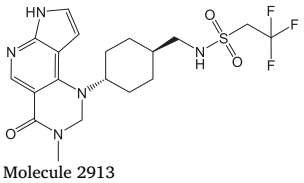
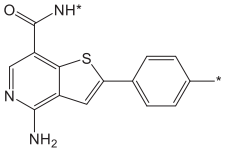
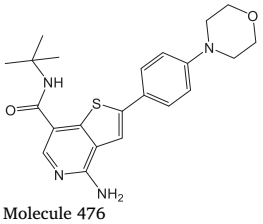
[a] (1) Represents that the molecule contained this MACCS key, (0) Represents that the molecule did not contain this MACCS key.
[b] The ratio of the number of highly active inhibitors and weakly active inhibitors in this class of inhibitors.
[c] In the MACCS key examples listed, the red is the matched structure, the bottom line was the SMARTS corresponding to the structure, and the meaning of the symbols involved:
A : Any valid periodic table element symbol; Q : Hetero atoms; any non-C or non-H atom; ! : Chain or non-ring bond; ! before a bond type specifies chain bond.
[d] The MACCS key pair corresponded to the partial substructure in the inhibitor, and the red was the matched structure.
[e] Representative molecules in this class of inhibitors, red was the matched substructure.

Table 6
Morgan fingerprint key and sub-structure analysis.

Groups	Morgan fingerprint Keys[a]	Subgroups	Highly/ Weakly[b]	Corresponding substructure[c]	Representative scaffold[c]	Typical compound[d]
Group 1	Morgan1772(0), Morgan2895(1), Morgan4078(1)		353/1	 Morgan 1772  Morgan 2895  Morgan 4078		 Itacitinib IC_{50} <= 5 nM
Group 2	Morgan1772(0), Morgan2895(0), Morgan1357(1), Morgan2722(1)	Subgroups 2A	358/1	 Morgan 1357  Morgan 2722		 Baricitinib IC_{50} = 4 nM
		Subgroups 2B	10/3			 Tofacitinib IC_{50} = 1.6 nM
		Subgroups 2C	6/0			 Molecule 1373 IC_{50} = 0.5 nM
		Subgroups 2D	72/3			 Molecule 2797 IC_{50} = 0.11 nM
		Subgroups 2E	13/0			 Molecule 2922 IC_{50} = 0.85 nM

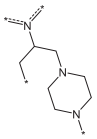
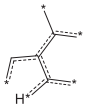
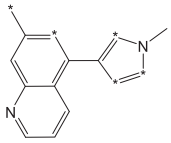
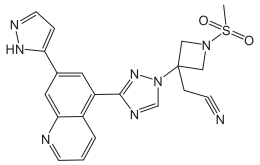
(*continued on next page*)

Table 6 (*continued*)

Groups	Morgan fingerprint Keys[a]	Subgroups	Highly/ Weakly[b]	Corresponding substructure[c]	Representative scaffold[c]	Typical compound[d]
Group 3	Morgan1772(0), Morgan2895(0), Morgan1357(1), Morgan2722(0)	Subgroups 3A	104/31			 Molecule 2960 IC_{50} = 0.07 nM
		Subgroups 3B	294/20			 Molecule 2791 IC_{50} = 0.16 nM
		Subgroups 3C	75/15			 Molecule 2848 IC_{50} = 1.44 nM
		Subgroups 3D	26/4			 Molecule 2280 IC_{50} = 3 nM
		Subgroups 3E	30/0			 Molecule 2177 IC_{50} < 20 nM
		Subgroups 3F	37/0			 Molecule 2913 IC_{50} = 0.42 nM
		Subgroups 3G	3/13			 Molecule 476 IC_{50} = 8820 nM

(*continued on next page*)

Table 6 (*continued*)

Groups	Morgan fingerprint Keys[a]	Subgroups	Highly/ Weakly[b]	Corresponding substructure[c]	Representative scaffold[c]	Typical compound[d]
Group 4	Morgan1772(1), Morgan3091(1), Morgan2962(0), Morgan2481(0)		12/415	 Morgan 3091  Morgan 2962  Morgan 2481		 Molecule 1209 IC_{50} = 8303 nM

[a] (1) Represents that the molecule contained the corresponding Morgan key structure; (0) Represents that the molecule did not contain the corresponding Morgan key structure.

[b] The ratio of the number of highly active inhibitors to weakly active inhibitors in the sub-category.

[c] *Represents any atom or group.

[d] The number behind 'Molecule' represents the ID of the molecule in our JAK1 inhibitors dataset. The JAK1 inhibitor dataset was displayed in JAK1_dataset.csv in the Supplementary materials.

parisons. When $p < 0.05$, the difference between two AUC values is thought to be significant at the 95% level. We calculated p-values across all models and found that model 3B had the largest number of significant differences from the other models. The results of multiple pairwise comparisons between Model 3B and other 19 models are shown in Table 4.

It can be seen from Table 4 that there are significant differences between Model 3B and Model 2C, Model 2A, Model 2B, Model 2D, Model 3C, Model 3A, Model 5C, Model 5A, Model 4A, Model 1A, Model 1C, Model 4C, Model 5D, Model 3D, and Model 3B is statistically superior to these models. Although there are differences with Model 1D, Model 4D, Model 1B, Model 5B and Model 4B, it is not statistically significant.

Analysis of MACCS fingerprints

Each MACCS key corresponds to a segment of SMARTS [126]. SMARTS was an abstract definition. It matches the corresponding structure through specific rules. A segment of SMARTS may match multiple different substructures, the same structure may also correspond to different SMARTS. Each letter in SMARTS has a different meaning. For example, 'A' represents any valid periodic table element symbol, 'Q' represents a hetero atom, then 'QNQ' means that there are two hetero atoms bond to the same nitrogen atom. Table 5 lists some examples of substructures corresponding to SMARTS.

We calculated a MACCS fingerprint with a length of 166 bits for each inhibitor. We deleted the bits with a variance less than the average variance, finally used a MACCS fingerprint with a length of 80 bits to build models. The top 80 MACCS fingerprints are shown in Table S1 in the Supplementary materials. In the analysis, we found that the combination of different important MACCS keys forms some important molecules with special scaffold features, which had an obvious activity possibility. The important MACCS keys were further divided into four classes which are summarized in Table 5.

In Class 1, all molecules contain MACCS129, MACCS155, MACCS94 and MACCS118. MACCS129 and MACCS118 correspond to a variety of cyclic structures; MACCS155 corresponds to some chain structures; MACCS94 corresponds to pyrazole, pyrrolotriazolopyrimidine, pyrazolopyrimidine and other structures with adjacent heteroatoms. There were 1261 inhibitors with the above characteristics, of which 1144 were highly active and 117 were weakly active. As shown in Table 5, Itacitinib hits all the above characteristics.

Class 2 contained MACCS129, MACCS155 and MACCS42 but did not contain MACCS94, which made them all contain fluorine atoms and cyclic structures but no adjacent non-carbon atoms. There were a total of 193 of these inhibitors, highly: weakly active inhibitors = 147: 46.

Class 3 contained MACCS129 and MACCS70, but did not contain MACCS155 and MACCS138. These inhibitors all contained six-membered aliphatic rings, but there were no heteroatoms in the aliphatic ring and no aliphatic chain structure. In addition, they contained three consecutive nitrogen atom structures, which made them all contain pyrrolotriazolopyrimidine. A total of 35 inhibitors meet the above characteristics, 33 of which were weakly active inhibitors.

There were 537 inhibitors in Class 4, and only 20 of them were highly active inhibitors, the others are weakly active inhibitors. These inhibitors all contained MACCS52, MACCS111 and MACCS116, which meant that they had a structure with two consecutive nitrogen atoms and did not have an alicyclic structure with more than five-element.

Analysis of Morgan fingerprints

In this study, we set the radius of the Morgan fingerprint to 4 and the length to 4096 bits, we screened the fingerprints to remove the bits whose variance was less than the average variance. Finally, 958 bits were reserved. According to the importance of the random forest model of all inhibitors, we listed the top 100 uniquely meaningful Morgan keys in Table S2 in the Supplementary materials.

Among these keys, the top ranking one contributes more to the model than the other ones in the behind. Subsequently, we analyzed some common scaffold structures with obvious regularities. As shown in Table 6, Group 1 (354 inhibitors) included Morgan2895 and Morgan4078 keys, yet excluded Morgan1772 keys. Among them, 353 compounds were highly active inhibitors and only one was weakly active. These inhibitors were analogs of Itacitinib and had an obvious common scaffold structure.

Group 2 (466 inhibitors) contained Morgan1357 and Morgan2722 keys, without Morgan1772 and Morgan2895 keys. These inhibitors

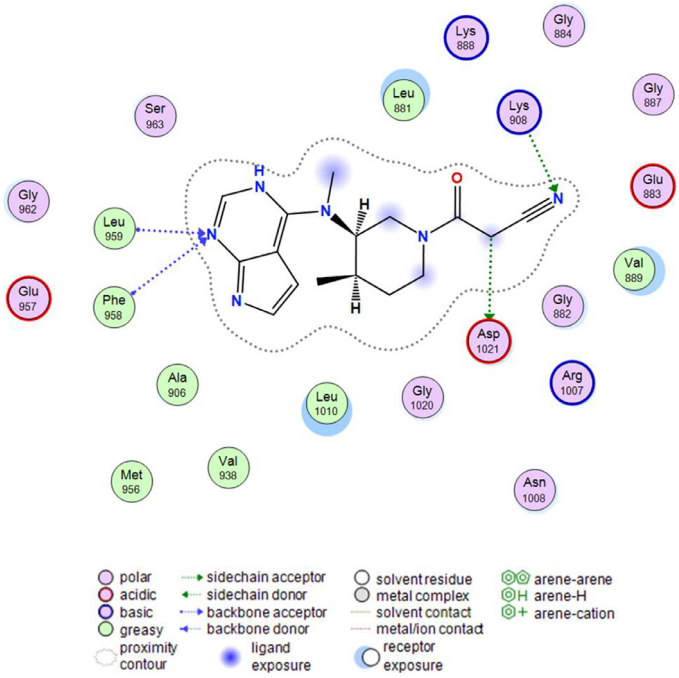


Figure 5. Interaction between Tofacitinib (CP-690, 550) and JAK1 (PDB ID: 3EYG).

would be divided into five subgroups. Subgroup 2A was composed of 359 Baricitinib analogs inhibitors, almost all inhibitors were highly active except one molecule with weak inhibition; There were 13 Tofacitinib analogs inhibitors in Subgroup 2B, of which 10 were highly active and three were weakly active; six highly active inhibitors formed Subgroup 2C; Subgroup 2D contained 72 highly active inhibitors and three weakly active inhibitors; 13 highly active inhibitors constituted Subgroup 2E.

A total of 652 inhibitors in Group 3 contained Morgan1357 keys and did not contain Morgan1772, Morgan2895, and Morgan2772 keys. They were divided into seven subgroups. Among them, Subgroup 3A contained 104 highly active inhibitors and 31 weakly active inhibitors. Subgroup 3B(294 highly active inhibitors and 20 weakly active inhibitors), Subgroup 3C(75 highly active inhibitors and 15 weakly active inhibitors) and Subgroup 3F(all of 37 inhibitors were highly active) had the same scaffold structure as Subgroup 2D, Subgroup 2B and Subgroup 3E in Group 2, but none of them contained a cyano structure. There were 30 inhibitors in Subgroup 3D, among them, 26 inhibitors were highly active and four inhibitors were weakly active. There were 16 inhibitors in Subgroup 3G, most of which were weakly active inhibitors. The high-weak activity ratios and corresponding sub-structures of all subgroups were shown in Table 6.

As shown in Table 6, Group 4 accommodated a total of 427 inhibitors. The inhibitors contained Morgan1772 and Morgan3091 keys but did not contain Morgan2962 and Morgan2481 keys. Among them, 415 inhibitors were weakly active molecules, the unfavorable substructures may reduce their activities.

Interaction between the inhibitor and JAK1 receptor

The interaction between Tofacitinib (CP-690, 550) and JAK1 (PDB ID: 3EYG) [127] was shown in Figure 5. It can be seen from Figure 5, Leu959, Phe958, Lys908 and Asp1021 are key residues of JAK1, since they develop direct interaction with Tofacitinib. The pyrrolopyrimidine group of the Tofacitinib generated the hydrogen bond with Leu959 and Phe958, which was consistent with the Morgan fingerprint analysis result, corresponding to Morgan1357 (Table 6); Tofacitinib is well accommodated in the JAK1 active site, change to this: located in and near the polar pocket containing Asp1021 and Lys90, Asp1021 and Lys90 form H-bonding interaction with carbon and nitrogen atoms of methyl cyanide group, respectively. This was consistent with the results of MACCS fingerprint analysis and Morgan fingerprint analysis, corresponding to MACCS155 and Morgan2722 (Table 6), indicating that our analysis of fingerprint descriptors was accurate and reliable.

Conclusions

In this study, several well-performed classification models for distinguishing highly/weakly active JAK1 inhibitors were developed based on a dataset of 2982 JAK1 inhibitors. We used MACCS and Morgan fingerprints to characterize molecules, then split the dataset into a training/test set by using SOM and random methods. We constructed classification models with SVM, DT, RF, XGBoost and DNN. The average ACC and MCC values of the traditional machine learning algorithm models reached 84.6% and 0.68 on the test set. Model 3B, built on Morgan fingerprints using the RF algorithm, is statistically significantly better than Model 3B and other 15 models, ACC of 93.6% and MCC of 0.87 were achieved on the test set. The classification models developed in the current work can be utilized to predict the bioactivities of JAK1 inhibitors.

Furthermore, We ranked the importance of the MACCS and Morgan fingerprints descriptors based on the RF models. We detected that the pyrazole group, the pyrrolotriazolopyrimidine group and the pyrazolopyrimidine group appeared frequently in highly active JAK1 inhibitors yet barely showed up in weak inhibitors. These observations may provide valuable clues for the design of potent JAK1 inhibitors.

Authors' contributions

ZWY implemented the method and evaluated the models, YJT, YK performed the analysis. AXY and YSZ provided the main idea of this work. All authors read and approved the final manuscript.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Acknowledgements

This work was supported by 'Chemical Grid Project' of Beijing University of Chemical Technology. We thank Molecular Networks GmbH, Erlangen, Germany, for providing the programs SONNIA.

Funding

YZ acknowledged the support from "the Fundamental Research Funds for the Central Universities, Beijing University of Chemical Technology (No: JD2103)".

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2022.100039.

References

[1] Kumar N, Kuang L, Villa R, Kumar P, Mishra J. Mediators Inflamm 2021;2021:6618924.
[2] Daoud S, Taha MO. J Mol Graph Model 2020;99:107615.
[3] Jisha RS, Aswathy L, Masand VH, Gajbhiye JM, Shibi IG. In Silico Pharmacol 2017;5:9.
[4] Keretsu S, Bhujbal SP, Cho SJ. J Biomol Struct Dyn 2021;39:753–65.
[5] Bajusz D, Ferenczy GG, Keseru GM. J Mol Graph Model 2016;70:275–83.
[6] Jasuja H, Chadha N, Kaur M, Silakari O. SAR QSAR Environ Res 2014;25:617–36.
[7] Itteboina R, Ballu S, Sivan SK, Manga V. J Recept Signal Transduct Res 2017;37:453–69.

[8] Spinelli FR, Colbert RA, Gadina M. Rheumatology (Oxford) 2021;60:ii3–ii10.
[9] Mascarenhas J, Hoffman R. Clin Cancer Res 2012;18:3008–14.
[10] Cherington CC, Gowin KL, Mesa RA. Expert Opinion on Orphan Drugs 2015;3:1085–96.
[11] Ali H, Salhotra A, Modi B, Nakamura R. Expert Rev Clin Immunol 2020;16:347–59.
[12] Gong X, Chen X, Kuligowski ME, Liu X, Liu X, Cimino E, McGee R, Yeleswaram S. Am J Clin Dermatol 2021;22:555–66.
[13] Yamaoka K. Expert Rev Clin Immunol 2019;15:577–88.
[14] Abdulrahim H, Sharlala H, Adebajo AO. Expert Opin Pharmacother 2019;20:1953–60.
[15] Colombel JF, Osterman MT, Thorpe AJ, Salese L, Nduaka CI, Zhang H, Lawendy N, Friedman GS, Quirk D, Su C, Reinisch W. Clin Gastroenterol Hepatol 2020. doi:10.1016/j.cgh.2020.10.004.
[16] Deodhar A, Sliwinska-Stanczyk P, Xu H, Baraliakos X, Gensler LS, Fleishaker D, Wang L, Wu J, Menon S, Wang C, Dina O, Fallon L, Kanik KS, van der Heijde D. Ann Rheum Dis 2021;80:1004–13.
[17] Mogul A, Corsi K, McAuliffe L. Ann Pharmacother 2019;53:947–53.
[18] Markham A. Drugs 2017;77:697–704.
[19] Duggan S, Keam SJ. Drugs 2019;79:1819–28.
[20] Tanaka Y. Mod Rheumatol 2020;30:779–87.
[21] Dhillon S, Keam SJ. Drugs 2020;80:1987–97.
[22] Peyrin-Biroulet L, Loftus Jr EV, Hibi T, Birchwood C, Yun C, Zhao S, Liu X, Rogler G, Danese S, Colombel JF, Feagan B. Journal of Crohn's and Colitis 2021;15:S395–6.
[23] Schroeder MA, Khoury HJ, Jagasia M, Ali H, Schiller GJ, Staser K, Choi J, Gehrs L, Arbushites MC, Yan Y, Langmuir P, Srinivas N, Pratta M, Perales MA, Chen YB, Meyers G, DiPersio JF. Blood Adv 2020;4:1656–69.
[24] Svoboda J, Barta S, Nasta S, Landsburg D, Gerson J, Ruella M, Waite T, King C, Emanuel SA, Ballard H, Schuster S. Hematological Oncology 2019;37:573–4.
[25] Forman SB, Pariser DM, Poulin Y, Vincent MS, Gilbert SA, Kieras EM, Qiu R, Yu D, Papacharalambous J, Tehlirian C, Peeva E. J Invest Dermatol 2020;140:2359–70.
[26] Kubo S, Nakayamada S, Tanaka Y. Expert Rev Clin Immunol 2019;15:693–700.
[27] You H, Xu D, Zhao J, Li J, Wang Q, Tian X, Li M, Zeng X. Clin Rev Allergy Immunol 2020;59:334–51.
[28] Crowley EL, Nezamololama N, Papp K, Gooderham MJ. Expert Rev Clin Immunol 2020;16:955–62.
[29] Stumpfe D, Bajorath J. MedChemComm 2016;7:1045–55.
[30] Muratov EN, Bajorath J, Sheridan RP, Tetko IV, Filimonov D, Poroikov V, Oprea TI, Baskin II, Varnek A, Roitberg A, Isayev O, Curtarolo S, Fourches D, Cohen Y, Aspuru-Guzik A, Winkler DA, Agrafiotis D, Cherkasov A, Tropsha A. Chem Soc Rev 2020;49:3525–64.
[31] Fratello M, Tagliaferri R. Decision Trees and Random Forests. Oxford: Academic Press; 2019.
[32] Wassan JT, Wang H, Zheng H. Machine Learning in Bioinformatics. Oxford: Academic Press; 2019.
[33] Qi J, Liang X, Xu R. Comput Intell Neurosci 2018;2018:1018789.
[34] Chen T, Guestrin C. presented in part at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. California, USA: San Francisco; 2016.
[35] Grossi E, Buscema M. Eur J Gastroenterol Hepatol 2007;19:1046–54.
[36] Sarithamol S, Pushpa VL, Divya V, Manoj KB. Chem Biol Drug Des 2020;95:503–19.
[37] Itteboina R, Ballu S, Sivan SK, Manga V. Comput Biol Chem 2016;64:33–46.
[38] Guarascio M, Manco G, Ritacco E. Deep Learning. Oxford: Academic Press; 2019.
[39] Sliwoski G, Kothiwale S, Meiler J, Lowe EW. Pharmacological Reviews 2014;66:334–95.
[40] Rogers D, Hahn M. J Chem Inf Model 2010;50:742–54.
[41] Davies M, Nowotka M, Papadatos G, Dedman N, Gaulton A, Atkinson F, Bellis L, Overington JP. Nucleic Acids Research 2015;43:W612–20.
[42] Reaxys Database, https://www.reaxys.com, (accessed 5.10, 2022).
[43] Gabrielson SW. Journal of the Medical Library Association 2018:106.
[44] Yin Y, Chen CJ, Yu RN, Shu L, Wang ZJ, Zhang TT, Zhang DY. Bioorg Chem 2020;98:103720.
[45] Su Q, Banks E, Bebernitz G, Bell K, Borenstein CF, Chen H, Chuaqui CE, Deng N, Ferguson AD, Kawatkar S, Grimster NP, Ruston L, Lyne PD, Read JA, Peng X, Pei X, Fawell S, Tang Z, Throner S, Vasbinder MM, Wang H, Winter-Holt J, Woessner R, Wu A, Yang W, Zinda M, Kettle JG. J Med Chem 2020;63:4517–27.
[46] Yin Y, Chen CJ, Yu RN, Shu L, Zhang TT, Zhang DY. Bioorg Med Chem 2019;27:1562–76.
[47] Yu RN, Chen CJ, Shu L, Yin Y, Wang ZJ, Zhang TT, Zhang DY. Bioorg Med Chem 2019;27:1646–57.
[48] Liang X, Zang J, Li X, Tang S, Huang M, Geng M, Chou CJ, Li C, Cao Y, Xu W, Liu H, Zhang Y. J Med Chem 2019;62:3898–923.
[49] Calbet M, Ramis I, Calama E, Carreno C, Paris S, Maldonado M, Orellana A, Calaf E, Pauta M, De Alba J, Bach J, Miralpeix M. J Pharmacol Exp Ther 2019;370:137–47.
[50] Bach J, Eastwood P, Gonzalez J, Gomez E, Alonso JA, Fonquerna S, Lozoya E, Orellana A, Maldonado M, Calaf E, Alberti J, Perez J, Andres A, Prats N, Carreno C, Calama E, De Alba J, Calbet M, Miralpeix M, Ramis I. J Med Chem 2019;62:9045–60.
[51] Wrobleski ST, Moslin R, Lin S, Zhang Y, Spergel S, Kempson J, Tokarski JS, Strnad J, Zupa-Fernandez A, Cheng L, Shuster D, Gillooly K, Yang X, Heimrich E, McIntyre KW, Chaudhry C, Khan J, Ruzanov M, Tredup J, Mulligan D, Xie D, Sun H, Huang C, D'Arienzo C, Aranibar N, Chiney M, Chimalakonda A, Pitts WJ, Lombardo L, Carter PH, Burke JR, Weinstein DS. J Med Chem 2019;62:8973–95.
[52] Fensome A, Ambler CM, Arnold E, Banker ME, Brown MF, Chrencik J, Clark JD, Dowty ME, Efremov IV, Flick A, Gerstenberger BS, Gopalsamy A, Hayward MM, Hegen M, Hollingshead BD, Jussif J, Knafels JD, Limburg DC, Lin D, Lin TH, Pierce BS, Saiah E, Sharma R, Symanowicz PT, Telliez JB, Trujillo JI, Vajdos FF, Vincent F, Wan ZK, Xing L, Yang X, Yang X, Zhang L. J Med Chem 2018;61:8597–612.
[53] Nakajima Y, Aoyama N, Takahashi F, Sasaki H, Hatanaka K, Moritomo A, Inami M, Ito M, Nakamura K, Nakamori F, Inoue T, Shirakami S. Bioorg Med Chem 2016;24:4711–22.
[54] Nakajima Y, Tojo T, Morita M, Hatanaka K, Shirakami S, Tanaka A, Sasaki H, Nakai K, Mukoyoshi K, Hamaguchi H, Takahashi F, Moritomo A, Higashi Y, Inoue T. Chem Pharm Bull (Tokyo) 2015;63:341–53.
[55] Yeleswaram K, Smith P, Hollis F. Gregory,*US Pat.* 2020(A1):WO2020132210.
[56] Brown Matthew F, Fensome A, Gerstenberger Brian S, Hayward Matthew M, Owen Dafydd R, Vajdos F, Xing Li H. US Pat. 2018(A1):WO2019034973.
[57] Van Der Plas S, Mammoliti O, Martina S, Claes P, Coti G, Annoot D, LÓPez Ramos M, Galien R, Amantini D, Brys R. BE Pat. 2019(A1):WO2019076716.
[58] Nakamura M, Yamanaka H, Mitsuya M, Harada T. JP Pat. 2018(A1):EP3330271 EP3330271 (A4).
[59] Nilsson K, ÅStrand A, Berggren A, Johansson J, LepistÖ M, Kawatkar S, Su Q, Kettle J. SE Pat. 2018(A1):WO2018134213.
[60] Xi N, Li X, Li M, Zhang T, Hu H, Wu Y. CN Pat. 2018(A1):WO2018169700.
[61] Menet Christel Jeanne M, Mammoliti O, Quinton E, Joannesse Caroline Martine A-M, De Blieck A, Blanc J. BE Pat. 2015(A1):WO2017012647.
[62] Fensome A, Gopalsamy A, Gerstenberger Brian S, Efremov Ivan V, Wan Z-K, Pierce B, Telliez J-B, Trujillo John I, Zhang L, Xing L. US Pat. 2015(B2):US9663526 US2016052930 (A1).
[63] Menet Christel Jeanne M, Mammoliti O, Blanc J, Orsulic M, Roscic M. BE Pat. 2015(B2):US9440929 US2015203455 (A1).
[64] Nakamura M, Yamanaka H, Shibata K, Mitsuya M, Harada T. JP Pat. 2015(A1):WO2015119126.
[65] Li YUNL, Zhuo J, Qian D-Q, Mei S, Cao G, Pan Y, Li Q, Jia Z. US Pat. 2014(A1):WO2014186706.
[66] Menet C, Schmitt B, Geney R, Doyle K, Peach J, Palmer N, Jones G, Hardy D, Duffy J. BE Pat. 2013(A1):WO2013117649.
[67] Rodgers James D, Shepard S, Zhu W, Shao L, Glenn J. US Pat. 2013(A1):WO2013173720.
[68] Rodgers James D, Zhu W, Glenn J. US Pat. 2012(A1):WO2012068440.
[69] Rodgers James D, Zhu W, Shao L, Glenn J. WO2012068450 (A1). US Pat. 2012.
[70] Paul REastwood, Rodriguez JGonzalez, Castillo EGomez, Bach Tana J. ES Pat. 2012(A1):WO2012160030.
[71] James DRodgers, Li YUNL, Shepard S, Wang H. US Pat. 2011(A1):WO2011028685.
[72] Paul REastwood, Gonzalez Rodriguez J, Bach Tana J, Lluis MPages Santacana, Taltavull Moll J, Juan FCaturla Javaloyes, Victor GMatassa. ES Pat. 2010(A1):WO2011076419.
[73] Bach Tana J, Pages Santacana Lluis M, Taltavull Moll J, Eastwood Paul R, Gonzalez Rodrigues J, Giulio Matassa V. ES Pat. 2011(A1):WO2011101161.
[74] Boys Mark L, Burgess Laurence E, Groneberg Robert D, Harvey Darren M, Huang L, Kercher T, Kraser Christopher F, Laird E, Tarlton E, Zhao Q. US Pat. 2011(A1):WO2011130146.
[75] Paul REastwood, Rodriguez JGonzalez, Castillo EGomez, Bach Tana J. ES Pat. 2011(A1):WO2011157397.
[76] Menet Christel Jeanne M, Jouannigot N, Blanc J, Van Rompaey Luc Juliaan C, Fletcher Stephen R. BE Pat. 2009(A1):WO2010010186.
[77] Combs Andrew P, Sparks Richard B, Yue Eddy W, Feng H, Bower Michael J, Zhu W. US Pat. 2009(B2):US8871753 US2009286778 (A1).
[78] Rodgers James D, Shepard S, Li YUNL, Zhou J, Liu P, Meloni D, Xia M. US Pat. 2009(A1):WO2009114512.
[79] Bourke David G, Bu X, Burns Christopher J, Cuzzupe Anthony N, Feutrill John T, Nero Tracy L, Blannin Beata M, Zeng J, Gaynor Shaun P. AU Pat. 2008(A1):WO2008092199.
[80] Malerich JP, Lam JS, Hart B, Fine RM, Klebansky B, Tanga MJ, D'Andrea A. Bioorg Med Chem Lett 2010;20:7454–7.
[81] Flanagan ME, Blumenkopf TA, Brissette WH, Brown MF, Casavant JM, Shang-Poa C, Doty JL, Elliott EA, Fisher MB, Hines M, Kent C, Kudlacz EM, Lillie BM, Magnuson KS, McCurdy SP, Munchhof MJ, Perry BD, Sawyer PS, Strelevitz TJ, Subramanyam C, Sun J, Whipple DA, Changelian PS. J Med Chem 2010;53:8468–84.
[82] Schenkel LB, Huang X, Cheng A, Deak HL, Doherty E, Emkey R, Gu Y, Gunaydin H, Kim JL, Lee J, Loberg R, Olivieri P, Pistillo J, Tang J, Wan Q, Wang HL, Wang SW, Wells MC, Wu B, Yu V, Liu L, Geuns-Meyer S. J Med Chem 2011;54:8440–50.
[83] Dugan BJ, Gingrich DE, Mesaros EF, Milkiewicz KL, Curry MA, Zulli AL, Dobrzanski P, Serdikoff C, Jan M, Angeles TS, Albom MS, Mason JL, Aimone LD, Meyer SL, Huang Z, Wells-Knecht KJ, Ator MA, Ruggeri BA, Dorsey BD. J Med Chem 2012;55:5243–54.
[84] Yang J, Wang LJ, Liu JJ, Zhong L, Zheng RL, Xu Y, Ji P, Zhang CH, Wang WJ, Lin XD, Li LL, Wei YQ, Yang SY. J Med Chem 2012;55:10685–99.
[85] Yamagishi H, Shirakami S, Nakajima Y, Tanaka A, Takahashi F, Hamaguchi H, Hatanaka K, Moritomo A, Inami M, Higashi Y, Inoue T. Bioorg Med Chem 2015;23:4846–59.
[86] Nakajima Y, Inoue T, Nakai K, Mukoyoshi K, Hamaguchi H, Hatanaka K, Sasaki H, Tanaka A, Takahashi F, Kunikawa S, Usuda H, Moritomo A, Higashi Y, Inami M, Shirakami S. Bioorg Med Chem 2015;23:4871–83.
[87] Vasbinder MM, Alimzhanov M, Augustin M, Bebernitz G, Bell K, Chuaqui C, Deegan T, Ferguson AD, Goodwin K, Huszar D, Kawatkar A, Kawatkar S, Read J, Shi J, Steinbacher S, Steuber H, Su Q, Toader D, Wang H, Woessner R, Wu A, Ye M, Zinda M. Bioorg Med Chem Lett 2016;26:60–7.
[88] Reddy MV, Akula B, Jatiani S, Vasquez-Del Carpio R, Billa VK, Mallireddigari MR, Cosenza SC, Venkata Subbaiah DR, Bharathi EV, Pallela VR, Ramkumar P, Jain R, Aggarwal AK, Reddy EP. Bioorg Med Chem 2016;24:521–44.

[89] Katoh T, Takai T, Yukawa T, Tsukamoto T, Watanabe E, Mototani H, Arita T, Hayashi H, Nakagawa H, Klein MG, Zou H, Sang BC, Snell G, Nakada Y. Bioorg Med Chem 2016;24:2466–75.
[90] Liang X, Huang Y, Zang J, Gao Q, Wang B, Xu W, Zhang Y. Bioorg Med Chem 2016;24:2660–72.
[91] Vazquez ML, Kaila N, Strohbach JW, Trzupek JD, Brown MF, Flanagan ME, Mitton-Fry MJ, Johnson TA, TenBrink RE, Arnold EP, Basak A, Heasley SE, Kwon S, Langille J, Parikh MD, Griffin SH, Casavant JM, Duclos BA, Fenwick AE, Harris TM, Han S, Caspers N, Dowty ME, Yang X, Banker ME, Hegen M, Symanowicz PT, Li L, Wang L, Lin TH, Jussif J, Clark JD, Telliez JB, Robinson RP, Unwalla R. J Med Chem 2018;61:1130–52.
[92] Yao L, Mustafa N, Tan EC, Poulsen A, Singh P, Duong-Thi MD, Lee JXT, Ramanujulu PM, Chng WJ, Yen JJY, Ohlson S, Dymock BW. J Med Chem 2017;60:8336–57.
[93] Grimster NP, Anderson E, Alimzhanov M, Bebernitz G, Bell K, Chuaqui C, Deegan T, Ferguson AD, Gero T, Harsch A, Huszar D, Kawatkar A, Kettle JG, Lyne P, Read JA, Rivard Costa C, Ruston L, Schroeder P, Shi J, Su Q, Throner S, Toader D, Vasbinder M, Woessner R, Wang H, Wu A, Ye M, Zheng W, Zinda M. J Med Chem 2018;61:5235–44.
[94] Yin Y, Chen CJ, Yu RN, Wang ZJ, Zhang TT, Zhang DY. Bioorg Med Chem 2018;26:4774–86.
[95] Hamaguchi H, Amano Y, Moritomo A, Shirakami S, Nakajima Y, Nakai K, Nomura N, Ito M, Higashi Y, Inoue T. Bioorg Med Chem 2018;26:4971–83.
[96] Wagner J, von Matt P, Sedrani R, Albert R, Cooke N, Ehrhardt C, Geiser M, Rummel G, Stark W, Strauss A, Cowan-Jacob SW, Beerli C, Weckbecker G, Evenou JP, Zenke G, Cottens S. J Med Chem 2009;52:6193–6.
[97] Huang T, Xue CHUB, Wang A, Kong Ling Q, Ye Hai F, Yao W, Rodgers James D, Shepard S, Wang H, Shao L, Li HUIY, Li Q. US Pat. 2011(B2):US8765734 US2011224190 (A1).
[98] Menet Christel Jeanne M, Smits Koen K. US Pat. 2011(B2):US8563545 US2012142678 (A1).
[99] Menet Christel Jeanne M, Blanc J. US Pat. 2010(B2):US8796457 US2010331359 (A1).
[100] Hayashi K, Watanabe T, Toyama K, Kamon J, Minami M, Uni M, Nasu M. JP Pat. 2012(B2):US9216999 US2014200344 (A1).
[101] Su WEIG, Deng W, Li J, Ji J. CN Pat. 2013(B2):US9346810 US2013210831 (A1).
[102] Ma H, Filip Sorin V. US Pat. 2010(B2):US8592415 US2011071149 (A1).
[103] Promo Michele A, Xie J, Acker Brad A, Hartmann Susan J, Wolfson Sergey G, Huang H-C, Jacobsen Eric J. US Pat. 2011(B2):US8633206 US2011136765 (A1).
[104] Takahashi K, Watanabe T, Hayashi K, Kurihara K, Nakamura T, Yamamoto A, Nishimura T, Kamiyama T, Hidaka Y. JP Pat. 2015(B2):US9475813 US2015368245 (A1).
[105] Combs Andrew P, Sparks Richard B, Yue Eddy Wai T, Feng H, Bower Michael J, Zhu W. US Pat. 2010(B2):US8765727 US2010190804 (A1); US2011251215 (A9).
[106] Allen S, Andrews Steven W, Condroski Kevin R, Haas J, Huang L, Jiang Y, Kercher T, Seo J. US Pat. 2010(B2):US8791123 US2012108568 (A1).
[107] Wang W, Diao Y, Li W, Luo Y, Yang T, Zhao Y, Qi T, Xu F, Ma X, Ge H, Liang Y, Zhao Z, Liang X, Wang R, Zhu L, Li H, Xu Y. Bioorg Med Chem Lett 2019;29:1507–13.
[108] Qi C, Tsui H, Zeng Q, Yang Z, Zhang X. CN Pat. 2020(A1):WO2020211839.
[109] Gade DR, Kunala P, Raavi D, Reddy PK, Prasad RV. J Recept Signal Transduct Res 2015;35:189–201.
[110] Dhanachandra Singh K, Karthikeyan M, Kirubakaran P, Nagamani S. J Mol Graph Model 2011;30:186–97.
[111] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Müller A, Nothman J, Louppe G. Journal of Machine Learning Research 2011;12:2825–30.
[112] Kohonen T. The Basic SOM. Berlin, Heidelberg: Springer Berlin Heidelberg; 2001.
[113] SONNIA, https://www.mn-am.com/products/sonnia, (accessed 5.10, 2022).
[114] RDKit 2020.03; Open-Source Cheminformatics Software, CA, USA: San Francisco; 2018. http://www.rdkit.org, (accessed 5.10, 2022).
[115] Wang H, Qin Z, Yan A. Mol Divers 2021;25:1597–616.
[116] Bemis GW, Murcko MA. J Med Chem 1996;39:2887–93.
[117] Willett P. Mol Inform 2014;33:403–13.
[118] Breiman L. Machine Learning 1996;24:123–40.
[119] Chauhan VK, Dahiya K, Sharma A. Artificial Intelligence Review 2018;52:803–55.
[120] Wu X, Kumar V, Quinlan JRoss, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou Z-H, Steinbach M, Hand DJ, Steinberg D. Knowledge and Information Systems 2007;14:1–37.
[121] Pytorch, https://github.com/pytorch/pytorch, (accessed 5.10, 2022).
[122] Matplotlib, https://matplotlib.org/, (accessed 5.10, 2022).
[123] DeLong ER, DeLong DM, Clarke-Pearson DL. Biometrics 1988:44.
[124] Nicholls A. J Comput Aided Mol Des 2016;30:103–26.
[125] Benjamini Y, Hochberg Y. Journal of the Royal Statistical Society: Series B (Methodological) 1995;57:289–300.
[126] Daylight Theory Manual (2006) SMARTS - A Language for Describing Molecular Patterns, https://www.ics.uci.edu/~dock/manuals/DaylightTheoryManual/theory.smarts.html, (accessed 5.10, 2022).
[127] Williams NK, Bamert RS, Patel O, Wang C, Walden PM, Wilks AF, Fantino E, Rossjohn J, Lucet IS. J Mol Biol 2009;387:219–32.