Supplement

To the paper of Hongbin Yang, Chaofeng Lou, Lixia Sun, Jie Li, Yingchun Cai, Zhuang Wang, Weihua Li, Guixia Liu and Yun Tang “admetSAR 2.0: web-service for prediction and optimization of chemical ADMET properties”

Contents

Description of preparation and features of molecules 1

Description of machine learning methods employed in admetSAR 1

Description of applicability domain 4

Description of ADMETopt 5

References 5

**Table 1.** Number of training molecules and the sources of each model 7

**Table 2.** The performance of the models in admetSAR 9

**Table 3.** The performance of the multi-classification models in admetSAR 10

**Table 4.** The performance of regression models in admetSAR 11

**Table 5.** The performance of multi-label models in admetSAR 12

**Figure S1.** Scheme of admetSAR 2.0 13

# Description of preparation and features of molecules

All datasets were processed with the following steps: firstly, molecular structures were checked for rationality and invalid ones were removed; secondly, all structures were desalted with inorganic counter-ions removed, and for mixtures or complexes, only the considered key component was remained; thirdly, inorganics were removed by checking the presence of carbon atoms and organic mental chemicals were also deleted; next, all tautomers were converted to correspondingly unique representations and then the stereochemistry information was removed to avoid duplicates as only the 2D molecular representations were used in our study; finally, all the duplicates were detected by using canonical smiles as their unique identifiers, and then the activity data of these duplicates were analyzed with ambiguous entries removed.

In order to construct predictive models, the molecules should be described by numeric features. In total 6 types of molecular fingerprints were calculated, including four topological fingerprints, i.e. RDKit fingerprint, Morgan fingerprint, Atom Pairs and Torsions fingerprint, and two predefined fingerprint, i.e. MACCS fingerprint and SubFP fingerprint.

The RDKit fingerprint identifies and hashes topological paths in the molecule and then uses them to set bits in a fingerprint of user-specified lengths. In this webserver, we used the default parameters. The path size is between 1-7. The number of bits set per hash is 2, and the fingerprint is folded down to 2048 bits.

The Atom Pairs fingerprint can be regarded as a huge size of vector in which each bit indicates the occurrence of a particular atom pair within a specific distance (Carhart, et al., 1985). The huge vectors are then folded into a much smaller size. In this webserver, the fingerprint size is 2048. Similarly, Torsion fingerprint is constructed by all kinds of four consecutively bonded non-hydrogen atoms (Nilakantan, et al., 1987). The fingerprint size is also set to 2048 bits.

The Morgan fingerprint is a kind of circular fingerprint that consider all possible fragment under a certain radius (Rogers and Hahn, 2010). In this webserver, the radius is set to 2 and the fingerprints were finally folded into 2048 bits.

The MACCS fingerprint is predefined by MDL containing 166 patterns represented as SMARTS (Durant, et al., 2002). Similarly, SubFP is another predefined substructure set containing 307 bits and is implemented by Open Babel (O'Boyle, et al., 2011). All the fingerprints except SubFP are calculated by RDKit (Landrum, 2017).

# Description of machine learning methods employed in admetSAR

## Single-label model

The models in admetSAR were mainly built by support vector machine (SVM), Random forest (RF), and k-nearest neighbor (kNN) that are widely used in classification and regression models. SVM, also known as support vector classifier (SVC) or support vector regression (SVR) in particular tasks, is to construct a hyperplane in a high dimensional space with the largest distance to the nearest training data points (Cortes and Vapnik, 1995). RF are derived from decision tree (Breiman, 2001), which can be viewed as bagging many decision trees that use a random subset of features and combine them via a voting system. kNN is one of the simplest algorithms (Cover and Hart, 1967), of which the creed is that compounds with similar structures have similar biological properties. In kNN, a sample is classified by the votes of the categories of its neighbors. The three algorithms were implemented via scikit-learn (Pedregosa, et al., 2011) in admetSAR. The models were evaluated by sensitivity (SE), specificity (SP), accuracy (ACC), and the area under the curve (AUC). SE, SP, and ACC can be calculated through the count of true positive (TP), false positive (FP), true negative (TN), and false negative (FN). AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

(1)

(2)

(3)

(4)

## Multi-label model

Unlike classification or regression in single-label models, multi-label classification (MLC) is a data mining approach in which each data instance can be assigned to multiple categories at once (Gibaja and Ventura, 2015; Tsoumakas, et al., 2010; Zhang and Zhou, 2014). There are three major approaches for multi-label learning: data transformation, method adaptation and ensembles of classifiers. Data transformation, including Binary Relevance (BR) (Godbole and Sarawagi, 2004), classifier chains (CC) (Read, et al., 2011) and Label Powerset (LP) (Boutell, et al., 2004), is to transform original multi-label dataset (MLD) to a set of binary datasets (BIDs) or one multi-class dataset (MCD) first, and then process them with traditional classification algorithms (Barot and Panchal, 2014). In this study, we used BR to build multi-label models for endocrine disruption chemicals and evaluated the models via the following metrics (Yang, et al., 2018).

(5)

(6)

(7)

(8)

(9)

where Yi represents the real label-set of the ith instance, and Zi the predicted one. n is the number of instances and k is the number of labels.

## Deep learning model

Graph convolutional neural network was used for building regression models. The graph convolution layer extends standard two-dimensional convolutions upon images to arbitrary graphs (Altae-Tran, et al., 2017; Duvenaud, et al., 2015). The molecules are viewed as undirected graphs in which atoms and bonds are viewed as nodes and edges. DeepChem (<https://deepchem.io>) is an open source python library devoted to providing a high quality toolchain to facilitate the use of DL in drug discovery and other fields. We implemented the deep learn algorithm via the “graph convolution” model defined in DeepChem. The models were evaluated by R2 and RMSE defined as following.

(10)

(11)

where xi is the experimental value, yi is the predicted value, are their corresponding means and N is the number of samples.

## Synthetic Minority Oversampling Technique (SMOTE)

This method is based on analyzing the similarity of minority class in feature space and synthesizing new fake minority data into the original set (Barua, et al., 2011). It can be defined as follows:

1. For each point p in S, compute its k nearest neighbors in S. The k variables have been optimized by using Python scripts which changing values from 1 to 10 interactively;

2. Randomly choose r ≤ k of the neighbors;

3. Choose a random point along the lines joining p and each of the r selected neighbors;

4. Add these synthetic points to the dataset S.

## Edited Nearest Neighbor (ENN) with SMOTE

This is an ensemble method to synthesize minority samples (Yang and Gao, 2013). Firstly, it adopts SMOTE algorithm to generate the balanced dataset T. Secondly, to decrease the potential of over-fitting, it uses k-Nearest Neighbor method to predict each sample in T set. If the prediction is not same as the true class, then delete the synthetic sample.

# Description of applicability domain

In admetSAR (version2), six physicochemical or topological properties are used to defined the applicability domain, which are molecular weight, alogP, number of atoms, number of rings, H-bond acceptors, and H-bond donors. We analyzed the distribution of these properties in all the training sets of the predictive models. The definition is as following:

Compounds with the molecular weight higher than 99% or lower than 99% of the training set will be regarded as warning, and if it is higher than the maximum of the training set, it will be tagged as out domain. Similarly, compounds with the AlogP higher than 99% or lower than 99% of the training set will be regarded as warning, and if it is higher than the maximum or lower than the minimum of the training set, it will be tagged as out domain. For the other four count properties, since their lower bound is 0 and is reasonable, we only tagged the compounds out of upper hound.

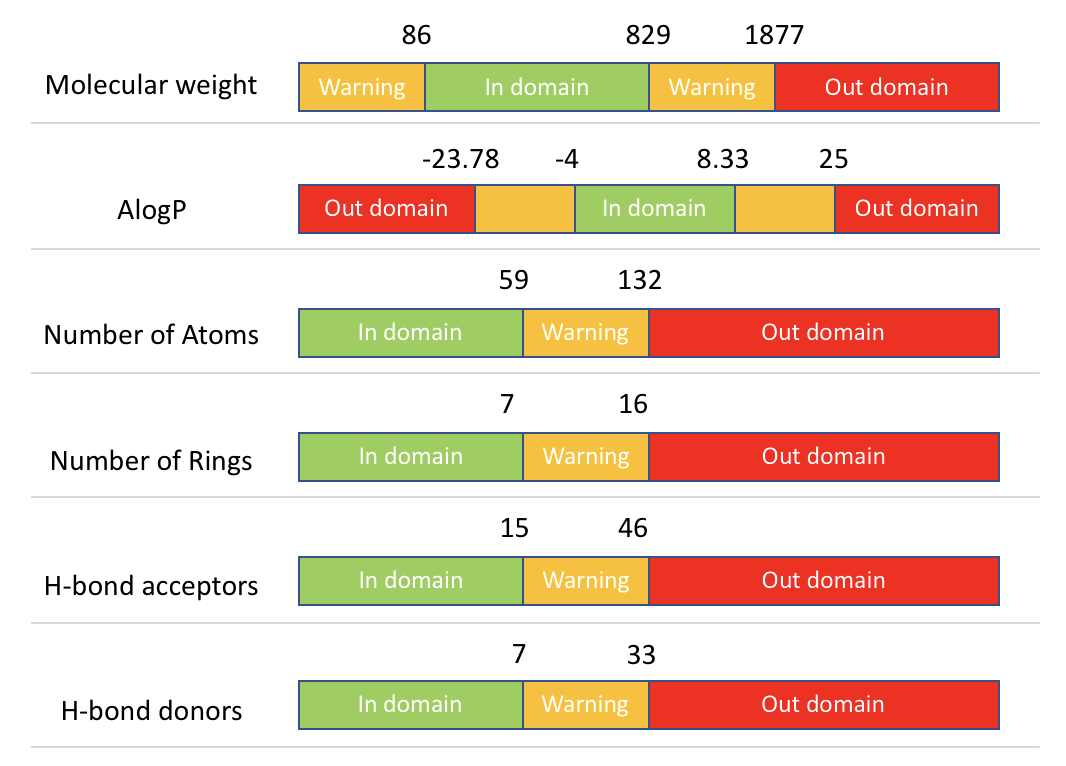


Figure 1. The definition of applicability domain.

# Description of ADMETopt

The ADMETopt is a module that can be used to optimize lead compounds using scaffold hopping and ADMET (Absorption, distribution, metabolism, excretion, and toxicity) screening.

## Definition of scaffold

The scaffold is defined as the following ring system, a single ring or a collection of fused rings or spiro rings, including exocyclic terminal bonds. The number of atoms in each smallest set of smallest rings must between 4 and 7. One compound may contain more than one scaffold and each time one scaffold can be selected to be replace by similar scaffolds.

## Scaffold hopping

Scaffold hopping is to replace the query scaffold of the query compound with by other similar ones. The similarity was calculated by Tanimoto similarity coefficient between the fingerprints of the two scaffolds. The scaffold fingerprint contains 15 statistics properties, 1 binary property and 4 kinds of vector properties, which are the number of bridges, spiro atoms, aromatic rings, rings, exocyclic bonds, carbon atoms, nitrogen atoms, oxygen atoms, sulfur atoms, phosphor atoms, sp3 hybridized carbons, sp3 hybridized carbons that are stereo centers, sp3 hybridized carbons with exit-vectors that are stereo centers, sp3 hybridized carbons without exit-vectors that are stereo centers, diversity points. The details of these descriptors can be found in this paper (Rabal, et al., 2015).

## ADMET properties as constrains

In total 15 properties can be set as constrains to optimize the query compound, including 7 physicochemical properties, i.e. molecular weight, partition coefficient, number of H-bond acceptors and donors, number of rotatable bonds, and number of halogens; and 8 machine learning based binary ADMET properties, i.e. blood brain barrier penetration, p-glycoprotein inhibitor, carcinogenicity, Ames mutagenicity, acute oral toxicity, hERG inhibitors, CYP450 inhibitory promiscuity, and human intestinal absorption.

# References

Altae-Tran, H.*, et al.* (2017) Low Data Drug Discovery with One-Shot Learning. *ACS Cent Sci*, 3(4):283-293.

Barot, P. and Panchal, M. (2014) Review on Various Problem Transformation Methods for Classifying Multi-Label Data. *International Journal of Data Mining & Emerging Technologies*:45-52.

Barua, S., Islam, M.M. and Murase, K. (2011) A Novel Synthetic Minority Oversampling Technique for Imbalanced Data Set Learning. *Neural Information Processing, Pt Ii*, 7063:735-+.

Boutell, M.R.*, et al.* (2004) Learning multi-label scene classification. *Pattern Recogn*, 37(9):1757-1771.

Breiman, L. (2001) Random forests. *Mach Learn*, 45(1):5-32.

Carhart, R.E., Smith, D.H. and Venkataraghavan, R. (1985) Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.*, 25(2):64-73.

Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. *Mach Learn*, 20(3):273-297.

Cover, T. and Hart, P. (1967) Nearest neighbor pattern classification. *IEEE Trans. Inf Theory*, 13(1):21-27.

Durant, J.L.*, et al.* (2002) Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.*, 42(6):1273-1280.

Duvenaud, D.*, et al.* Convolutional Networks on Graphs for Learning Molecular Fingerprints. In, *ArXiv e-prints*. 2015.

Gibaja, E. and Ventura, S. (2015) A Tutorial on Multilabel Learning. *Acm Computing Surveys*, 47(3):1-38.

Godbole, S. and Sarawagi, S. (2004) Discriminative methods for multi-labeled classification. *Advances in Knowledge Discovery and Data Mining, Proceedings*, 3056:22-30.

Landrum, G. RDKit. In.; 2017.

Nilakantan, R.*, et al.* (1987) Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.*, 27(2):82-85.

O'Boyle, N.M.*, et al.* (2011) Open Babel: An open chemical toolbox. *J. Cheminform.*, 3:33.

Pedregosa, F.*, et al.* (2011) Scikit-learn: Machine Learning in Python. *J Mach Learn Res*, 12:2825-2830.

Rabal, O., Amr, F.I. and Oyarzabal, J. (2015) Novel Scaffold FingerPrint (SFP): applications in scaffold hopping and scaffold-based selection of diverse compounds. *J. Chem. Inf. Model.*, 55(1):1-18.

Read, J.*, et al.* (2011) Classifier chains for multi-label classification. *Mach Learn*, 85(3):333-359.

Rogers, D. and Hahn, M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, 50(5):742-754.

Tsoumakas, G., Katakis, I. and Vlahavas, I. Mining Multi-label Data. In: Maimon, O. and Rokach, L., editors, *Data Mining and Knowledge Discovery Handbook*. Boston, MA: Springer US; 2010. p. 667-685.

Yang, H.*, et al.* (2018) In Silico Prediction of Chemical Toxicity for Drug Design Using Machine Learning Methods and Structural Alerts. *Front Chem*, 6:30.

Yang, Z.P. and Gao, D.Q. (2013) Classification for Imbalanced and Overlapping Classes Using Outlier Detection and Sampling Techniques. *Appl Math Inform Sci*, 7:375-381.

Zhang, M.L. and Zhou, Z.H. (2014) A Review on Multi-Label Learning Algorithms. *IEEE Trans. Knowl. Data Eng.*, 26(8):1819-1837.

**Table 1.** Number of training molecules and the sources of each model.

|  |  |  |  |
| --- | --- | --- | --- |
| Endpoint name (abb) | Numbers  Postive/negative  I/II/III … | Total number | Model type |
| Blood brain barrier (BBB) | 1438/401 | 1839 | Binary |
| Plasma protein binding (PPB) | - | 1209 | Regression |
| Caco-2 (caco2) | 303/371 | 674 | Binary |
| Human intestinal absorption (HIA) | 500/78 | 578 | Binary |
| Water solubility (WS) | - | 1708 | Regression |
| Human oral bioability (HOB) | 509/486 | 995 | Binary |
| Subcellular localization (Subc) |  | 614 | Quaternary |
| CYP1a2 inhibitor (cyp1a2i) | 7415/7488 | 14903 | Binary |
| CYP2d6 inhibitor (cyp2d6i) | 3060/11681 | 14741 | Binary |
| CYP2c9 inhibitor (cyp2c9i) | 4978/9731 | 14709 | Binary |
| CYP2c19 inhibitor (cyp2c19i) | 6041/8535 | 14576 | Binary |
| CYP3a4 inhibitor (cyp3a4i) | 6707/11854 | 18561 | Binary |
| CYP2d6 substrate (cyp2d6s) | 191/480 | 671 | Binary |
| CYP2c9 substrate (cyp2c9s) | 142/531 | 673 | Binary |
| CYP3a4 substrate (cyp3a4s) | 357/317 | 674 | Binary |
| P-glycoprotein substrate (Pgps) | 718/847 | 1565 | Binary |
| P-glycoprotein inhibitor (Pgpi) | 1172/771 | 1943 | Binary |
| Breast Cancer Resistance Protein inhibitor (BCRPi) | 432/538 | 970 | Binary |
| Bile Salt Export Pump inhibitor (BSEPi) | 317/290 | 607 | Binary |
| Organic Cation Transport Protein 1 inhibitor (OCT1i) | 60/124 | 184 | Binary |
| Organic Cation Transport Protein 2 inhibitor (OCT2i) | 244/633 | 904 | Binary |
| Multidrug and toxin extrusion transporter 1 inhibitor (MATE1) | 80/738 | 818 | Binary |
| OATP1b1 inhibitor1 (oatp1b1i) | 1657/198 | 1855 | Binary |
| OATP1b3 inhibitor1 (oatp1b3i) | 1743/130 | 1873 | Binary |
| OATP2b1 inhibitor1 (oatp2b1i) | 44/175 | 219 | Binary |
| Acute oral toxicity (AO) |  | 10207 | Quaternary &  regression |
| Hepatotoxicity (Heap) | 3115/593 | 3708 | Binary |
| Micronucleus (MN) | 237/339 | 576 | Binary |
| Human either-a-go-go (hERG) | 717/261 | 978 | Binary |
| Ames mutagenesis (Ames) | 4866/3482 | 8348 | Binary |
| Eye Irritation (EI) | 3874/1346 | 5220 | Binary |
| Eye Corrosion (EC) | 887/1412 | 2299 | Binary |
| Carcinogenesis (Carc) | 476/440 | 916 | Ternary &  Binary |
| Fish aquatic toxicity (FAQ) | 366/188 | 554 | Binary |
| Crustacean toxicity (Crust) | 336/324 | 660 | Binary |
| Bio-degradation (Biod) | 591/1014 | 1605 | Binary |
| Avian toxicity (Avian) | 94/484 | 574 | Binary |
| Tetrahymena pyriformis toxicity (TPT) | - | 1571 | Regression |
| Honey bee toxicity (HBT) | 99/96 | 195 | Binary |
| Androgen Receptor (AR) | 1059/4860 | 5919 | Binary2 |
| Peroxisome Proliferator-Activated Receptor gamma (PPARg) | 576/4880 | 5456 | Binary2 |
| Aromatase Inhibitor (Aro) | 697/5457 | 6154 | Binary2 |
| Glucocorticoid Receptor Modulator (GR) | 601/5454 | 6055 | Binary2 |
| Thyroid Receptor Modulator (TR) | 373/4999 | 5372 | Binary2 |
| Estrogen Receptor Modulator (ER) | 1338/4252 | 6055 | Binary2 |

1OATP: organic anion-transporting polypeptide. 2Multi-label model.

**Table 2.** The performance of the binary models in admetSAR.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Binary model | AUC | Accuracy | Sensitivity | Specificity |
| BBB | 0.944 | 0.907 | 0.921 | 0.861 |
| caco2 | 0.857 | 0.768 | 0.73 | 0.799 |
| HIA | 0.958 | 0.965 | 0.978 | 0.885 |
| HOB | 0.752 | 0.697 | 0.739 | 0.654 |
| cyp1a2i | 0.8832 | 0.8147 | 0.7985 | 0.8307 |
| cyp2d6i | 0.84 | 0.855 | 0.446 | 0.96 |
| cyp2c9i | 0.858 | 0.802 | 0.637 | 0.886 |
| cyp2c19i | 0.8712 | 0.8054 | 0.748 | 0.846 |
| cyp3a4i | 0.848 | 0.645 | 0.865 | 0.525 |
| cyp2d6s | 0.772 | 0.775 | 0.476 | 0.895 |
| cyp2c9s | 0.625 | 0.779 | 0.234 | 0.924 |
| cyp3a4s | 0.695 | 0.66 | 0.676 | 0.641 |
| Pgps | 0.865 | 0.802 | 0.76 | 0.837 |
| Pgpi | 0.931 | 0.861 | 0.901 | 0.8 |
| BCRPi | 0.903 | 0.85 | 0.808 | 0.884 |
| BSEPi | 0.9 | 0.827 | 0.82 | 0.838 |
| OCT1i | 0.914 | 0.87 | 0.68 | 0.96 |
| OCT2i | 0.814 | 0.808 | 0.481 | 0.927 |
| MATE1i | 0.714 | 0.907 | 0.125 | 0.992 |
| oatp1b1i | 0.815 | 0.896 | 0.962 | 0.343 |
| oatp1b3i | 0.76 | 0.927 | 0.971 | 0.338 |
| oatp2b1i | 0.786 | 0.885 | 0.386 | 0.96 |
| Heap | 0.719 | 0.833 | 0.943 | 0.255 |
| MN | 0.937 | 0.87 | 0.819 | 0.906 |
| hERG | 0.811 | 0.804 | 0.876 | 0.57 |
| Ames | 0.914 | 0.843 | 0.849 | 0.835 |
| EI | 0.975 | 0.949 | 0.972 | 0.883 |
| EC | 0.992 | 0.963 | 0.949 | 0.972 |
| Carc | 0.847 | 0.896 | 0.583 | 0.982 |
| Faq | 0.892 | 0.839 | 0.863 | 0.793 |
| Crust | 0.842 | 0.766 | 0.741 | 0.793 |
| Biod | 0.893 | 0.834 | 0.746 | 0.885 |
| Avian | 0.85 | 0.897 | 0.415 | 0.992 |
| HBT | 0.851 | 0.785 | 0.768 | 0.802 |

**Table 3.** The performance of the multi-classification models in admetSAR.

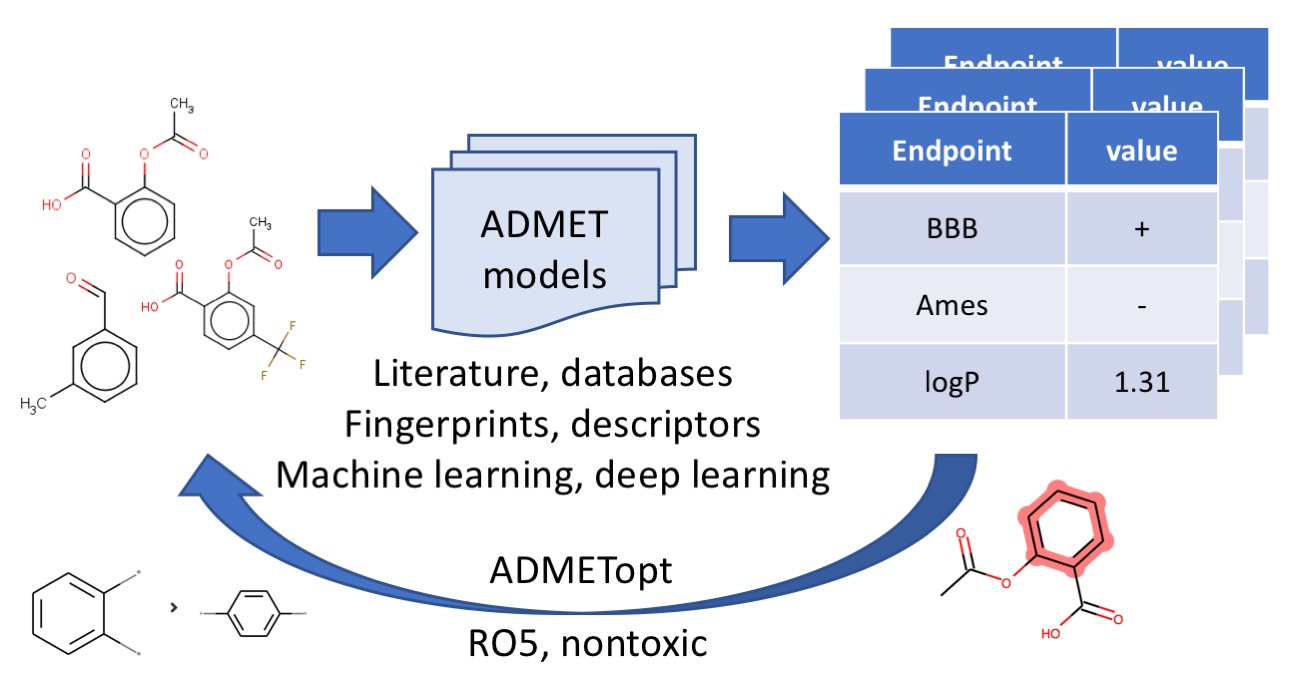
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Endpoint | Q1 | Q2 | Q3 | Q4 | Qtotal |
| Ao | 0.782 | 0.767 | 0.912 | 0.583 | 0.832 |
| Carc | 0.844 | 0.659 | 0.886 | - | 0.816 |
| Subc | 0.720 | 0.811 | 0.679 | 0.795 | 0.762 |

**Table 4.** The performance of regression models in admetSAR.

|  |  |  |
| --- | --- | --- |
| Endpoint | R2 | RMSE |
| WS | 0.810 | 0.823 |
| TPT | 0.822 | 0.447 |
| AO | 0.522 | 0.672 |
| PPB | 0.668 | 0.191 |

**Table 5.** The performance of multi-label models in admetSAR.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | AUC | Accuracy | Sensitivity | Specificity |
| AR | 0.886 | 0.831 | 0.788 | 0.84 |
| ER | 0.880 | 0.795 | 0.803 | 0.793 |
| GR | 0.879 | 0.829 | 0.750 | 0.837 |
| Aro | 0.886 | 0.705 | 0.879 | 0.684 |
| TR | 0.838 | 0.790 | 0.761 | 0.793 |
| PPARg | 0.818 | 0.761 | 0.766 | 0.761 |
|  |  |  |  |  |
| Subset\_accuracy | Hamming\_loss | Jaccard\_similarity | Recallmicro | Precisionmicro |
| 0.096 | 0.384 | 0.468 | 0.863 | 0.523 |



**Figure S1.** Scheme of admetSAR 2. ADMET data are collected from literature and databases, represented by fingerprints and descriptors, and the models were built by machine (deep) learning methods. ADMETopt can be used to optimize the ADMET properties of a query compound by scaffold hopping.