Advanced Drug Delivery Reviews 86 (2015) 83–100

Contents lists available at ScienceDirect

Advanced Drug Delivery Reviews

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

Recent progresses in the exploration of machine learning methods as in-silico ADME prediction tools☆

L. Tao [a,b], P. Zhang [b], C. Qin [b], S.Y. Chen [b], C. Zhang [b], Z. Chen [c], F. Zhu [d], S.Y. Yang [a], Y.Q. Wei [a], Y.Z. Chen [b,*]

[a] State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, West China Medical School, Sichuan University, Chengdu 610041, China
[b] Bioinformatics and Drug Design Group, Department of Pharmacy, Center for Computational Science and Engineering, National University of Singapore, Singapore 117543, Singapore
[c] Zhejiang Key Laboratory of Gastro-intestinal Pathophysiology, Zhejiang Hospital of Traditional Chinese Medicine, Zhejiang Chinese Medical University, No. 54 Youdian Road, Hangzhou 310006, China
[d] Innovative Drug Research Centre and College of Chemistry and Chemical Engineering, Chongqing University, Chongqing, China

ARTICLE INFO

Article history:
Received 14 November 2014
Received in revised form 18 March 2015
Accepted 22 March 2015
Available online 30 May 2015

Keywords:
ADME
Absorption
Distribution
Metabolism
Excretion
Drug discovery
Machine learning
Molecular descriptors
QSAR

ABSTRACT

In-silico methods have been explored as potential tools for assessing ADME and ADME regulatory properties particularly in early drug discovery stages. Machine learning methods, with their ability in classifying diverse structures and complex mechanisms, are well suited for predicting ADME and ADME regulatory properties. Recent efforts have been directed at the broadening of application scopes and the improvement of predictive performance with particular focuses on the coverage of ADME properties, and exploration of more diversified training data, appropriate molecular features, and consensus modeling. Moreover, several online machine learning ADME prediction servers have emerged. Here we review these progresses and discuss the performances, application prospects and challenges of exploring machine learning methods as useful tools in predicting ADME and ADME regulatory properties.

Contents

☆ This review is part of the *Advanced Drug Delivery Reviews* theme issue on "In silico ADMET predictions in pharmaceutical research".

* Corresponding author at: Bioinformatics and Drug design Group, Department of Pharmacy, Center for Computational Science and Engineering, National University of Singapore, S16, Level 8, 6 Science Drive 2, Singapore 117546, Singapore. Tel.: +65 6874 6877; fax: +65 6774 6756.
E-mail address: phacyz@nus.edu.sg (Y.Z. Chen).

http://dx.doi.org/10.1016/j.addr.2015.03.014
0169-409X/

1. Introduction

The discovery and optimization of therapeutic agents with desirable pharmacodynamics, pharmacokinetic toxicological properties is the key focus of drug development efforts [1]. Predictive tools for accurately assessing pharmacokinetic and toxicological properties as well as pharmacodynamic properties in early development stages are highly useful for increased productivity in drug discovery processes [1–3]. As part of the efforts for developing these tools, computational methods have been developed and improved for the prediction of compound absorption, distribution, metabolism, and excretion (ADME) properties [4,5]. In particular, machine learning (ML) methods have shown promising potential in predicting ADME properties by correlating these properties to molecular features and by establishing the complex structure–property relationships for diverse ranges of molecular structures and mechanisms [6,7].

More recently, efforts have been directed at the development and refinement of ML models for improved prediction and more extensive coverage of various ADME properties particularly excretion [8–10] and distribution [11,12] properties, and for the prediction of regulators of drug metabolism [13–16] and excretion [8] implicated in drug–drug interactions and multi-drug resistance respectively. Efforts have also been made to further explore consensus modeling for improved prediction of the ADME properties and ADME regulatory properties of drug candidates [8,13]. Moreover, online machine learning ADME and ADME regulatory property prediction servers have emerged [15,17]. Here we review these progresses and discuss the performances, application prospects and challenges of exploring ML methods as tools for predicting ADME and ADME regulatory properties.

2. Molecular descriptors for representing compounds in ADME prediction

Molecular descriptors have been extensively used for representing structural and physicochemical properties of compounds from their molecular structures. The compounds associated with a specific ADME property are typically of high structural and mechanistic diversity. Therefore, the prediction of various ADME properties requires different sets of molecular descriptors that adequately cover the relevant molecular features. A large variety of >3000 molecular descriptors can be computed from such software as DRAGON [18], E-DRAGON [19], Molconn-Z [20], JOELib [21], MODEL [22] and PaDEL [23]. These descriptors are broadly divided into 18 classes, which include constitutional descriptors such as molecular weight, geometrical descriptors such as surface areas, topological descriptors such as topological index, RDF descriptors representing such features as inter-atomic distances [24], molecular walk counts [25], 3D-MoRSE descriptors describing such properties as polarizabilities [26], BCUT descriptors representing such information as connectivity [27], WHIM descriptors describing such features as molecular symmetries [28], Galvez topological charge indices and charge descriptors [29], GETAWAY descriptors [30], 2D autocorrelations, functional groups, atom-centered descriptors, aromaticity indices [31], Randic molecular profiles [32], electrotopological state descriptors [33], and linear solvation energy relationship descriptors [34].

Not all molecular descriptors are necessary for predicting an ADME property. The relevant descriptors can be selected by either intuition [35] or feature selection methods. The commonly used feature selection methods include recursive feature eliminations (RFE) [36], genetic algorithm-based approaches [37], and simulated annealing-based methods [38]. Some methods such as RFE have gained popularity due to their effectiveness for discovering features informative of ADME properties [39–41]. The general feature selection strategy is outlined as follows: First, a ML model is generated by using either all or a few descriptors. This model is further used to rank the predictive contribution of the descriptors with the least contributing ones eliminated. In the next step, a new ML model is constructed by using either the reduced set of descriptors for the all-descriptors model or the retained set of descriptors plus newly added descriptors for the few-descriptors model. This new model is subsequently used to rank and eliminate/add descriptors. This iteration process continues until all of the irrelevant descriptors are eliminated or added. In many cases, it is difficult to uniquely select an optimal set of descriptors due to the high redundancy and overlapping of many descriptors [42]. Separate sets of descriptors with redundant and/or overlapping feature coverage have been found to give similar predictive accuracies [43]. The interpretation of the predictive results should be more appropriately conducted at the feature class level where redundant and overlapping descriptors are grouped into one class [44–46].

The currently available descriptors, though capable of representing a wide range of molecular features, seem to be insufficient to describe certain molecular features relevant for ADME prediction, leading to higher false positive rates in the prediction of some ADME properties [39,47]. Examples of the inadequately-described molecular features are inflexible multi-rings, highly polar tetrazole rings, complex two ring system with multiple heteroatoms, polycyclic aromatic structures, long flexible chains, hydrazine group, and multiple ionisable groups. There is a need to develop new molecular descriptors covering these and other molecular features.

3. Commonly used machine learning methods for developing classification models

A number of ML methods have been used for developing ADME predictive tools. These include *Linear Discriminant Analysis* (LDA), *k Nearest Neighbor* (kNN), *Artificial Neural Network* (ANN), *Probabilistic Neural Network* (PNN), *Support Vector Machine* (SVM), *Decision Tree* (DT), *Recursive Partitioning* (RP), *Random Forest* (RF), *Naïve Bayesian* (NB), *Multiple Linear Regression* (MLR), *Partial Least Squares Regression* (PLSR), *kNN Regression* (kNNR), *Support Vector Regression* (SVR), *Random Forest Regression* (RFR), and combined classifier approaches. Websites for the freely downloadable codes of these methods are given in Table 1.

Table 1
Some websites that contain freely downloadable codes of machine learning methods.

Decision tree	
Simple Decision Tree	https://sites.google.com/site/simpledecisiontree/
OC1	http://www.cbcb.umd.edu/~salzberg/announce-oc1.html
SMILES	http://users.dsic.upv.es/~flip/smiles/
PC4.5	http://www.cs.nyu.edu/~binli/pc4.5/
YaDT	http://www.di.unipi.it/~ruggieri/software.html
C4.5 and C5.0	http://www.rulequest.com/Personal/
Random forests	
Random Forests	http://www.stat.berkeley.edu/~breiman/RandomForests/
randomForest R package	http://cran.r-project.org/web/packages/randomForest/index.html
FastRandomForest	https://code.google.com/p/fast-random-forest/
KNN	
kNN classifier	http://www.fit.vutbr.cz/~bartik/Arcbc/kNN.htm
k Nearest Neighbor demo	http://www.cs.cmu.edu/~zhuxj/courseproject/knndemo/KNN.html
GPU-FS-kNN	http://sourceforge.net/projects/gpufsknn/
GA/KNN	http://www.niehs.nih.gov/research/resources/software/biostatistics/gaknn/
Dense K Nearest Neighbor	http://www.autonlab.org/autonweb/10522.html
Neural network	
BrainMaker	http://www.calsci.com/
fann	http://leenissen.dk/fann/
NuClass	http://www.uta.edu/faculty/manry/new_software.html
sciengyrpf	http://sourceforge.net/projects/sciengyrpf/
Sharky Neural Network	http://sharktime.com/us_SharkyNeuralNetwork.html
SVM	
LIBSVM	http://www.csie.ntu.edu.tw/~cjlin/libsvm/
SVM light	http://svmlight.joachims.org/
M-SVM	http://www.loria.fr/~guermeur/
mySVM	http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/index.html
e1071 R package	http://cran.r-project.org/web/packages/e1071/index.html
BSVM	http://www.csie.ntu.edu.tw/~cjlin/bsvm/
LS-SVMlab	http://www.esat.kuleuven.be/sista/lssvmlab/

3.1. Linear discriminant analysis (LDA)

LDA [48] (Fig. 1) separates two classes of vectors by constructing a hyperplane defined by the following linear discriminant function $L = \sum_i^k w_i x_i$, where L is the resultant classification score and w_i is the weight associated with the corresponding descriptor x_i. A positive or negative L value indicates that a vector x belongs to the positive or negative class respectively.

3.2. k nearest neighbor (kNN) and kNN regression (kNNR)

In kNN (Fig. 2), the Euclidean distance $D = \sqrt{||x - x_i||^2}$ between an unclassified vector x and each individual vector x_i in the training set is measured [49]. A total of k number of vectors nearest to the unclassified vector x are used to determine the class of that unclassified vector. The class of the majority of the k nearest neighbors is chosen as the predicted class of the unclassified vector x. The activity of the studied compound is determined by averaging the activity values of a total of k number of training compounds nearest to that compound $\hat{y} = (\sum_{i=1}^{k} y_i)/k$.

3.3. Artificial neural network (ANN)

An ANN (Fig. 3) is an information-processing paradigm inspired by the way the densely interconnected, parallel structure of the mammalian brain processes information. ANN consists of a set of highly interconnected entities, called *nodes* or *units*. Each unit is designed to mimic its biological counterpart, the neuron, mathematically. Each node accepts a weighted set of inputs and responds with an output respectively [50].

3.4. Probabilistic neural network (PNN)

PNN (Fig. 4) is a form of neural network that uses Bayes optimal decision rule $h_i c_i f_i(x) > h_j c_j f_j(x)$ for classification [51], where h_i and h_j are the prior probabilities, c_i and c_j are the costs of misclassification and $f_i(x)$ and $f_j(x)$ are the probability density function for class i and j respectively. An unknown vector x is classified into population i if the product of all the three terms is greater for class i than for any other class j (not equal to i). In most applications, the prior probabilities and costs of misclassifications are treated as being equal. The probability density function for each class for a multivariate case can be estimated by using the Parzen's nonparametric estimator $g(x) = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-\sum_{j=1}^{p}\left(\frac{x_j - x_{ij}}{\sigma_j}\right)^2\right)$ [52], where n is the number of samples in the population, p is the number of features, x_j is the jth feature of an unclassified sample, x_{ij} is the jth feature of the ith sample in the population, and σ_j is the smoothing factor of this feature. Traditional neural networks such as feed-forward back-propagation neural network rely on multiple parameters and network architectures to be optimized. In contrast, PNN only has a single adjustable parameter, a smoothing factor σ for the radial basis function in the Parzen's nonparametric estimator. Thus the training process of PNN is usually orders of magnitude faster than those of the traditional neural networks.

3.5. Support vector machine (SVM) and support vector regression (SVR)

SVM is illustrated in Fig. 5. Linear SVM constructs a hyperplane separating two different classes of feature vectors with a maximum margin [53]. This hyperplane is constructed by finding a vector w and a parameter b that minimizes $||w||$ which satisfies the following conditions:

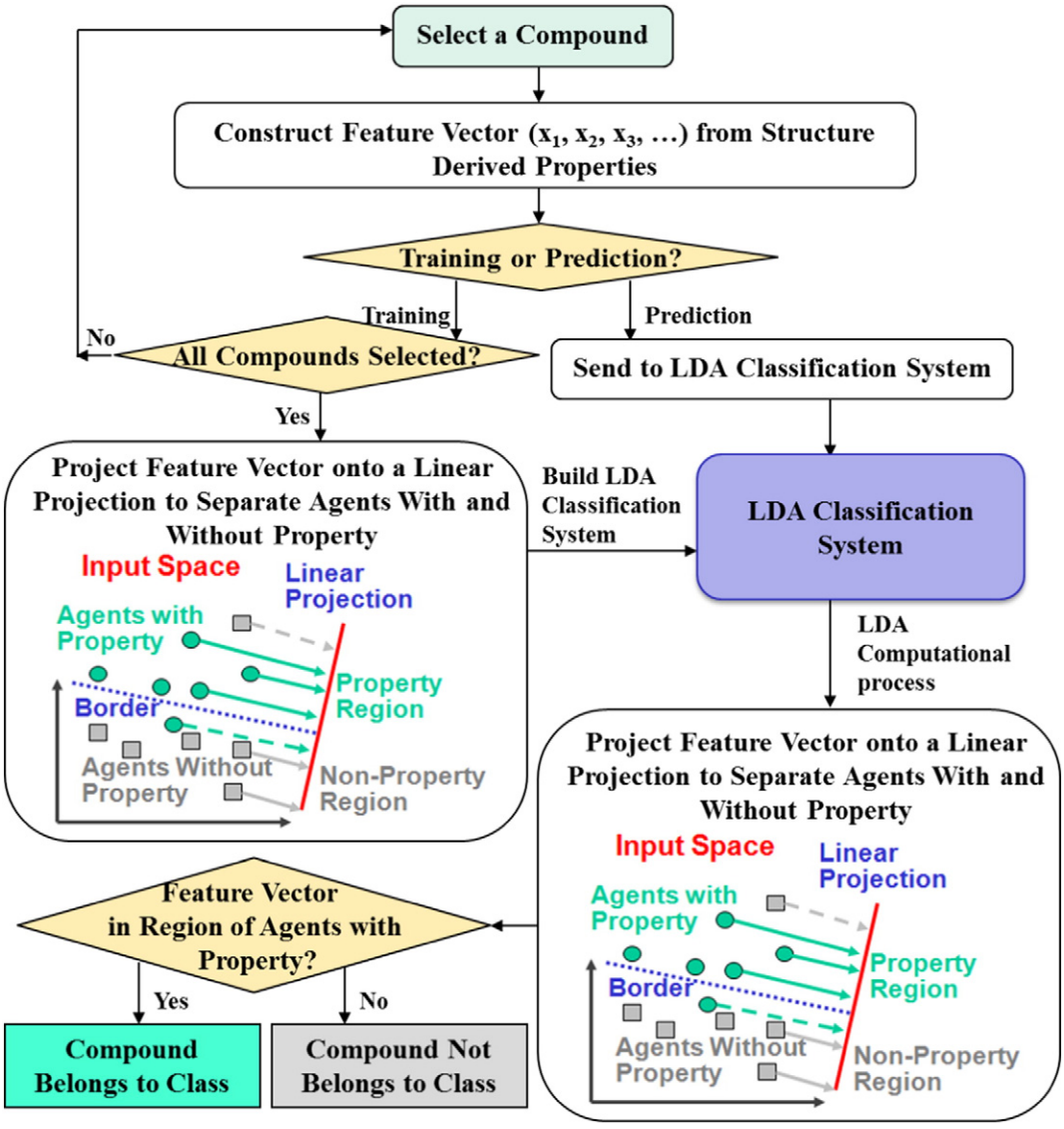

Fig. 1. Schematic diagram illustrating the processing of the prediction of compounds of a particular property from their structure by using a machine learning method — linear discriminative analysis (LDA). Feature vector $(x_1, x_2, x_3, \ldots)$ represents structural and physiochemical properties such as hydrophobicity, volume, polarizability, etc.

$w \cdot x_i - b \geq 1$ (positive class) and $w \cdot x_i - b \leq -1$ (negative class). Here x_i is a feature vector, w is a vector normal to the hyperplane, $\frac{b}{||w||}$ is the perpendicular distance from the hyperplane to the origin and $||w||$ is the Euclidean norm of w. Nonlinear SVM projects feature vectors into a high dimensional feature space by using a kernel function such as $k(x_i, x_j) = \exp(-\gamma|x_i - x_j|^2)$, for $\gamma > 0$, the linear SVM procedure is then applied to the feature vectors in this feature space. After the determination of w and b, a given vector x can be classified by using $f(x) = sign(\sum_{i=1}^{n} a_i^0 y_i K(x, x_i) + b)$, where the coefficients α_i^0 and b are determined by maximizing the following Lagrangian expression $\sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j K(x_i, x_j)$, under conditions $a_i \geq 0$ and $\sum_{i=1}^{n} a_i y_i = 0$. A positive or negative $f(x)$ value indicates that the vector x belongs to the positive or negative class respectively.

SVR is an extension of support vector machine (SVM) to solve nonlinear regression problems by introducing an ε-insensitive loss function [53–55]. A kernel function (in the form of a polynomial, gaussian, or sigmoidal function) is used to map the input vectors into a higher dimensional feature space and then a linear regression model is conducted in this feature space. The quality of estimation is measured by the ε-insensitive loss function $L(y, f(x, \omega)) = 0$ if $|y - f(x, w)| \leq \varepsilon$, otherwise $L(y, f(x, \omega)) = |y - f(x, \omega)| - \varepsilon$. The optimal regression function can be represented by $\hat{y} = \sum_{i=1}^{Nsv} (a_i - a_i^*) K(x_i, x) + b$ under the conditions $0 \leq a_i, a_i^* \leq C$ and $\sum_{i=1}^{n} (a_i + a_i^*) = 0$, where $\hat{y}$ represents the predicted activity value of a specific property, Nsv is the number of support vectors, constant C determines the trade-off between the flatness of function f and the amount up to which deviations larger than ε are tolerated and K is the kernel function, normally Gaussian kernel function $K(x_i, x_j) = e^{-\left(|x_j - x_i|^2\right)/(2\sigma^2)}$ is used.

3.6. C4.5 decision tree (C4.5 DT)

C4.5 DT (Fig. 6) is a branch-test-based classifier [56]. A branch of the decision tree corresponds to a group of classes and a leaf represents a specific class. A decision node specifies a test on a single attribute value, with one branch and its subsequent classes as possible outcomes. C4.5 decision tree uses recursive partitioning to examine every attribute of the data and rank them according to their ability to partition the remaining data, thereby constructing a decision tree. A vector x is classified by starting at

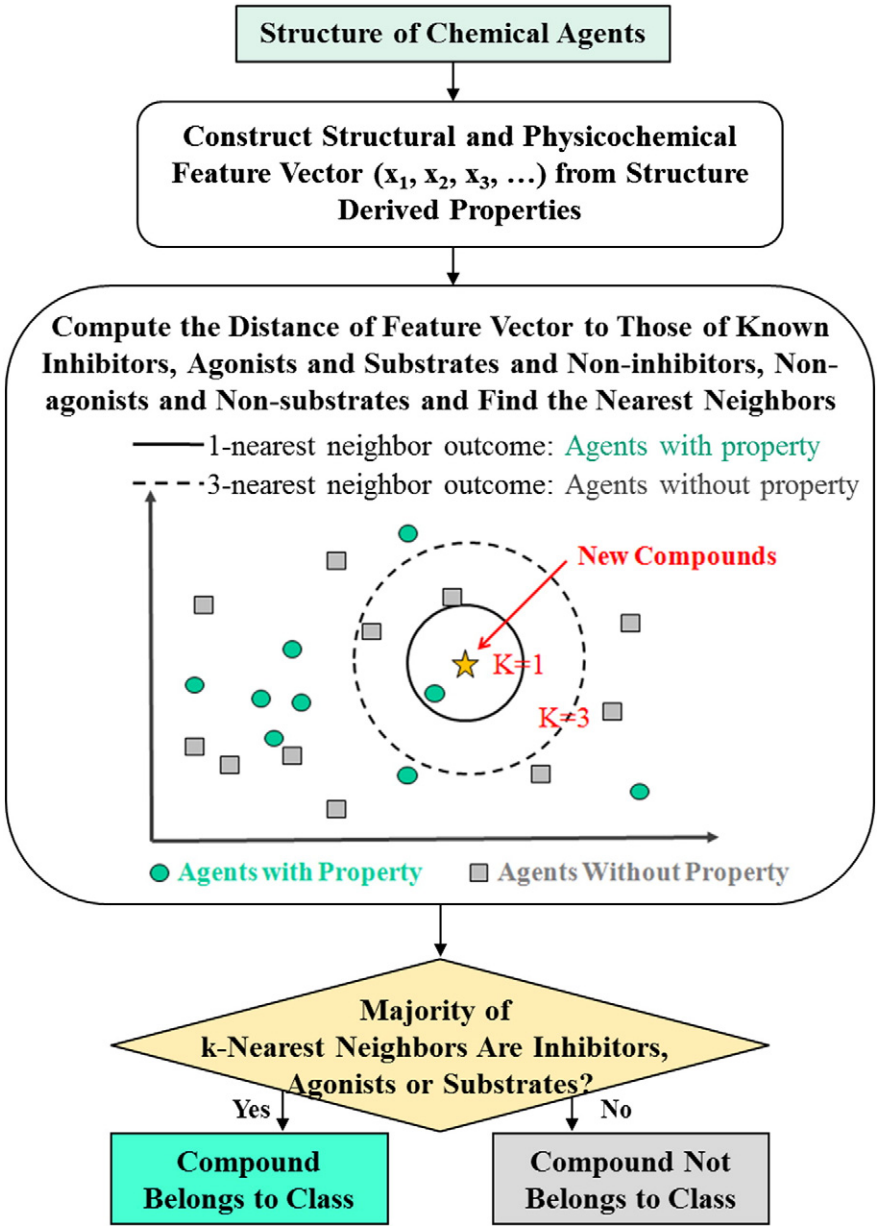

Fig. 2. Schematic diagram illustrating the processing of the prediction of compounds of a particular property from their structure by using a machine learning method — k nearest neighbors (kNN). Feature vector (x_1, x_2, x_3, …) is the same as that in Fig. 1.

the root of the tree and moving through the tree until a leaf is encountered. At each non-leaf decision node, a test is conducted to move into a branch. Upon reaching the destination leaf, the class of the vector x is predicted to be that of the leaf. This process continues to allow the tree to grow to the full size, which is then pruned back to an appropriate size based on the evaluation of its overall prediction performance.

The estimation criterion in the decision tree algorithm is the selection of an attribute to test at each decision node in the tree. The goal is to select the attribute that is most useful for classifying examples. A good quantitative measure of the worth of an attribute is a statistical property called information gain $Gain(S,A) = S - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} S_v$ that measures how well a given attribute separates the training examples according to their target classification. Here $S = \sum_{i=1}^{n} - p_i \log_2 p_i$, S is called entropy that characterizes the (im)purity of an arbitrary collection of examples, p_i is the proportion of S belonging to class I, values(A) is the set of all possible values for attribute A, and S_v is the subset of S for which attribute A has value v (i.e., $S_v = \{s\hat{I}S|A(s) = v\}$).

3.7. Recursive partitioning (RP) classifiers

RP creates a decision tree to classify compounds into separate classes based on a set of predefined variables (e.g. descriptors). RP models are constructed by successively splitting a dataset into increasingly homogeneous subsets until they can no longer be split, based on a set of "stopping rules". At each splitting point, the RP algorithm searches a pool of independent variables (e.g. descriptors) and identifies a single variable and the corresponding splitting value that best purify the group of compounds entering the node. The splitting process continues until either no further improvement can be achieved, or the number of compounds in each purified group is too small to justify further splitting.

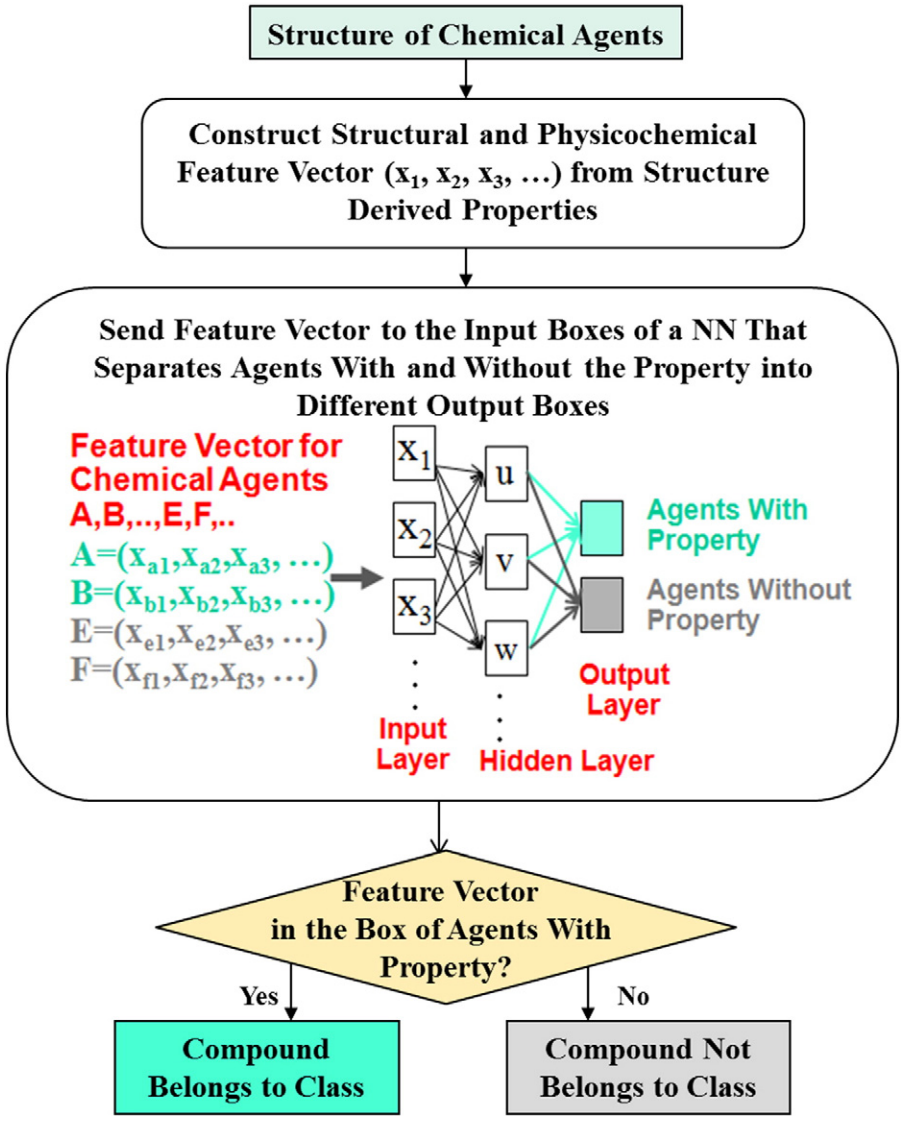

Fig. 3. Schematic diagram illustrating the processing of the prediction of compounds of a particular property from their structure by using a machine learning method — artificial neural network (ANN) or neural network (NN). A, B are the agents with this property, E, F are the agents without this property, and their feature vector ($x_1, x_2, x_3, \ldots$) is the same as that in Fig. 1.

3.8. Random forest (RF)

RF (Fig. 7) is a non-linear multiple classification algorithm, by making prediction based on majority voting from an ensemble of unpruned decision trees. Each decision tree is grown by a bootstrap sampling of the training dataset. Each node is split by a subset of features randomly chosen at that node, where the best node split is selected based on the criterion to minimize the variance within the branches. Each leaf in each decision tree is an entry in the training data.

3.9. Naïve Bayesian classifiers

A naive Bayesian classifier is a simple probabilistic classifier based on the Bayes' theorem with strong (naive) independence assumptions [57]. The probability model for a classifier is a conditional model $P(Y|x_1, \ldots, x_n)$ over a dependent class variable C with a small number of outcomes or classes (e.g. $C = +$ for compounds associated with an ADME property, $C = -$ for compounds not associated with the property), conditional on feature variables x_1 through x_n (e.g. molecular descriptors). When Bayes's theorem is used: $P(+|x_1, \ldots, x_n) = \frac{P(x_1, \ldots, x_n|+)P(+)}{P(x, \ldots, x_n)}$ where $P(x_1, \ldots, x_n|+)$ is the conditional probability of a particular compound being classified as a member of the $+$ class, $P(+)$ is the prior probability, a probability induced from a set of compounds in the training set; $P(x_1, \ldots, x_n)$ is the marginal probability of the given descriptors that will occur in the training set.

3.10. Multiple linear regression (MLR)

MLR [58] is one of the most commonly used and simplest methods for constructing QSPR models. A MLR model is constructed under the assumption that a linear relationship exists between a set of molecular descriptors of a compound (which is represented by a feature vector x with each descriptor as its component) and a specific ADME activity (which is represented by a quantity y). A MLR model can be described using the following equation $\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$, where $\{x_1, \ldots, x_n\}$ are molecular descriptors, β_0 is the regression model constant, β_1 to β_n are the coefficients for individual

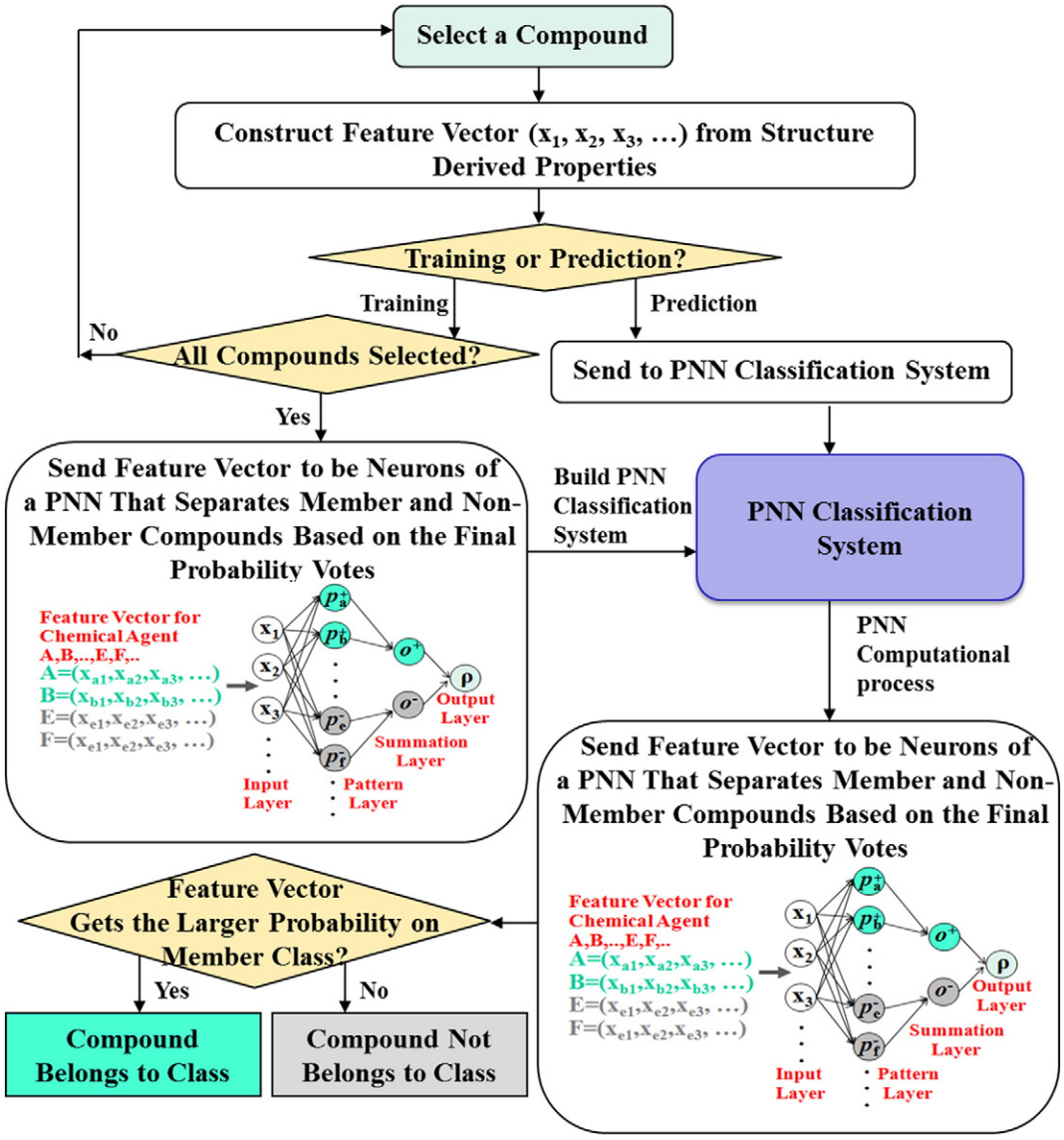

Fig. 4. Schematic diagram illustrating the processing of the prediction of compounds of a particular property from their structure by using a machine learning method — probabilistic neural network (PNN). A, B, E, F and feature vector (x_1, x_2, x_3, …) are the same as those in Figs. 1 and 3.

descriptor x_1 to x_n. The values for β_0 to β_n are chosen by minimizing the sum of squares of the residuals between the observed and predicted values defined by the equation so as to give the best prediction of y from x.

The advantage of MLR is its simplistic form and easily interpretable mathematical expression. The sign of the coefficients β_1 to β_n indicates whether each molecular descriptor contributes positively or negatively to a specific activity and their magnitudes indicates the relative importance of each descriptor to that activity. However, MLR works well only when the structure–property relationship is linear in nature, the set of molecular descriptors are mathematically independent (orthogonal) of each another, and the number of compounds in the training set exceeds the number of molecular descriptors by at least a factor of five [59]. It has been found that, when collinear descriptors are used, the derived coefficients β_1 to β_n tend to be larger than the real values and sometimes have opposite signs [60]. Therefore, the assumption of a linear relationship between a set of molecular descriptors and a specific activity may not always be appropriate, especially in the cases involving multiple mechanisms.

3.11. Partial least squares regression (PLSR) and logistic PLS

PLSR is a generalization of MLR [61]. It has more enhanced capability than MLR in classifying data with strongly collinear (correlated), noisy, and numerous descriptors, and in simultaneously predicting the multiple responses. Given a feature vector x composed of molecular descriptors $\{x_1, \ldots, x_n\}$, a linear PLSR model finds a few "new" variables, called X-scores and denoted by t_a ($a = 1,2,\ldots,A$), for predicting the response y as linear combinations of $\{x_1, \ldots, x_n\}$ with the weights W_{na} ($a = 1,2,\ldots,A$), namely $t_{ia} = \sum_n W_{na} \times x_{in}$.

Logistic PLS (Fig. 8) is a variation of ordinary PLS, possessing all its useful features in combination with the ability to analyze binary data. The predicted value is the logit transformation of probability for a compound to possess a specific ADME property (p) which is calculated as the sum of the contributions of all descriptors: $logit(p) = \ln\left(\frac{p}{1-p}\right) = \sum_i a_i f_i + c$, where p is the probability for a compound to exhibit an ADME property; f_i is the occurrence sum of the ith descriptor, a_i is the statistical coefficient of the descriptor determined using logistic PLS, and c is the intercept.

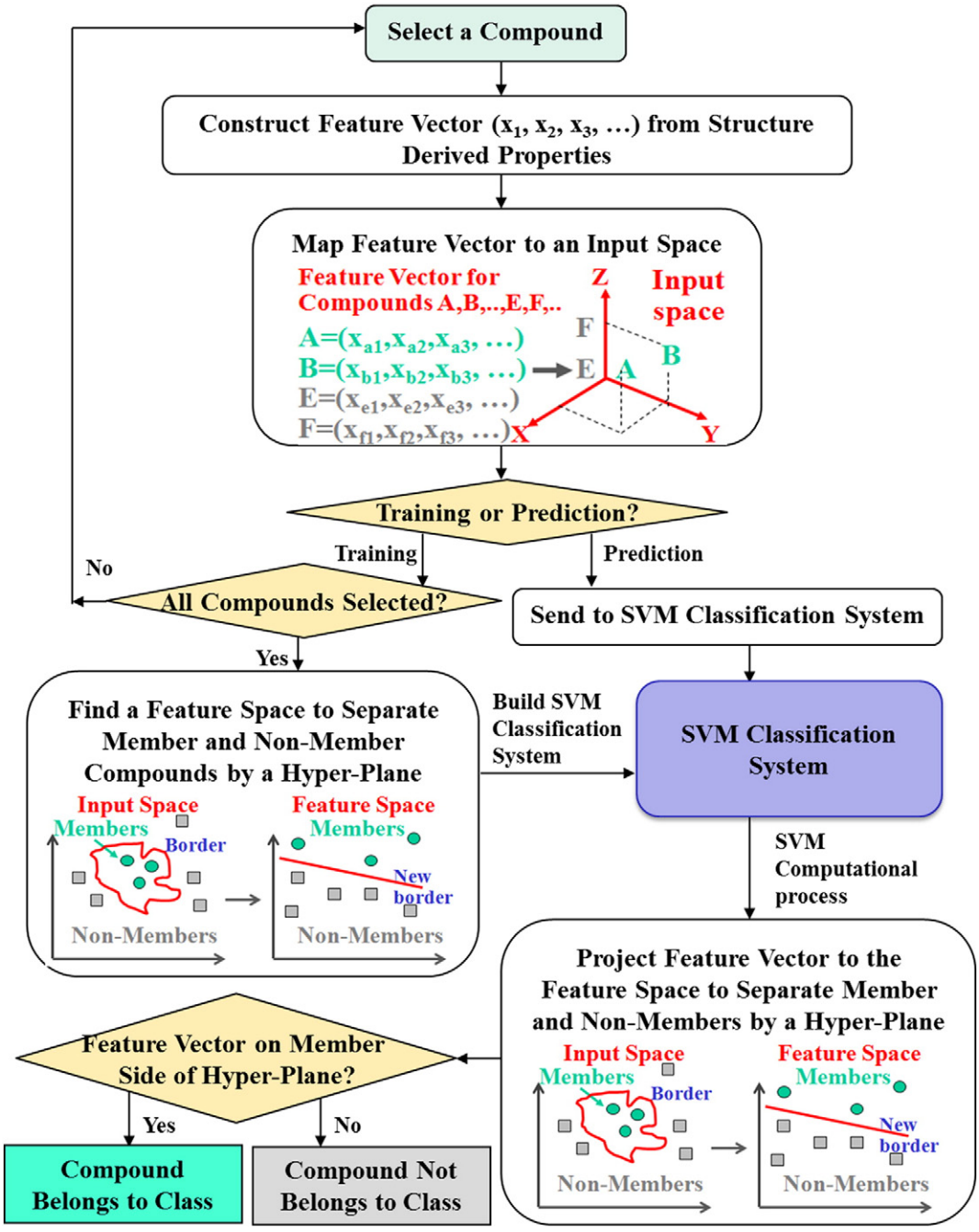

Fig. 5. Schematic diagram illustrating the processing of the prediction of compounds of a particular property from their structure by using a machine learning method — support vector machine (SVM). A, B, E, F and feature vector (x_1, x_2, x_3, …) are the same as those in Figs. 1 and 3.

3.12. Combined classifiers approach

The combined classifier approach, illustrated in Fig. 9, explores the collective predictive power of multiple ML classifiers, with the probability output of each independent ML model processed by a decision network [13]. This decision network consists of two layers of units, an input layer and an output layer. First, each individual ML model is trained separately. Then the predictive outcomes of the multiple ML models are evaluated to obtain the probability output (P_{i+1} and P_{i-1}, $i = 1,2,3,4$). These probability outputs are used as new descriptors to develop a neural network model, such as a Back-Propagation ANN model [13], that generates the final combination decision probability ($P_C{}^{+1}$ and $P_C{}^{-1}$). In predicting a specific ADME property of a compound, it is first classified by each ML model and then put into the neural network model developed by the training set to make the final prediction.

4. The exploration of machine learning classification methods for predicting ADME properties

ML classification methods classify compounds into one of the two opposing classes, one associated with a property (e.g. an ADME property) and the other not associated with the property. Because of their ability in classifying compounds of diverse range of structures and physicochemical properties, ML classification methods have been extensively explored for predicting various ADME properties that are

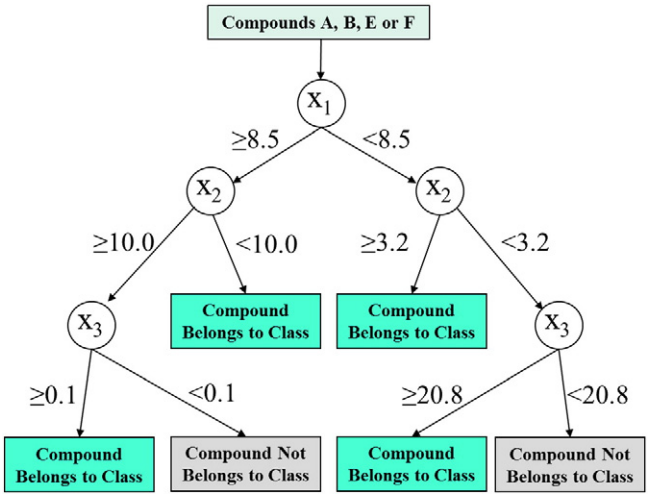

Fig. 6. Schematic diagram illustrating the processing of the prediction of compounds of a particular property from their structure by using a machine learning method — C4.5 decision tree. A, B, E, and F are the same as those in Fig. 3.

typically associated with compounds of diverse structures (e.g. substrates of a drug transporter) and in some cases multiple mechanisms (e.g. blood brain barrier crossing), Table 2 summarises the recently developed ML classification models and consensus models for predicting compounds of various ADME properties. Most of these models are for the prediction of transporter-mediated excretion properties, reflecting the recent efforts for more extensive coverage of key drug efflux and influx transporters and for improved predictive performance by expanded training datasets and appropriate selection of molecular descriptors [8–10]. Several models are for the prediction of distribution properties in plasma protein binding and blood brain barrier crossing, reflecting continuous efforts in developing improved predictive models by expanded training datasets and appropriate selection of molecular descriptors [17,62].

The ADME predictive performance of these ML models has typically been measured by the sensitivity SE, specificity SP, and overall accuracy AC, which measure the predictive accuracy for the compounds associated with an ADME property, compounds not associated with the property, and all compounds respectively. As shown in Table 2, the developed ML models showed good predictive capability in terms of these performance indicators. Specifically, the SEs and SPs in predicting human serum albumin substrates and blood–brain barrier crossing in ADME category E are in ranges of 81%–96%, 70%–83%, respectively. The SEs, SPs and ACs in predicting P-glycoprotein substrates in ADME category D are in ranges of 61–78%, 56%–76%, and 65%–79% respectively. The ACs in predicting BCRP, MRP1, MRP2, MRP3, MRP4, ASBT, OATP2B1, OCT1, and PEPT1 substrates in category D are in ranges of 66%–87%, 76%–96%, 76%–96%, 82%–100%, 67%–97%, 77%–100%, 50%–90%, 65%–96, and 69%–93% respectively. The overall SEs, SPs and ACs are in ranges of 61%–96%, 56%–83% and 50%–100%, with the majority of data concentrated between 74% and 92%, 66% and 76%, 72% and 92%, respectively. These further demonstrate that ML classification methods are capable of predicting ADME properties at reasonably good accuracy levels. The slightly lower SPs than SEs possibly reflects the tendency to optimize SE at a slight expense of SP.

5. The exploration of machine learning classification methods for predicting ADME regulatory properties

ML classification methods have also been extensively used for predicting regulators of drug ADME properties, particularly the inhibitors of drug efflux and influx transporters for regulating multidrug resistance (Table 3) [8,64,65] and the inhibitors of drug metabolism enzymes for assessing drug–drug interactions (Table 4) [13,14,66]. These studies have primarily focused on the extended coverage of drug transporters (9 transporters) [8] and metabolism enzymes (5 CYP enzymes CYP 1A2, 2C9, 2C19, 2D6, 3A4) [13,15,16], improvement of predictive performance by such strategies as the use of expanded training datasets [8,64–67], and both objectives [13,16].

The SEs, SPs, and ACs in predicting P-glycoprotein inhibitors are in ranges of 58%–99%, 47%–91%, and 62%–94% respectively. The ACs in predicting BCRP, MRP1, MRP2, MRP4, ASBT, MCT1, OATP2B1, OCT1, PEPT1, and hERG inhibitors are in ranges of 71%–87%, 78%–94%, 73%–98%, 48%–78%, 80%–97%, 100%–100%, 66%–89%, 76%–98%, 45%–87%, and 86%–89% respectively. The overall SEs, SPs and ACs in predicting inhibitors of these drug efflux and influx transporters are in ranges of 58%–99%, 47%–91% and 45%–100%, with the majority of data concentrated between 84% and 97%, 53% and 73%, 70% and 95%, respectively.

The SEs, SPs and ACs for predicting inhibitors of drug metabolism enzymes are in ranges of 26%–87%, 60%–96% and 64%–90%, with the majority of data concentrated between 73% and 87%, 65% and 88%, 73% and 85%, respectively. Specifically, the SEs, SPs and ACs in predicting CYP1A2 inhibitors are in ranges of 73%–87%, 65%–88%, and 73%–88 respectively, those in predicting CYP2C9 inhibitors are in ranges of 56%–84%, 69%–87%, and 67%–83% respectively, CYP2C19 inhibitors 52%–86%, 67%–86%, and 68%–85%, CYP2D6 inhibitors 26%–75%, 65%–96%, and 73%–90%, and CYP3A4 inhibitors 39%–84%, 60%–86%, and 64%–84%, respectively. These further demonstrate the capability of ML classification methods in predicting ADME properties. The training datasets (inhibitors of drug transporters and metabolism enzymes) in the majority of these studies are substantially larger than those of the substrates of drug transporters and metabolism enzymes. But their prediction performance is not markedly improved over that of the substrates of drug transporters and metabolism enzymes. One possible reason is the inadequate representation of the non-inhibitors of a specific drug transporter or metabolism enzyme. The number of the inhibitors in the published studies is typically in the range of a few hundred or less, which is unlikely to be sufficient to fully represent the vast non-inhibitor chemspace of millions of compounds in the current versions of the chemical databases.

6. The exploration of machine learning regression methods for predicting ADME and ADME regulatory properties

ML regression methods are intended for estimating the affinity/activity level in addition to the determination of whether or not a compound possesses or regulates a specific ADME property. Table 5 summarises the performance of the recently developed ML regression methods for predicting the affinity/activity level of ADME and ADME regulatory properties. Partly because of the limited availability of experimental affinity/activity levels, ML regression models have been developed for a limited variety of ADME and ADME regulatory properties and most of them have been trained by a significantly smaller training dataset than those of the ML classification models for predicting ADME and ADME regulatory properties. These models enable the prediction of female genital tract penetrators [71] in category A, apparent volume of distribution [11] and blood–brain barrier crossing [62] in category D, and intrinsic clearance [11,72] in category E. They also enable the prediction of inhibitors of P-glycoprotein [73,74] and 8 drug metabolism enzymes (CYP 1A2, 2C8, 2C9, 2A6, 2C19, 2D6, 3A4, and 17) [75].

The performance of these models has primarily been evaluated by the R^2 value, which measures the variance between the computed and experimental activity levels. Moreover, RMSE values have also been frequently computed for measuring the root mean square errors of the developed models. The computed R^2 values are ~0.4 in predicting female genital tract penetrators in ADME category A, 0.56–0.74 and 0.42–0.69 in predicting apparent volume of distribution and blood–brain barrier

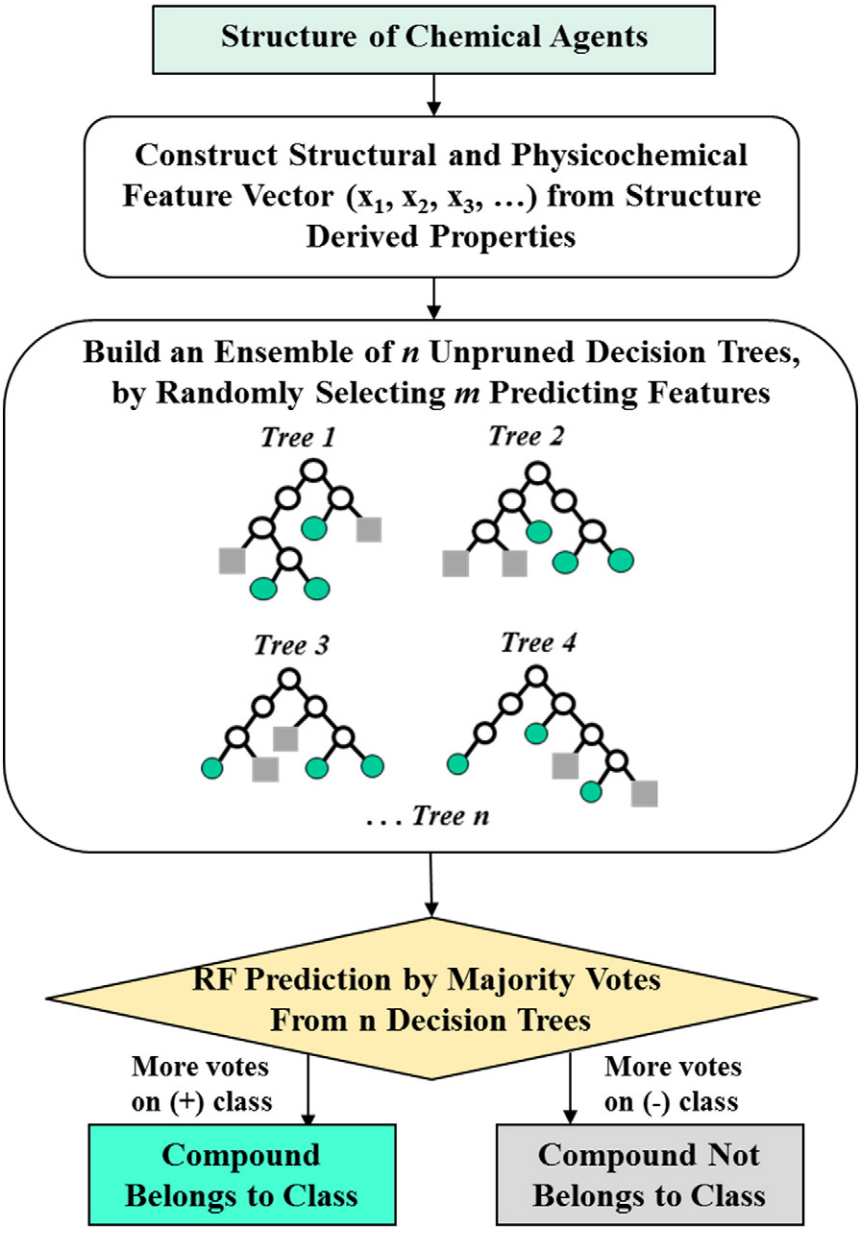

Fig. 7. Schematic diagram illustrating the processing of the prediction of compounds of a particular property from their structure by using a machine learning method — random forest (RF). Feature vector (x_1, x_2, x_3, …) is the same as that in Fig. 1.

crossing in category D, 0.43–0.96 and 0.64 in predicting intrinsic clearance and systemic clearance in category E, 0.69–0.87 in predicting P-glycoprotein inhibitors, and 0.65, 0.96, 0.99, 0.99, 0.99, 0.98, 0.99, 0.99 in predicting CYP2D6, CYP1A2, CYP3A4, CYP2A6, CYP2C9, CYP2C8, CPY2C19 and CYP17 inhibitors respectively. The majority of R^2 values are in the range of 0.55–0.85, which is comparable to the range of 0.51 to 0.88 of the conventional QSAR and QSPR studies [76,77]. These suggest that ML regression methods are capable of predicting the activity values of ADME and ADME regulatory properties at accuracy levels comparable to conventional QSAR and QSPR methods for pharmacodynamics and toxicological properties.

7. The trends in the development of machine learning models for predicting ADME and ADME regulatory properties

There are noticeable trends in the recent efforts for developing ML models to predict ADME and ADME regulatory properties. In developing ML classification models for predicting ADME and ADME regulatory properties, three ML methods support vector machines (SVM, 38 models), random forest (RF, 27 models) and k nearest neighbor (kNN, 25 models) have been more frequently used than other ML regression methods (4 models). These three methods have also been used for developing all the consensus ML models for predicting ADME and ADME regulatory properties. These three methods have been widely used because of their consistently superior performances [6,78–80], robustness in accommodating diverse structures and sample redundancy, and the lower over-fitting risks [81,82].

On the other hand, more variety of methods has been used for developing ML regression models to predict ADME and ADME regulatory properties. These are support vector machines regression (SVR, 4 models), multiple linear regression (MLR, 4 models), random forest regression (RFR, 3 models), neural network regression (NNR, 2 models), partial least-squares (PLS, 2 models) and principal component analysis (PCA, 1 model). Partly because of the limited availability of the activity

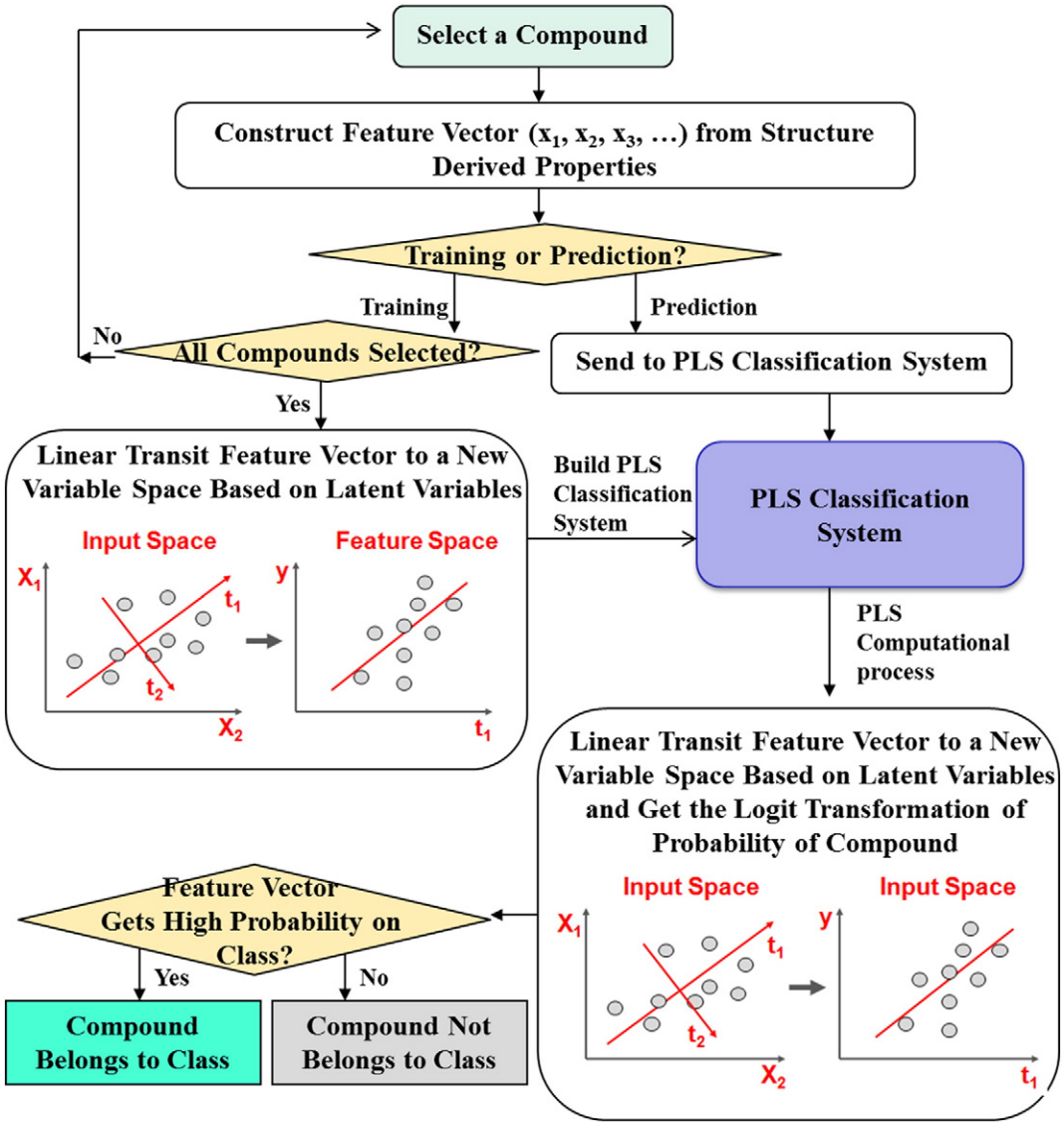

Fig. 8. Schematic diagram illustrating the processing of the prediction of compounds of a particular property from their structure by using a machine learning method — logistic partial least squares (PLS). Feature vector (x_1, x_2, x_3, …) is the same as that in Fig. 1.

data, the datasets for training ML regression models are significantly smaller in numbers (38–569 for each model, majority 38–246) and thus the less structural diversity than those for training ML classification models (53–5838 for each model, majority 100–2516). The significantly less structural diversity makes it easier to explore both the conventional (e.g. MLR and PLS) and more sophisticated ML (e.g. SVR and NNR) regression methods for developing QSAR models [80].

For better facilitating the discovery and optimization of bioactive compounds with good ADME properties, recent efforts have been directed at the more extensive exploration of ML methods with both good predictive capability and easily interpretable rules/models for the prediction of ADME and ADME regulatory properties. For instance, the recursive partitioning and naive Bayesian classifiers have been developed for identifying the important structural features necessary for classifying P-glycoprotein inhibitors and non-inhibitors [68] and for differentiating hERG potassium channel blockers and non-blockers [69], which are useful for facilitating the design of potent P-glycoprotein inhibitors and the identification of potential drug adverse reactions induced by hERG potassium channel blockages. There have also been efforts for developing more effective and robust predictive models based on combinatorial modeling strategies. In particular, the combined ML models can be developed such that multiple ML models are fused by a neural network platform, which have been shown to generate highly robust predictive models with less sensitivity on the choices of the ML models and parameters [13].

Proper representation of the structural and physicochemical features of the compounds is a key to the development of good ML regression models [40,83]. Given the structural diversity of the compounds associated with a specific ADME property, a significant number of molecular descriptors are needed to comprehensively represent their diverse structural and physicochemical properties. Indeed, the recently developed ML classification models are based on 45–650 molecular descriptors, most of which by 166–650 descriptors. Although a number of software packages and servers are available for computing a large number of molecular descriptors [18,21,22,83,84], most of the recently developed ML classification models have been generated by using Dragon, MOE, and MACCS likely due to their reputation and the reported good performances in developing predictive models for ADME as well as for pharmacodynamic and toxicological properties [78,85–87].

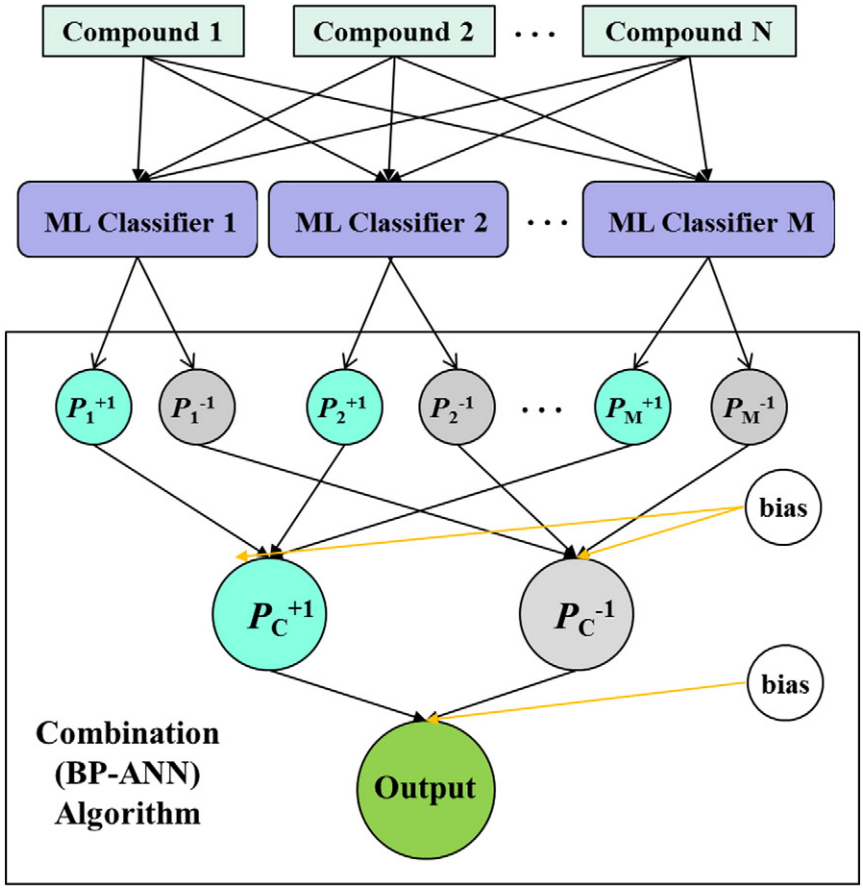

Fig. 9. Schematic diagram illustrating the processing of the prediction of compounds of a particular property from their structure by using a combined machine learning classifier approach.

8. Application scope of the developed machine learning models

The recently and previously [79] developed ML classification models broadly cover compound metabolism (by 6 different CYP enzymes) [79], efflux (by 6 different transporters) [8] and influx (by 4 different transporters) [8] at reasonably good predictive accuracies. The SEs, SPs and ACs of the majority of the ML classification models are in ranges of 74%–92%, 66%–76% and 72%–92% respectively. The SEs are close to but the SPs are substantially lower than the SEs (~90%) and SPs (~90%) of ML virtual screening models [88] that have been used to discover kinase inhibitors [89]. Therefore, these ML models have some capacity in predicting metabolism, efflux and influx properties but need to further improve the SPs to the ~90% levels for reducing false prediction rates. The R^2 values of the majority of the recently and previously [80] developed ML regression models are in the range of 0.55–0.85, which is comparable to the typical R^2 values of QSAR models used for drug lead discovery and optimization studies [90]. Therefore, the developed ML regression models are also useful for predicting the activity levels of the ADME and ADME regulatory properties. Other highly evaluated ADME properties covered by the recently and previously [80,90] developed ML models are human intestine absorption, female genital tract penetration [71] in category A and apparent volume of distribution [11], plasma protein binding [17], and blood–brain barrier crossing [62] in category D, and clearance [72] in category E.

While many ML methods are capable of predicting ADME and ADME regulatory properties at good performance levels, recent studies have suggested useful strategies for exploring their individual and collective capability for achieving better predictive performances. In the cases of the limited availability of training data and the inadequate coverage of the molecular features by the existing molecular descriptors, it is desirable to develop more robust predictive models that are less sensitive to the choices of training data, parameters and descriptors. A highly useful strategy for developing robust predictive models is to use a combination of multiple ML models fused by a neural network platform [13]. Another useful strategy is to develop ML models based on a selected set of descriptors and/or parameters most relevant to the prediction of a specific ADME or ADME regulatory property, which can be obtained by using such methods as the genetic algorithm based feature selection approach and conjugate gradient method respectively [9]. This combinatorial modeling approach is also useful for the prediction of the affinity/activity levels of the ADME and ADME regulatory properties. For instance, the combination of ML regression and ML classification models outperform several individual ML methods in the prediction of the activity of CYP2C19 inhibitors [75]. Moreover, the combination of ML models with other predictive models such as pharmacophore ensemble models also performs well in the prediction of the activity of P-glycoprotein inhibitors [73].

9. Challenges in the exploration of machine learning methods

The performance of ML methods critically depends on the diversity and representativeness of in the training datasets and the appropriate representation of their structural and physicochemical properties. The training datasets used in the most of the ML models described in Tables 2–5 are not expected to be fully representative of the compounds associated with each specific ADME property. This is particularly true for compounds not possessing a specific ADME property, which is likely an important factor for the substantially lower SPs than SEs produced by the recent ML classification models. There is a need to further expand

Table 2
Performance of ML classification methods for predicting compounds of various ADME properties. Abbreviations: BCRP — breast cancer resistance protein; MRP — multidrug resistance-associated protein; ASBT — apical sodium-dependent bile acid transporter; OATP — organic anion transporting polypeptide; OCT — organic cation transporter.

ADME class	Target	Method (testing accuracies)	Data set	Descriptors	Testing method	Ref
D	Human serum albumin	SVM (SE 81%, SP 83%)	100 substrates, 63 non-substrates	45 Dragon descriptors	Randomly divided test set	[17]
	Blood–brain barrier	SVM (SE 85%–92%, SP 70%–79%) RF (SE 85%–96%, SP 66%–81%)	246 ($f_{u,p}$, $K_{p,brain}$, & $V_{u,brain}$ are available)	196 2D, 3D descriptors	External test set of marketed CNS drugs	[62]
E	P-Glycoprotein	SVM (SE 77%, SP 74%)	99 substrates, 98 non-substrates	DragonX descriptors	Clustering derived test set	[63]
		SVM (AC 65%–72%) RF (AC 79%–79%) kNN (AC 70%–77%)	294 substrates, 250 non-substrates	286–650 Dragon or 136–148 MOE descriptors	5-Fold cross validation	[8]
		SVM (SE 61%–63%, SP 56%–66%) kNN (SE 74%–77%, SP 59%–69%) RF (SE 72%–74%, SP 68%–76%)	243 substrates, 241 non-substrates	>200 checkmol molecular fingerprints	10-Fold cross validation, and DODM derived test sets	[64]
		Consensus kNN, RF, and SVM (AC 74%–78%)	294 substrates, 250 non-substrates	286–650 Dragon or 136–148 MOE descriptors	5-Fold cross validation	[8]
	BCRP	Modified SVM GA-CG-SVM method (AC 85%)	120 substrates, 57 non-substrates	>1000 molecular descriptors	Independent validation set	[9]
		LDA (SE 70.4%, SP 76%)	262 substrates and non-substrates (human wild-type BCRP)	180 (0D–2D Dragon descriptors	Clustering derived test set	[10]
		SVM (AC 72%–86%) RF (AC 66%–85%) kNN (AC 67%–85%) Consensus kNN, RF, and SVM (AC 73%–87%)	76 substrates, 70 non-substrates	286–650 Dragon or 136–148 MOE descriptors	5-Fold cross validation	[8]
	MRP1	SVM (AC 76%–96%) RF (AC 82%–94%) kNN (AC 84%–92%) Consensus kNN, RF, and SVM (AC 89%–93%)	87 substrates, 81 non-substrates (37 assumed)	286–650 Dragon or 136–148 MOE descriptors	5-Fold cross validation	[8]
	MRP2	SVM (AC 76%–95%) RF (AC 76%–96%) kNN (AC 78%–94%) Consensus kNN, RF, and SVM (AC 79%–95%)	101 substrates, 87 non-substrates			
	MRP3	SVM (AC 86%–100%) RF (AC 82%–100%) kNN (AC 88%–100%) Consensus kNN, RF, and SVM (AC 95%–100%)	31 substrates, 31 non-substrates (22 assumed)			
	MRP4	SVM (AC 81%–94%) RF (AC 67%–95%) kNN (AC 79%–95%) Consensus kNN, RF, and SVM (AC 87%–97%)	46 substrates, 46 non-substrates (25 assumed)			
	ASBT	SVM (AC 77%–97%) RF (AC 87%–97%) kNN (AC 85%–100%) Consensus kNN, RF, and SVM (AC 89%–97%)	50 substrates, 50 non-substrates (42 assumed)			
	OATP 2B1	SVM (AC 50%–81%) RF (AC 55%–87%) kNN (AC 52%–90%) Consensus kNN, RF, and SVM (AC 61%–89%)	30 substrates, 23 non-substrates			
	OCT1	SVM (AC 73%–96%) RF (AC 65%–94%) kNN (AC 65%–96%) Consensus kNN, RF, and SVM (AC 83%–95%)	39 substrates, 39 non-substrates			
	PEPT1	SVM (AC 69%–85%) RF (AC 72%–92%) kNN (AC 70%–86%) Consensus kNN, RF, and SVM (AC 75%–93%)	79 substrates, 79 non-substrates			

the training datasets incorporating newly studied compounds and by more extensively mining the literatures and databases.

Moreover, the currently available molecular descriptors are insufficient in representing certain structural and physicochemical features, particularly the complex structural or chemical configurations [39]. Examples of the inadequately represented features are large rigid structure combined with a short flexible hydrophilic tail, and contain multi-rings with various hetero atoms such as nitrogen, oxygen, sulfur, fluorine and chlorine. Therefore, there is a need to explore new molecular descriptors. Several sets of new molecular descriptors and molecular representations have emerged in recent years [91–93]. The usefulness of these and other new descriptors for representing ADME relevant structural and physicochemical properties needs to be evaluated and explored. Another approach is to explore appropriate combination of existing molecular descriptors for expanded representation of certain ADME relevant structural and physicochemical properties.

Table 3
Performance of ML classification methods for predicting the inhibitors of drug efflux and influx transporters for regulating multi-drug resistance. DODM stands for D-optimal onion design multivariate method.

Target	Method (testing accuracies)	Data set	Descriptors	Testing method	Ref
P-Glycoprotein	Binary QSAR (SE 58%, SP 74%) SVM (SE 97%, SP 62%)	1076 inhibitors (IC_{50} ≤ 15 μM or >25–30% of inhibition), 532 non-inhibitors (IC_{50} > 100 μM or <10–12% of inhibition)	46 descriptors (11 MOE 2D, 16 MACCS fingerprints and 19 substructure fingerprints) selected by using BestFirst search algorithm	DODM derived test set	[65]
	SVM (SE 93.8%, SP 73.8%)	666 inhibitors, 609 non-inhibitors (molecular weight > 700 are excluded)	87 PreADMET molecular descriptors	10-Fold cross validation	[67]
	SVM (SE 86%–90%, SP 47%–48%)	1280 inhibitors (IC_{50} ≤ 15 μM & >25–30% of inhibition), 655 non-inhibitors (IC_{50} > 100 μM & <10–12% of inhibition)	>200 checkmol molecular fingerprints	10-Fold cross validation, and DODM derived test set	[64]
	SVM (AC 88%–93%)	743 inhibitors (IC_{50} ≤ 10 μM), 828 non-inhibitors	286–650 Dragon or 136–148 MOE descriptors after removing redundant and high correlation ones	5-Fold cross validation	[8]
	RF (SE 99%, SP 57%)	1076 inhibitors (IC_{50} ≤ 15 μM or >25–30% of inhibition), 532 non-inhibitors (IC_{50} > 100 μM or <10–12% of inhibition)	46 descriptors (11 MOE 2D, 16 MACCS fingerprints and 19 substructure fingerprints) selected by using BestFirst search algorithm	DODM derived test set	[65]
	RF (SE 84%–90%, SP 63%–65%)	1280 inhibitors (IC_{50} ≤ 15 μM & >25–30% of inhibition), 655 non-inhibitors (IC_{50} > 100 μM & <10–12% of inhibition)	>200 checkmol molecular fingerprints	10-Fold cross validation, and D DODM derived test set	[64]
	RF (AC 90%–94%)	743 inhibitors (IC_{50} ≤ 10 μM), 828 non-inhibitors	286–650 Dragon or 136–148 MOE descriptors after removing redundant and high correlation ones	5-Fold cross validation	[8]
	kNN (SE 64%, SP 72%)	1076 inhibitors (IC_{50} ≤ 15 μM or >25–30% of inhibition), 532 non-inhibitors (IC_{50} > 100 μM or <10–12% of inhibition)	46 descriptors (11 MOE 2D, 16 MACCS fingerprints and 19 substructure fingerprints) selected by using BestFirst search algorithm	DODM derived test set	[65]
	kNN (SE 86%–90%, SP 53%–58%)	1280/655 (IC_{50} ≤ 15 μM & >25–30% of inhibition as inhibitors, IC_{50} > 100 μM & <10–12% of inhibition as non-inhibitors)	>200 checkmol molecular fingerprints	10-Fold cross validation, and DODM derived test set	[64]
	kNN (AC 88%–94%)	743 inhibitors (IC_{50} ≤ 10 μM), 828 non-inhibitors	286–650 Dragon or 136–148 MOE descriptors after removing redundant and high correlation ones	5-Fold cross validation	[8]
	RP (SE 83.5%, SP 67.0%) Naïve Bayesian (SE 83.5%, SP 73.2%, AC ~80%)	797/476 (MDRR ratio > 5.0 as inhibitor, <4.0 as non-inhibitor)	13 molecular descriptors, FCFP, ECFP, LCFP, FPFP, EPFP, and LEFP fingerprints	300 external test set	[68]
	SVM + Docking (SE 86.7%, SP 91.2%)	666/609 (molecular weight > 700 are excluded)	87 PreADMET molecular descriptors	10-Fold cross validation	[67]
	Consensus kNN, RF and SVM (AC 62%–78%)	743 inhibitors (IC_{50} ≤ 10 μM), 828 non-inhibitors	286–650 Dragon or 136–148 MOE descriptors after removing redundant and high correlation ones	5-Fold cross validation	[8]
BCRP	SVM (AC 75%–85%) RF (AC 75%–83%) kNN (AC 71%–85%) Consensus kNN, RF, and SVM (AC 79%–87%)	167 inhibitors (IC_{50} ≤ 10 μM), 215 non-inhibitors	286–650 Dragon or 136–148 MOE descriptors after removing redundant and high correlation ones	5-Fold cross validation	[8]
MRP1	SVM (AC 78%–90%) RF (AC 83%–91%) kNN (AC 81%–91%) Consensus kNN, RF, and SVM (AC 86%–94%)	224 inhibitors (IC_{50} ≤ 10 μM), 194 non-inhibitors			
MRP2	SVM (AC 78%–96%) RF (AC 81%–97%) kNN (AC 73%–95%) Consensus kNN, RF, and SVM (AC 80%–98%)	48 inhibitors (IC_{50} ≤ 10 μM), 48 non-inhibitors			
MRP4	SVM (AC 48%–78%) RF (AC 52%–72%) kNN (AC 48%–74%) Consensus kNN, RF, and SVM (62%–78%)	32 inhibitors (IC_{50} ≤ 10 μM), 32 non-inhibitors			
ASBT	SVM (AC 80%–92%) RF (AC 82%–94%) kNN (AC 82%–94%) Consensus kNN, RF, and SVM (AC 87%–97%)	75 inhibitors (IC_{50} ≤ 10 μM), 75 non-inhibitors			
MCT1	SVM (AC 100%) RF (AC 100%) kNN (AC 100%) Consensus kNN, RF, and SVM (AC 100%)	47 inhibitors (IC_{50} ≤ 10 μM), 20 non-inhibitors			

Table 3 (*continued*)

Target	Method (testing accuracies)	Data set	Descriptors	Testing method	Ref
OATP 2B1	SVM (AC 66%–86%) RF (AC 67%–89%) kNN (AC 70%–88%) Consensus kNN, RF, and SVM (AC 74%–86%)	70 inhibitors (IC_{50} ≤ 100 μM), 66 non-inhibitors			
OCT1	SVM (AC 76%–95%) RF (AC 83%–97%) kNN (AC 81%–95%) Consensus kNN, RF, and SVM (AC 86%–98%)	87 inhibitors (IC_{50} ≤ 100 μM), 112 non-inhibitors			
PEPT1	SVM (AC 61%–83%) RF (AC 47%–77%) kNN (AC 45%–84%) Consensus kNN, RF, and SVM (AC 57%–87%)	40 inhibitors (IC_{50} ≤ 100 μM), 40 non-inhibitors			
hERG	RP (SE 86.9%, SP 76.3%) Naïve Bayesian (AC 89.4% for WOMABAT-PK, AC 86.1% for PubChem)	806 (IC_{50} in 1, 5, 10, 20, 30, and 40 μM as threshold range to identify the blockers and non-blockers	14 molecular descriptors, FCFP, ECFP, LCFP, FPFP, EPFP, and LEFP fingerprints	120 test molecules 66 external molecules from WOMBAT-PK and 1953 external molecules from PubChem	[69]

However, excessive use of existing descriptors, which are substantially overlapping and redundant, may introduce noise as well as the expanded representation of ADME relevant structural and physicochemical properties. Hence it may be more desirable to introduce new descriptors for more appropriately representing these structural and physicochemical properties.

10. Perspectives

Both classification-based and regression-based ML methods have consistently shown promising capability in predicting a variety of ADME and ADME regulatory properties for diverse ranges of structures at accuracy levels comparable to those practically used in drug lead

Table 4
Performance of ML classification methods for predicting the inhibitors of drug metabolism enzymes for assessing drug–drug interactions.

Target	Method (testing accuracies)	Data SET	Descriptors	Testing method	Ref
CYP 1A2, 2C9, 2C19, 2D6, 3A4	SVM (SE 75%–87%, SP 83%–88%)	17,143 compounds tested by HTS	Bioclipse molecular signatures	Unspecified	[15]
	SVM (SE 36.9%–79.4%, SP 80.2%–96.4%, AC 76.2%–84.3%) Naïve Bayes (SE 48.5%–76.9%, SP 60.0%–86.1%, AC 63.7%–78.2%) kNN (SE 51.0%–84.7%, SP 64.6%–88.7%, AC 68.3%–80.7% C4.5 DT (SE 46.2%–74.2%, SP 75.5%–88.3%, AC 70.5%–79.4% SVM + kNN (SE SE 39.0%–80.0%, SP 80.7%–95.6%, AC 75.1%–83.8%) SVM + C4.5 DT (SE 39.6%–79.2%, SP 80. 2%–95.5%, AC 75.1%–83.7%)	2516–5663 inhibitors (AC_{50} ≤ 10 μM), 6436–9365 non-inhibitors (AC_{50} > 57 μM)	166 MACCS and 307 FP4 molecular fingerprints	Independent set of 544–2070 inhibitors and 1052–4955 non-inhibitors from PubChem BioAssay	[13]
	SVM (AC 82.5%–88.3%, AUC 0.88–0.94) RF (AC 87.5%–88.6%, AUC 0.93–0.95) kNN (AC 79.5%–80.4%, AUC 0.865–0.868)	2545–5838 inhibitors (AC_{50} ≤ 10 μM as inhibitors)	Bioclipse molecular signatures	External set of 12,634–13,276 inhibitors and non-inhibitors	[16]
	SVM (AC 80.6%–89.5%, AUC 0.87–0.93)	17,143 compounds tested by HTS (with class 1.1, 1.2, 2.1 as active, with class 4 curves as inactive)	264 molecular descriptors	7-Fold cross validation	[14]
CYP 3A4	RF (SE 76.8%, SP 86.0%) SVM (SE 79.7%, SP 80.1%) logistic PLS (SE 60.8%, SP 84.4%)	4605 inhibitors (IC_{50} < 40 μM), 12,394 non-inhibitors (IC_{50} > 60 μM)	379 fragmental descriptors	External set of 8528 compounds with experimental CYP3A6 inhibition activity	[66]
CYP 2D6, 3A4	SVM (SE 26.0%–39.0%, SP 74.0%–85.0%) Multiple LDA (SE 41.9%–54.7%, SP 64.5%–68.5%)	160–216 inhibitors, 386–442 non-inhibitors	353 mold molecular descriptors ^{13}C- and ^{15}N-NMR spectra	10-Fold cross validation	[70]

Table 5
Performance of ML regression methods for predicting the activity level of ADME and ADME regulatory properties.

ADME class	Target	Action on target	Method (testing accuracies)	Data Set	Descriptors	Testing method	Ref
A	Female genital tract	Penetrator	KNNR (R^2 ~ 0.4) RFR (R^2 ~ 0.4)	38 penetrators & 20 poor-penetrators (TPR in the 0.00–0.49 range as poor penetrators, 0.50–1.49 as good, and ≥1.5 as excellent)	Dragon & SiRMS descriptors	5-Fold cross validation	[71]
D	Apparent volume of distribution	Human VDss value	MLR (R^2 0.74) SVR (Q^2 0.55)	569 compounds with VDss value	89 molecular descriptors	Randomly splited	[11]
			ANN (R^2 0.62, RMSE 0.32) MLR (R^2 0.56, RMSE 0.32) SVR (R^2 0.58, RMSE 0.31)	121 (VD_{ss} in the range of 0.1 to 21 L/kg)	7 E-Dragon descriptors	10-Fold cross validation	[12]
D	Blood–brain barrier	Penetration	RFR (R^2 0.42–0.69 RMSE 0.32–0.59) SVR (R^2 0.41–0.6 RMSE 0.39–0.58)	246 ($f_{u,p}$, $K_{p,brain}$, & $V_{u,brain}$ were available)	196 2&3D molecular descriptors	Randomly splited	[62]
E	Intrinsic clearance	CLint values	Orthogonal PLS (R^2 0.59, Q^2 0.48) Principals (R^2 0.43, Q^2 0.35) RFR (R^2 0.96, Q^2 0.48)	244 (CL_{int} in 1–1,400,000 mL/min range, <1500 mL/min as stable)	93 Selma molecular descriptors	7-Fold cross validation	[72]
E	Systemic clearance	Human iv CL value	MLR (R^2 0.64)	525 compounds with iv CL value	89 molecular descriptors	Randomly splited	[11]
ER	P-glycoprotein	Inhibitor	SVR (R^2 0.87, RMSE 0.39)	180 compounds with activity value	HypoGen pharmacophore ensemble	Deliberately divided	[73]
ER	P-glycoprotein	Inhibitor	BP-ANN (R^2 0.81–0.87) MLR (R^2 0.69)	88 (40 flavones, 1 isoflavone, 22 chalcones, 5 silybins, 14 aurones, and 6 xanthones)	118 PaDEL descriptors	Randomly splited	[74]
MR	CYP 1A2, 2C8, 2C9, 2A6, 2C19, 2D6, 3A4, 17	Inhibitor	PLSR (R^2 0.65–0.99)	54–209 compounds with activity value	6–15 E-DRAGON descriptors	Unspecified	[75]

discovery and optimization, making the developed ADME and ADME regulatory prediction models potentially useful tools for assessing ADME properties and predicting ADME regulatory properties. In spite of the significant efforts, the recently developed ML models only cover a limited variety of ADME and ADME regulatory properties. There is a need to develop the in-silico prediction tools for covering more diverse range of ADME and ADME regulatory properties.

The application potential of these ML models is constrained by the limited knowledge and information about the compounds associated with certain ADME or ADME regulatory properties. Existing molecular descriptors are not fully representative of some of the structural and physiochemical properties. Efforts are needed to further expand the knowledge and coverage of ADME and ADME regulatory properties and the associated compounds, and to develop and select more appropriate sets of molecular descriptors for representing the structural and physicochemical properties relevant to ADME and ADME regulatory properties, so as to develop these ML models into useful tools for facilitating the in-silico assessment ADME and ADME regulatory properties.

Acknowledgements

We acknowledge the support by Major State Basic Research Development Program of China 2013CB967204 and Singapore Academic Research Fund R148000181112.

References

[1] J. Drews, Drug discovery: a historical perspective, Science 287 (2000) 1960–1964.
[2] S. Ekins, B.J. Ring, J. Grace, D.J. McRobie-Belle, S.A. Wrighton, Present and future in vitro approaches for drug metabolism, J. Pharmacol. Toxicol. Methods 44 (2000) 313–324.
[3] R.E. White, High-throughput screening in drug metabolism and pharmacokinetic support of drug discovery, Annu. Rev. Pharmacol. Toxicol. 40 (2000) 133–157.
[4] H. van de Waterbeemd, E. Gifford, ADMET in silico modelling: towards prediction paradise? Nat. Rev. Drug Discov. 2 (2003) 192–204.
[5] D. Stepensky, Prediction of drug disposition on the basis of its chemical structure, Clin. Pharmacokinet. 52 (2013) 415–431.
[6] M. Trotter, S. Holden, Support vector machines for ADME property classification, QSAR Comb. Sci. 22 (2003) 533–548.
[7] Y. Sakiyama, The use of machine learning and nonlinear statistical tools for ADME prediction, Expert Opin. Drug Metab. Toxicol. 5 (2009) 149–169.
[8] A. Sedykh, D. Fourches, J. Duan, O. Hucke, M. Garneau, H. Zhu, P. Bonneau, A. Tropsha, Human intestinal transporter database: QSAR modeling and virtual profiling of drug uptake, efflux and interactions, Pharm. Res. 30 (2013) 996–1007.
[9] L. Zhong, C.Y. Ma, H. Zhang, L.J. Yang, H.L. Wan, Q.Q. Xie, L.L. Li, S.Y. Yang, A prediction model of substrates and non-substrates of breast cancer resistance protein (BCRP) developed by GA-CG-SVM method, Comput. Biol. Med. 41 (2011) 1006–1013.
[10] M.E. Gantner, M.E. Di Ianni, M.E. Ruiz, A. Talevi, L.E. Bruno-Blanch, Development of conformation independent computational models for the early recognition of breast cancer resistance protein substrates, Biomed Res. Int. 2013 (2013) 863592.
[11] V.K. Gombar, S.D. Hall, Quantitative structure–activity relationship models of clinical pharmacokinetics: clearance and volume of distribution, J. Chem. Inf. Model. 53 (2013) 948–957.
[12] B. Louis, V.K. Agrawal, Prediction of human volume of distribution values for drugs using linear and nonlinear quantitative structure pharmacokinetic relationship models, Interdiscip. Sci. 6 (2014) 71–83.
[13] F. Cheng, Y. Yu, J. Shen, L. Yang, W. Li, G. Liu, P.W. Lee, Y. Tang, Classification of cytochrome P450 inhibitors and noninhibitors using combined classifiers, J. Chem. Inf. Model. 51 (2011) 996–1011.
[14] H. Sun, H. Veith, M. Xia, C.P. Austin, R. Huang, Predictive models for cytochrome p450 isozymes based on quantitative high throughput screening data, J. Chem. Inf. Model. 51 (2011) 2474–2481.
[15] M. Rostkowski, O. Spjuth, P. Rydberg, WhichCyp: prediction of cytochromes P450 inhibition, Bioinformatics 29 (2013) 2051–2052.
[16] M. Lapins, A. Worachartcheewan, O. Spjuth, V. Georgiev, V. Prachayasittikul, C. Nantasenamat, J.E. Wikberg, A unified proteochemometric model for prediction of inhibition of cytochrome p450 isoforms, PLoS One 8 (2013) e66566.
[17] F. Zsila, Z. Bikadi, D. Malik, P. Hari, I. Pechan, A. Berces, E. Hazai, Evaluation of drug–human serum albumin binding interactions with support vector machine aided online automated docking, Bioinformatics 27 (2011) 1806–1813.
[18] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, DRAGON, 2005.
[19] I.V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V.A. Palyulin, E.V. Radchenko, N.S. Zefirov, A.S. Makarenko, V.Y. Tanchuk, V.V. Prokopenko, Virtual computational chemistry laboratory — design and description, J. Comput. Aided Mol. Des. 19 (2005) 453–463.
[20] L.H. Hall, G.E. Kellogg, D.N. Haney, Molconn-Z, in, eduSoft, LC, 2002.
[21] J.K. Wegner, JOELib/JOELib2, 2005.
[22] Z.R. Li, L.Y. Han, Y. Xue, C.W. Yap, H. Li, L. Jiang, Y.Z. Chen, MODEL-molecular descriptor lab: a web-based server for computing structural and physicochemical features of compounds, Biotechnol. Bioeng. 97 (2007) 389–396.
[23] C.W. Yap, PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints, J. Comput. Chem. 32 (2011) 1466–1474.
[24] M.C. Hemmer, V. Steinhauer, J. Gasteiger, Deriving the 3D structure of organic molecules from their infrared spectra, Vib. Spectrosc. 19 (1999) 151–164.

[25] G. Rücker, C. Rücker, Counts of all walks as atomic and molecular descriptors, J. Chem. Inf. Comput. Sci. 33 (1993) 683–695.
[26] J.H. Schuur, P. Setzer, J. Gasteiger, The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure–spectra correlations and studies of biological activity, J. Chem. Inf. Comput. Sci. 36 (1996) 334–344.
[27] R.S. Pearlman, K.M. Smith, Metric validation and the receptor-relevant subspace concept, J. Chem. Inf. Comput. Sci. 39 (1999) 28–35.
[28] G. Bravi, E. Gancia, P. Mascagni, M. Pegna, R. Todeschini, A. Zaliani, MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: a comparative 3D QSAR study in a series of steroids, J. Comput. Aided Mol. Des. 11 (1997) 79–92.
[29] J. Galvez, R. Garcia, M.T. Salabert, R. Soler, Charge indexes. New topological descriptors, J. Chem. Inf. Comput. Sci. 34 (1994) 520–525.
[30] V. Consonni, R. Todeschini, M. Pavan, Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors, J. Chem. Inf. Comput. Sci. 42 (2002) 682–692.
[31] M. Randic, Graph theoretical approach to local and overall aromaticity of benzenoid hydrocarbons, Tetrahedron 31 (1975) 1477–1481.
[32] M. Randic, Molecular profiles. Novel geometry-dependent molecular descriptors, New J. Chem. 19 (1995) 781–791.
[33] L.B. Kier, L.H. Hall, Molecular Structure Description: The Electrotopological State, Academic Press, San Diego, 1999.
[34] J.A. Platts, D. Butina, M.H. Abraham, A. Hersey, Estimation of molecular free energy relation descriptors using a group contribution approach, J. Chem. Inf. Comput. Sci. 39 (1999) 835–845.
[35] M.S. Meskin, E.J. Lien, QSAR analysis of drug excretion into human breast milk, J. Clin. Hosp. Pharm. 10 (1985) 269–278.
[36] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (2002) 389–422.
[37] C.B. Lucasius, G. Kateman, Understanding and using genetic algorithms. Part 1. Concepts, properties and context, Chemom. Intell. Lab. Syst. 19 (1993) 1–33.
[38] J.M. Sutter, J.H. Kalivas, Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection, Microchem. J. 47 (1993) 60–66.
[39] Y. Xue, C.W. Yap, L.Z. Sun, Z.W. Cao, J.F. Wang, Y.Z. Chen, Prediction of P-glycoprotein substrates by a support vector machine approach, J. Chem. Inf. Comput. Sci. 44 (2004) 1497–1505.
[40] H. Li, C.W. Yap, C.Y. Ung, Y. Xue, Z.W. Cao, Y.Z. Chen, Effect of selection of molecular descriptors on the prediction of blood–brain barrier penetrating and nonpenetrating agents by statistical learning methods, J. Chem. Inf. Model. 45 (2005) 1376–1384.
[41] M. Iyer, R. Mishra, Y. Han, A.J. Hopfinger, Predicting blood–brain barrier partitioning of organic molecules using membrane-interaction QSAR analysis, Pharm. Res. 19 (2002) 1611–1621.
[42] P. Gramatica, P. Pilutti, E. Papa, Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting into training-test sets and consensus modeling, J. Chem. Inf. Comput. Sci. 44 (2004) 1794–1802.
[43] S. Izrailev, D.K. Agrafiotis, A method for quantifying and visualizing the diversity of QSAR models, J. Mol. Graph. Model. 22 (2004) 275–284.
[44] C.W. Yap, Y.Z. Chen, Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines, J. Chem. Inf. Model. 45 (2005) 982–992.
[45] H. Li, C.W. Yap, C.Y. Ung, Y. Xue, Z.W. Cao, Y.Z. Chen, Effect of selection of molecular descriptors on the prediction of blood–brain barrier penetrating and non-penetrating agents by statistical learning methods, J. Chem. Inf. Model. 45 (2005) 1376–1384.
[46] J. Serra, E. Thompson, P. Jurs, Development of binary classification of structural chromosome aberrations for a diverse set of organic compounds from molecular structure, Chem. Res. Toxicol. 16 (2003) 153–163.
[47] C.W. Yap, Y.Z. Chen, Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines, J. Chem. Inf. Model. 45 (2005) 982–992.
[48] C.J. Huberty, Applied Discriminant Analysis, John Wiley & Sons, New York, 1994.
[49] E. Fix, J.L. Hodges, Discriminatory Analysis: Non-parametric Discrimination: Consistency Properties, USAF School of Aviation Medicine, Texas, 1951.
[50] I. Aleksander, H. Morton, An Introduction to Neural Computing, 2nd ed International Thomson Computer Press, London, 1995.
[51] D.F. Specht, Probabilistic neural networks, Neural Netw. 3 (1990) 109–118.
[52] E. Parzen, On estimation of a probability density function and mode, Ann. Math. Stat. 33 (1962) 1065–1076.
[53] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[54] A.J. Smola, B. Scholkopf, A tutorial on support vector regression, in: NeuroCOLT2 Technical Report Series, 1998.
[55] Z. Yuan, B.X. Huang, Prediction of protein accessible surface areas by support vector regression, Proteins 57 (2004) 558–564.
[56] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, Calif, 1993.
[57] P. Domingos, M. Pazzani, Beyond independence: conditions for the optimality of the simple Bayesian classifier, in: L. Saitta (Ed.), Proceedings of the Thirteenth International Conference on Machine Learning, Morgan Kaufmann, San Francisco 1996, pp. 105–112.
[58] P. Allison, Multiple Regression, Pine Forge Press, Thousand Oaks, CA, 1999.
[59] J.G. Topliss, R.P. Edwards, Chance factors in studies of quantitative structure–activity relationships, J. Med. Chem. 22 (1979) 1238–1244.
[60] L. Eriksson, J. Jaworska, M. Cronin, A. Worth, P. Gramatica, R. McDowell, Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs, Environ. Health Perspect. 111 (2003) 1361–1375.
[61] S. Wold, M. Sjostrom, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.
[62] H. Chen, S. Winiwarter, M. Friden, M. Antonsson, O. Engkvist, In silico prediction of unbound brain-to-plasma concentration ratio using machine learning algorithms, J. Mol. Graph. Model. 29 (2011) 985–995.
[63] Z. Bikadi, I. Hazai, D. Malik, K. Jemnitz, Z. Veres, P. Hari, Z. Ni, T.W. Loo, D.M. Clarke, E. Hazai, Q. Mao, Predicting P-glycoprotein-mediated drug transport based on support vector machine and three-dimensional crystal structure of P-glycoprotein, PLoS One 6 (2011) e25815.
[64] V. Poongavanam, N. Haider, G.F. Ecker, Fingerprint-based in silico models for the prediction of P-glycoprotein substrates and inhibitors, Bioorg. Med. Chem. 20 (2012) 5388–5395.
[65] F. Klepsch, P. Vasanthanathan, G.F. Ecker, Ligand and structure-based classification models for prediction of P-glycoprotein inhibitors, J. Chem. Inf. Model. 54 (2014) 218–229.
[66] R. Didziapetris, J. Dapkunas, A. Sazonovas, P. Japertas, Trainable structure–activity relationship model for virtual screening of CYP3A4 inhibition, J. Comput. Aided Mol. Des. 24 (2010) 891–906.
[67] W. Tan, H. Mei, L. Chao, T. Liu, X. Pan, M. Shu, L. Yang, Combined QSAR and molecule docking studies on predicting P-glycoprotein inhibitors, J. Comput. Aided Mol. Des. 27 (2013) 1067–1073.
[68] L. Chen, Y. Li, Q. Zhao, H. Peng, T. Hou, ADME evaluation in drug discovery. 10. Predictions of P-glycoprotein inhibitors using recursive partitioning and naive Bayesian classification techniques, Mol. Pharm. 8 (2011) 889–900.
[69] S. Wang, Y. Li, J. Wang, L. Chen, L. Zhang, H. Yu, T. Hou, ADMET evaluation in drug discovery. 12. Development of binary classification models for prediction of hERG potassium channel blockage, Mol. Pharm. 9 (2012) 996–1010.
[70] B. McPhail, Y. Tie, H. Hong, B.A. Pearce, L.K. Schnackenberg, W. Ge, L.G. Valerio, J.C. Fuscoe, W. Tong, D.A. Buzatu, J.G. Wilkes, B.A. Fowler, E. Demchuk, R.D. Beger, Modeling chemical interaction profiles: I. Spectral data–activity relationship and structure–activity relationship models for inhibitors and non-inhibitors of cytochrome P450 CYP3A4 and CYP2D6 isozymes, Molecules 17 (2012) 3383–3406.
[71] C.G. Thompson, A. Sedykh, M.R. Nicol, E. Muratov, D. Fourches, A. Tropsha, A.D. Kashuba, Short communication: cheminformatics analysis to identify predictors of antiviral drug penetration into the female genital tract, AIDS Res. Hum. Retrovir. 30 (2014) 1058–1064.
[72] Y.W. Hsiao, U. Fagerholm, U. Norinder, In silico categorization of in vivo intrinsic clearance using machine learning, Mol. Pharm. 10 (2013) 1318–1321.
[73] M.K. Leong, H.B. Chen, Y.H. Shih, Prediction of promiscuous p-glycoprotein inhibition using a novel machine learning scheme, PLoS One 7 (2012) e33829.
[74] J. Shen, Y. Cui, J. Gu, Y. Li, L. Li, A genetic algorithm-back propagation artificial neural network model to quantify the affinity of flavonoids toward P-glycoprotein, Comb. Chem. High Throughput Screen. 17 (2014) 162–172.
[75] O. Dagliyan, I.H. Kavakli, M. Turkay, Classification of cytochrome P450 inhibitors with respect to binding free energy and pIC50 using common molecular descriptors, J. Chem. Inf. Model. 49 (2009) 2403–2411.
[76] M. Lobell, V. Sivarajah, In silico prediction of aqueous solubility, human plasma protein binding and volume of distribution of compounds from calculated pKa and AlogP98 values, Mol. Divers. 7 (2003) 69–87.
[77] T. Hou, X. Xu, ADME evaluation in drug discovery. 3. Modeling blood–brain barrier partitioning using simple molecular descriptors, J. Chem. Inf. Comput. Sci. 43 (2003) 2137–2152.
[78] E. Byvatov, U. Fechner, J. Sadowski, G. Schneider, Comparison of support vector machine and artificial neural network systems for drug/nondrug classification, J. Chem. Inf. Comput. Sci. 43 (2003) 1882–1889.
[79] C.W. Yap, Y. Xue, H. Li, Z.R. Li, C.Y. Ung, L.Y. Han, C.J. Zheng, Z.W. Cao, Y.Z. Chen, Prediction of compounds with specific pharmacodynamic, pharmacokinetic or toxicological property by statistical learning methods, Mini Rev. Med. Chem. 6 (2006) 449–459.
[80] C.W. Yap, H. Li, Z.L. Ji, Y.Z. Chen, Regression methods for developing QSAR and QSPR models to predict compounds of specific pharmacodynamic, pharmacokinetic and toxicological properties, Mini Rev. Med. Chem. 7 (2007) 1097–1107.
[81] N. Pochet, F. De Smet, J.A. Suykens, B.L. De Moor, Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction, Bioinformatics 20 (2004) 3185–3195.
[82] F. Li, Y. Yang, Analysis of recursive gene selection approaches from microarray data, Bioinformatics 21 (2005) 3741–3747.
[83] Y. Xue, Z.R. Li, C.W. Yap, L.Z. Sun, X. Chen, Y.Z. Chen, Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents, J. Chem. Inf. Comput. Sci. 44 (2004) 1630–1638.
[84] L. Hall, G. Kellogg, D. Haney, Molconn-Z, 2002.
[85] A.K. Saxena, P. Prathipati, Comparison of MLR, PLS and GA-MLR in QSAR analysis, SAR QSAR Environ. Res. 14 (2003) 433–445.
[86] A. Koutsoukas, S. Paricharak, W.R. Galloway, D.R. Spring, A.P. Ijzerman, R.C. Glen, D. Marcus, A. Bender, How diverse are diversity assessment methods? A comparative analysis and benchmarking of molecular descriptor space, J. Chem. Inf. Model. 54 (2014) 230–242.
[87] J.K. Wegner, H. Frohlich, A. Zell, Feature selection for descriptor based classification models. 2. Human intestinal absorption (HIA), J. Chem. Inf. Comput. Sci. 44 (2004) 931–939.
[88] X.H. Liu, X.H. Ma, C.Y. Tan, Y.Y. Jiang, M.L. Go, B.C. Low, Y.Z. Chen, Virtual screening of Abl inhibitors from large compound libraries by support vector machines, J. Chem. Inf. Model. 49 (2009) 2101–2110.
[89] C. Zhang, C. Tan, X. Zu, X. Zhai, F. Liu, B. Chu, X. Ma, Y. Chen, P. Gong, Y. Jiang, Exploration of (S)-3-aminopyrrolidine as a potentially interesting scaffold for discovery of novel Abl and PI3K dual inhibitors, Eur. J. Med. Chem. 46 (2011) 1404–1414.

[90] M. Shen, C. Beguin, A. Golbraikh, J.P. Stables, H. Kohn, A. Tropsha, Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds, J. Med. Chem. 47 (2004) 2356–2364.
[91] L. Xue, J.W. Godden, F.L. Stahura, J. Bajorath, Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme, J. Chem. Inf. Comput. Sci. 43 (2003) 1151–1157.
[92] M. Posa, Heuman indices of hydrophobicity of bile acids and their comparison with a newly developed and conventional molecular descriptors, Biochimie 97 (2014) 28–38.
[93] F. Berenger, A. Voet, X.Y. Lee, K.Y. Zhang, A rotation-translation invariant molecular descriptor of partial charges and its use in ligand-based virtual screening, J. Cheminform. 6 (2014) 23.