J Comput Aided Mol Des (2008) 22:843–855
DOI 10.1007/s10822-008-9225-4

Considerations and recent advances in QSAR models for cytochrome P450-mediated drug metabolism prediction

Haiyan Li · Jin Sun · Xiaowen Fan · Xiaofan Sui · Lan Zhang · Yongjun Wang · Zhonggui He

Received: 19 December 2007 / Accepted: 8 June 2008 / Published online: 24 June 2008

Abstract Quantitative structure–activity relationships (QSAR) methods are urgently needed for predicting ADME/T (absorption, distribution, metabolism, excretion and toxicity) properties to select lead compounds for optimization at the early stage of drug discovery, and to screen drug candidates for clinical trials. Use of suitable QSAR models ultimately results in lesser time-cost and lower attrition rate during drug discovery and development. In the case of ADME/T parameters, drug metabolism is a key determinant of metabolic stability, drug–drug interactions, and drug toxicity. QSAR models for predicting drug metabolism have undergone significant advances recently. However, most of the models used lack sufficient interpretability and offer poor predictability for novel drugs. In this review, we describe some considerations to be taken into account by QSAR for modeling drug metabolism, such as the accuracy/consistency of the entire data set, representation and diversity of the training and test sets, and variable selection. We also describe some novel statistical techniques (ensemble methods, multivariate adaptive regression splines and graph machines), which are not yet used frequently to develop QSAR models for drug metabolism. Subsequently, rational recommendations for developing predictable and interpretable QSAR models are made. Finally, the recent advances in QSAR models for cytochrome P450-mediated drug metabolism prediction, including in vivo hepatic clearance, in vitro metabolic stability, inhibitors and substrates of cytochrome P450 families, are briefly summarized.

Keywords QSAR · Drug metabolism prediction · Dataset diversity · Variable selection · Random Forest · Multivariate adaptive regression splines · Graph machine

H. Li · J. Sun (✉) · X. Fan · X. Sui · L. Zhang · Y. Wang · Z. He (✉)
Department of Biopharmaceutics, School of Pharmacy, Shenyang Pharmaceutical University, No. 59 Mailbox, No. 103 of Wenhua Road, Shenyang 110016, China
e-mail: sunjin66@21cn.com

Z. He
e-mail: hezhonggui@gmail.com

Background

The early prediction of drug ADME/T properties is of great importance to shorten the time and increase the chances of success during drug discovery and development [1, 2]. Due to the great efforts made by researchers and the pharmaceutical industry involving in-depth study of pharmacokinetic performance between 1991 and 2000, the percentage of drug development failures due to pharmacokinetic and bioavailability problems has fallen markedly from 40% (in 1991) to 10% (in 2000) [3]. In 2005, three main reasons for terminating development of new drugs were safety, efficacy and cost [4]. However, the combination of activity, safety, and acceptable ADME/T properties, rather than one specific factor, will dictate the overall success of drug discovery and development [4].

Among the ADME/T properties, drug metabolism is a key determinant of several important drug processes in vivo, such as metabolic stability, drug–drug interactions and drug toxicity [5]. However, prediction models of drug metabolism are the most difficult to establish, because of the complexity of biochemical reactions catalyzed by different metabolic enzymes, the extensive range of binding sites, mechanisms of drug biotransformation, and the influence of transporters [6–8].

In recent years, lots of in silico models have been applied to predict drug metabolism and these are mainly structure-based approaches and quantitative structure–activity relationship (QSAR) methods [9]. The structure-based approaches, such as 3D molecular modeling [10–18], quantum mechanical methods [19–22] and pharmacophore modeling [23–25], are generally limited by the requirement for experimental X-ray structures of cytochrome P450 families (CYPs) and the assumption of a unique mode for substrate binding. In these cases, QSAR can be a useful tool to predict interactions between drugs and metabolic enzymes and to identify molecules that may give rise to problems in their metabolism because of their chemical structure [26].

Based on different endpoints, significant progress has been made in QSAR models to predict drug metabolism using various statistical techniques, such as multiple linear regression (MLR), partial least squares (PLS), naïve bayse classifier (NBC), *k*-nearest neighbor (*k*-NN), self-organizing map (SOM), recursive partition (RP), artificial neutral network (ANN), and support vector machine (SVM). However, most of these QSAR models have poor predictability for novel drugs and cannot generally be applied [27]. One of the reasons is inadequate consideration of the accuracy/consistency, and the representation and diversity of the data sets (training set and test set) [28]. In addition, most of the statistical techniques used are difficult to interpret, particularly the effect of certain molecular properties on drug metabolism.

So, the purpose of this review is firstly to guide the reader through the considerations in building QSAR models for drug metabolism, including the accuracy/consistency of the entire data set, representation and diversity of training and test set, and variable selection. Secondly, some novel statistical techniques (ensemble methods, MARS (multivariate adaptive regression splines) and graph machines), less frequently used for developing QSAR models in drug metabolism will be introduced. Subsequently, rational recommendations for developing QSAR models in drug metabolism will be provided. Finally, recent advances in QSAR models for cytochrome P450-mediated drug metabolism prediction, including the models for in vivo hepatic clearance and in vitro metabolic stability, inhibitors and substrates of CYPs, will be briefly summarized.

Considerations in QSAR models for drug metabolism prediction

As shown in Fig. 1 (bold words), the QSAR models used in predicting drug metabolism can be divided into four steps, i.e., determination or collection of metabolism parameters of interest, molecular descriptor generation and variable selection to extract desirable independent variables, model construction and evaluation with training and test sets using linear or non-linear statistical methods and, finally, prediction of the metabolism of new compounds using an

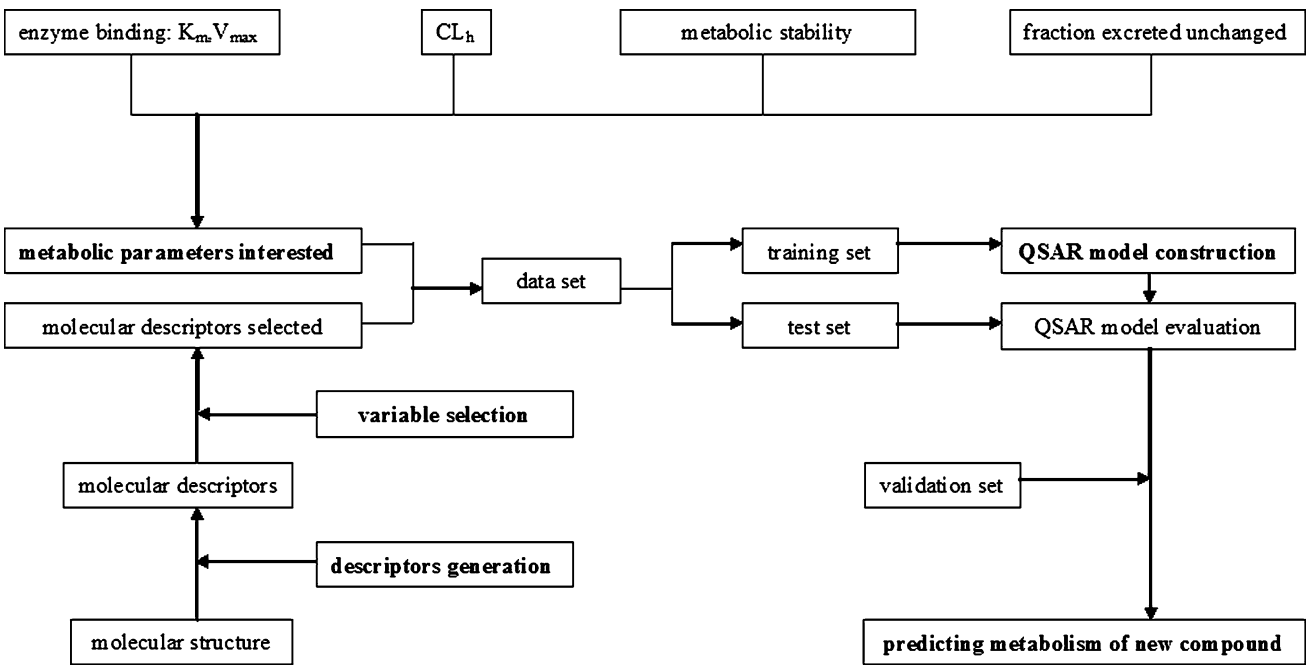

Fig. 1 Scheme of construction of QSAR models for drug metabolism prediction

external validation set. The first three will now be described in detail.

Determination or collection of the metabolism parameters of interest

Data collection: which end point to select and its accuracy/consistency?

It is essential to choose the type of end point available to construct the QSAR models for drug metabolism prediction [29]. Unlike the prediction of absorption and toxicity, for which the end point is relatively easy to select, such as the fraction absorbed from the intestine (or membrane permeability coefficient) and LD_{50} (medial lethal dose), there is no straightforward assay process for metabolism [30]. Basically, for the categorical model, outputs may be inhibitors (or non-inhibitors) and substrates (or non-substrates); while the IC_{50} (50% inhibiting concentration), V_{max} (maximum rate of metabolism), hepatic metabolic clearance (CL_h) and $CL_{int,\ in\ vitro}$ (in vitro intrinsic clearance) can be used as end points for a quantitative metabolism model.

Furthermore, the accuracy and consistency of the metabolic parameters are the key to the predictability of the model. Different results may be obtained from different enzyme sources (recombinant CYP enzymes, human liver microsomes or human hepatocytes), different experimental conditions (type of probe substrate, analysis method) or different laboratories, as described below.

For CYPs inhibition studies using different enzyme sources, Di Marco et al. observed higher IC_{50} values in rat microsomes compared with hepatocytes for a range of CYP2D inhibitors, and speculated that cell-associated inhibitor concentrations were higher [31]. After that, McGinnity et al. found the $IC_{50,\ apparent}$ values of CYP2C9 generated in human hepatocytes were systematically higher than those determined with rP450s (recombinant cytochrome P450 enzymes) [32]. They suggested that it is predominantly due to greater nonspecific binding in hepatocytes compared with rP450s. After correcting for nonspecific binding, there was an excellent correlation ($R^2 = 0.88$, $p < 0.0001$) of the $IC_{50,\ unbound}$ values generated in two different milieus. They proposed that rP450s had the advantage of allowing the use of nonselective probes such, as naproxen *O*-demethylation, for CYP2C9. While, for compounds concentrated in the liver due to cellular transport, intact hepatocytes would be a better choice for cytochrome P450 inhibition testing.

With the development of high throughput screening (HTS), recombinant CYP enzymes, together with fluorogenic probe substrates are frequently employed by many pharmaceutical companies in the early stages of drug discovery [33]. Cohen et al. have examined the correlation between IC_{50} values obtained with fluorogenic and conventional probes for structurally diverse inhibitors of CYP 3A4, 1A2, 2D6, 2C9 and 2C19. However, the correlations were poor with significant numbers of compounds being missed as inhibitors by both types of probe [34]. Zlokarnik et al. discussed how HTS methods based on fluorescence and luminescence substrates did not provide the same answers as assays based on selective substrates and liver microsomes [33]. Furthermore, Refsgaard et al. have observed a poor correlation between inhibition data obtained by application of human liver microsomes (HLM) with conventional substrates and the recombinant enzyme CYP2D6 with fluorogenic substrates [28].

Experimental results also vary among laboratories. The $CL_{int,\ in\ vitro}$ has been generally regarded as a rapidly and easily determined parameter more amenable to HTS than its in vivo counterpart [35, 36]. However, this assertion has been challenged from time to time [37–39]. Nagilla et al. contrasted published $CL_{int,\ in\ vitro}$ values from different laboratories [40]. They found that data available in the literature generally varied significantly from laboratory to laboratory, and $CL_{int,\ in\ vitro}$ ranges of more than 10-fold were not uncommon. They suggested that attention should be paid to the experimental methods used (i.e., substrate disappearance versus metabolite generation measurement methods), differences in scaling factors, in particular the incubation matrix applied (addition of plasma or blood serum in hepatocyte experiments), the presence or absence and type of cofactors in the incubation medium. So, when constructing QSAR models for drug metabolism, the metabolic parameters should be measured by the same standard procedure, collected from literature in the same laboratory, so that the data on different compounds can be reliably compared with each other. This accuracy/consistency of the datasets will eliminate the effect of the data on the accuracy and predictability of the model.

Selection of training and test sets

The rational division of the entire dataset into training and test sets is one of the most important steps governing the predictability of a QSAR model [41]. Golbraikh et al. argued that training and test sets must satisfy the following criteria: (i) representative points of the test set must be close to those of the training set; (ii) representative points of the training set must be close to representative points of the test set; (iii) the training set must have sufficient diversity [42].

Fundamentally, the whole dataset can be divided into training and test sets in a random manner [43]. Another approach is based on sorted biological activity data. The whole range of activities is divided into bins, and

compounds belonging to each bin are randomly (or in some regular way) assigned to the training or test set [44, 45]. Obviously, these methods cannot guarantee that compounds in the training set represent the entire descriptor space of the original dataset, and that each compound-point of the test set is close to at least one point of the training set. Other methods, such as cluster-based methods, dissimilarity-based methods, cell-based methods, stochastic techniques, statistical experimental designs and neural networks [46–48] have also been used to create diverse training sets and representative test sets. Often, the *K*-means clustering algorithm (a cluster-based method) was used to choose the representative subsets of compounds in the QSAR process.

For example, Leonard and Roy compared the predictability of QSAR models based on three methods for three data sets of different sizes (n = 35, 56 and 87) [49]. When the training and test sets were generated by random division and sorted biological activity data, predictive models were not obtained in most cases. However, good external validation statistics were obtained when training and test sets were selected based on *K*-means clusters of factor scores of the descriptor space along with/without the biological activity values. In contrast, little research has been performed to investigate the effects of different dataset splitting methods on model predictability in drug metabolism.

Diversity of the dataset

Predictive accuracy of statistical learning systems is known to be strongly affected by the diversity of the samples used in the training and test sets [50, 51]. The structural diversity of a dataset can be evaluated by calculating the diversity index (DI), the average value of the similarity between all the pairs of compounds in a dataset [52]. The smaller the DI value, the higher the diversity of the datasets. The similarity between two compounds is commonly described by the Tanimoto coefficient [53]. The mean Tanimoto coefficient of the compounds in datasets A and B can be used as a representativeness index (RI) to measure the level of representativeness of dataset A by dataset B. Dataset B is more representative of dataset A if the RI value between datasets A and B is higher.

For instance, Yap and Chen explored the use of different statistical learning methods for predicting inhibitors and substrates of CYP3A4, CYP2D6 and CYP2C9 [54]. In their study, the DI values of the six training sets and six test sets were in the range 0.001–0.005 and 0.002–0.020, respectively, suggesting that these datasets were sufficiently diverse. The RI values between each of the training sets and its corresponding test set ranged from 0.446 to 0.511, suggesting that these test sets were suitable for evaluating the model developed by the training sets. As a result, for inhibitors classification of CYP2C9, the best consensus support vector machine model gave accuracy (ratio of correctly classified compounds) of 88.9% for inhibitors and 96.3% for non-inhibitors. For substrates classification of CYP3A4, CYP2D6, and CYP2C9, values of accuracy were 98.2 and 90.9%, 96.6 and 94.4%, and 85.7 and 98.8%, respectively. In their later successful study of the prediction of total clearance with QSAR models using three different statistical learning methods [55], the DI values of the training and test sets were 0.067 and 0.068, respectively, while the DI values in earlier studies ranged from 0.274 to 0.584, and the RI was 0.881. This suggested that the datasets were sufficiently diverse and the test set was representative of the training set. However, until now, most of the QSAR studies involving drug metabolism prediction have not used the DI and RI values to evaluate the diversity of their modeling datasets.

Molecular descriptor generation and variable selection

Molecular descriptor generation

In the process of QSAR model construction, various rationally designed molecular descriptors are needed to examine molecular structures. Also, the molecular structures can be drawn or searched for within databases, such as PubChem, Crossfire Beilstein, Drugbank and Scifinder. After that, some free on-line software (such as PreADME/T [56], molecular descriptor lab (MODEL) [57], chemistry development kit (CDK) [58] and E-DRAGON [59]) can be employed to generate the molecular descriptors.

Different descriptors emphasize different chemical properties implicit in the molecular structure, usually divided into 2D and 3D descriptors [60]. The 2D descriptors are independent of the 3D orientation of drugs, including constitutional, electronic, quantum chemical, topological, geometrical descriptors, fragment-based descriptors and fingerprints. Generation of 3D descriptors initially involves determination of the molecular conformation. The conformation is then refined by minimizing the energy [61] and, subsequently, alignment of the conformers uniformly in space. Finally, the space with immersed conformer is probed computationally for various descriptors. Comparative molecular field analysis (CoMFA) with electrostatic and steric energy fields [62] and comparative molecular similarity indices (CoMSIA) [63] with steric, electrostatic, and hydrophobic properties are commonly used for alignment-dependent 3D descriptors. Some 3D descriptors are derived independently of the molecular alignment, such as comparative molecular moment analysis (CoMMA) [64], weighted holistic invariant molecular (WHIM) descriptors [65], VolSurf [66] approach, and grid-independent descriptors (GRIND) [67].

Variable selection

A large number of structural descriptors can be generated by various forms of software. Therefore, selection of proper and interpretable descriptors to establish QSAR models is a very important step to reduce over-fitting, speed up training, improve the overall model predictability, and to interpret the QSAR model. At the same time, this is also a challenging and difficult step.

The correlation coefficient-based method, also called the "univariate" approach [68], may serve as a preliminary filter for discarding inter-correlated descriptors. This can be done by estimating correlations between every pair of descriptors and, if it exceeds a threshold (usually 0.85), then removing one of descriptor pairs. The removal can be random or based on the correlation coefficient between the descriptors and biological activity. In addition, "sequential" approaches, including forward variable selection [69–71] and backward feature elimination methods [72–77], select the best variable first, and then the best pair formed by the first and second and so on in a forward or backward progression. Recently, some stochastic methods, such as genetic algorithms (GA) [78–82] and simulated annealing (SA) [83], evolutionary programming [84], artificial ants [85, 86], and particle swarms [87] have been widely employed in QSAR variable selection, but less frequently used in QSAR for drug metabolism.

Genetic Algorithms (GA): GA methods, an efficient strategy to search for the global optimum of the solutions, were developed to mimic some of the processes observed in natural evolution. In the variable selection process, GA methods have shown their ability to reduce the noise introduced by the descriptors that did not have a relationship with the studied structure–property relationship [88]. GA methods, in combination with linear and non-linear statistical approaches (GA-PLS [89–92], GA-SVM and GA-ANN [93, 94]) used as fitness functions, have been applied to perform variable selection of QSAR models. For example, Yap and Chen used GA-SVM to remove descriptors irrelevant in the prediction of CYP 450 inhibitors and substrates [54]. In the descriptor selection process, the Matthews correlation coefficient derived from the SVM classification was used as the fitness function for genetic algorithm optimization. As described in section "Diversity of the dataset", the models offered highly accurate predictions.

When developing QSAR models, three strategies can be considered: (i) linear variable selection/linear modeling; (ii) linear variable selection/nonlinear modeling and (iii) nonlinear variable selection/nonlinear modeling. For modeling the processes with linear characteristics, variable selection can easily be solved with strategy (i). However, the performance of strategy (ii) applied for nonlinear systems may not always be productive. The linear variable selection strategy may ignore some information important for the nonlinear model. In the case of strategy iii (for GA-ANN), it is a complex procedure due to the optimization of GA together with ANN parameters [95]. A novel nonlinear variable selection method, called genetic algorithm-kernel partial least square (GA-KPLS), integrated GA as a powerful optimization method and KPLS as a robust nonlinear statistical method for variable selection, and was reported first by Jalali-Heravi and Kyani [96]. They combined GA-KPLS with ANN to develop a nonlinear QSAR model for predicting carbonic anhydrase II inhibiting activities of a series of substituted aromatic sulfonamides. As a result, the R^2 (root mean square error for leave-one-out validation, $RMSE_{cv}$) for GA-KPLS-ANN and MLR models ($n = 47$) were 0.913 (0.101) and 0.728 (0.291), respectively. The R^2 ($RMSE_{cv}$) for GA-KPLS-ANN and GA-PLS-ANN models ($n = 114$) were 0.899 (0.229) and 0.851 (0.277), respectively. This suggests the superiority of this method (GA-KPLS-ANN) to the linear one (MLR). Also, GA-KPLS is a powerful method for variable selection in nonlinear systems compared with GA-PLS-ANN.

Simulated Annealing (SA): SA is another widely used stochastic technique for variable selection in QSAR. The principles of SA optimization of variable selection have been described in detail elsewhere [97]. SA is derived from a thermodynamic process called annealing and it proceeds stepwise through a search space defined by all possible subsets to find global optimum subsets of the QSAR model.

Although SA has only been employed in QSAR studies of drug discovery for a few years [98, 99], it has been found that SA can deal with highly nonlinear models, as well as chaotic and noisy data. For example, Ng et al. studied the prediction of the distribution volume at steady-state (V_{ss}) and the clearance (CL) of 44 antimicrobial agents in humans [100]. SA-based variable selection followed by *k*-NN (SA-*k*-NN) and PLS analysis were employed to develop the QSAR models. The best SA-*k*-NN model was highly predictive of Vss ($R^2 = 0.80$) and CL ($R^2 = 0.94$), respectively. For V_{ss}, the model provided an average x-fold error value of 1.00 (1.05 for CL), and 93% of the compounds were within a two-fold range of their actual values. By contrast, PLS methods (linear regression technique) resulted in inferior QSAR models ($R^2 = 0.53$ for V_{ss} and 0.35 for CL).

The major advantages of SA over other global search methods (such as GA) are its ability to avoid being trapped in local optimization [101]. Recently, Jung et al. performed variable selection of tacrine derivatives against acetylcholinesterase activity using the SA-MLR, stepwise MLR, and GA-MLR methods [102]. The best equation was obtained from SA-MLR. The $R^2_{\text{traning set}}$ for SA-MLR, stepwise MLR, and GA-MLR methods were 0.959, 0.957,

and 0.955 while the $R^2_{\text{test set}}$ for these models were 0.844, 0.731 and 0.765, respectively. Therefore, the SA approach may be a useful tool for variable selection in drug metabolism prediction.

Ensemble methods: With the development of ensemble modeling approaches, multiple models of different types have been developed resulting in different descriptor subsets for each model type. To the best of our knowledge, the linear regression model is generally more interpretable than a neural network model but it is also generally less accurate than such a model. However, recent work on the interpretability of neural network QSAR models allows us to provide detailed interpretations of structure–activity trends, in a manner similar to the interpretation of linear regression models [103–105]. Dutta et al. used the method, called ensemble descriptor selection, to select a single, optimal and consistent subset of descriptors for multiple QSAR models. This approach searched for the descriptor subsets using the GA method, the fitness function of which was a linear combination of the root mean square error (RMSE) for the different model types [106]. This method was applied to three data sets, covering both regression and classification. The results obtained showed that the limitation of forcing different model types to use the same set of descriptors did not lead to a significant loss in predictive ability for the individual models (the actual magnitude of the decrease in performance was usually under 12%). For example, for the prediction of the IC_{50} values of 79 platelet-derived growth factor inhibitors (the first data sets), the RMSE of the training set (test set) was 0.33 (0.52) for ANN, 0.54 (0.32) for MLR, 0.37 (0.55) for the ensemble ANN and 0.58 (0.41) for the ensemble MLR. The ensemble approaches could find common variables in ANN and MLR models. Also, the effect of these descriptors on the inhibitor performance could be interpreted in detail.

Therefore, the ensemble variable selection method can be applied to establish QSAR models in drug metabolism and elucidate the main descriptors influencing interactions between drugs and the metabolic enzymes.

Hybrid methods: Various hybrid methods have been used to select the desired variables in QSAR studies. On the one hand, a rapid objective method is used as a preliminary filter to reduce the feature set. Next, the more accurate, but slower, subjective method is used [107]. On the other hand, different stochastic approaches should be applied simultaneously to obtain the essential descriptors influencing the chosen pharmacokinetic parameter [108, 109].

Novel statistical techniques less frequently used in drug metabolism QSAR

Machine learning techniques used for in silico modeling of drug metabolism have been summarized in recent reviews [30, 110–113]. In addition, other novel statistical approaches, such as the line-walking method [114] and the Gaussian kernel weighted k-nearest neighbor method, [115] have also been used to predict drug interactions with cytochrome P450 enzyme. However, most of these methods are difficult to interpret and easily fall into over-fitting. In the following sections, we will provide a brief overview of some novel, robust and interpretable statistical approaches, which have been successfully applied to QSAR models in drug development, but have rarely been used for drug metabolism prediction.

Ensemble methods: Random Forest (RF), Bagging, and Adaboost

Traditional approaches to develop QSAR analysis focused on constructing a single predictive model. Recently, methods using a combination or ensemble of models for improving the predication have been proposed.

Random Forest (RF): In the RF method [116], sub-data sets are generated by a re-sampling method called bootstrapping. The input features of the same number are randomly chosen for each sub-data set. Accordingly, RF can be applied to data sets with many more variables than data (observations). The classification trees of the sub-datasets are fully grown and are un-pruned so as to obtain low bias trees. Each tree gives a classification (vote) and the RF chooses the classification having the most votes. RF is robust with respect to noise features and unlikely to overlearn the training data. The RF method has recently been widely used in building QSAR models, including P-glycoprotein transport activity [117], the distribution volume of drugs in humans [118], and mutagenicity [119], and provided better results than the single decision tree and PLS. Bruce et al. demonstrated that RF was an interpretable and powerful model, far better than SVM and *k*-NN [120]. Sakiyama et al. derived QSAR models for predicting human microsomal metabolic stability with RF, SVM, logistic regression, and RP methods. In this, 535 stable ($CL_{\text{int, in vitro}} < 20$ ml/min/kg) and 1417 unstable ($CL_{\text{int, in vitro}} \geq 20$ ml/min/kg) compounds were assigned in the training set, with 133 stable and 354 unstable compounds in the test set [121]. The results demonstrated that all classifiers yielded satisfactory results, with an accuracy > 0.8, and sensitivity > 0.9 (also known as Recall, the ratio of correctly classified compounds, the true state of which is positive), specificity > 0.6 (ratio of correctly classified compounds, the true state of which is negative) and precision > 0.8 (positive precision: ratio of correctly classified compounds, which are predicted to be positive compounds; negative precision: ratio of correctly classified compounds, which are predicted to be negative compounds; the precision here means the average of positive

and negative precision). However, logistic regression yielded a relatively lower negative precision (0.74) than RF (0.90), SVM (0.90) and RP (0.83), while RP produced a relatively lower specificity (0.61) than RF (0.68), SVM (0.71) and logistic regression (0.68). These results indicate that the RF and SVM methods have a slightly higher performance and reliability than logistic regression and RP. The RF model was useful in identifying compounds with the potential risk of being metabolically unstable, which should be a warning in the early drug discovery process, thus avoiding the high costs of late-stage failure.

Bagging: Bagging, a meta-learning method, is used in combination with other algorithms [122]. In Bagging, multiple classifiers are generated using the multiple learning data sets obtained by bootstrap re-sampling based on the experimental distribution. In this, each classifier is trained by a different data set (but overlapping) and these classifiers are integrated for the final classification.

Adaboost: In modeling, the Adaboosting method uses all the descriptors and all the training compounds. According to prediction error, the weight of each training data set with a wrong result is repeatedly adjusted in each round. Thus, classifiers are trained sequentially using the given training data. The final classification is done by a weighted-majority vote based on the multiple weak classifiers obtained in each round. The advantage of Adaboosting is that it allows the use of relatively simple and erroneous base models. Similar to the SVM classifier, the power of boosting stems from its ability to create decision boundaries maximizing the margin [123].

Koike et al. compared the performance of five variable-selection methods for predicting chemical compound-protein binding affinities, including GA, the recursive feature elimination method, the information-gain method, the Fisher discriminant ratio and the Odds ratio. Six classification methods were used to establish QSAR models for six series of compounds, containing cytochrome P450 2C9 inhibitors (data gathered from the study of Yap and Chen [54]) [124]. It was found that GA was superior to the other variable-selection methods, and its combination with RF and Adaboost or Bagging gave almost the same performance as SVM and better than the other classifiers. For CYP2C9 inhibitor prediction, using GA as the variable selection method, the overall accuracy for the training set (test set) of SVM, AdaBoost + RF, Bagging + RF, RF, C4.5 and Naïve Bayse was 98.0% (95%), 99.8% (95%), 99.5% (93%), 99.3% (95%), 93.1% (87%) and 42.0% (34%), respectively. They suggested that the lipophilicity-related descriptors, such as SlogP and SlogP_VSA5, the number of double bounds and polarizability-related descriptors, such as the Wiener polarity number, were probably essential for the classification of CYP2C9 inhibitors/non-inhibitors.

Multivariate adaptive regression splines (MARS)

MARS can be considered as a generalization of classification and regression trees (CART) which is able to overcome some limitations of CART [125]. MARS is a local regression method and uses a series of local so-called basis functions to model the complex (non-linear) relationships. The space of the predictors is split into several (over-lapping) regions in which so-called spline functions are fitted. Then, the global MARS model consists of the weighted sum of the local models. In general, the MARS methodology consists of three steps. Firstly, a constructive phase, in which basis functions are introduced in several regions of the predictors and are combined in a weighted sum to define the global model. This model often contains too many basis functions, which leads to over-fitting. Therefore, the constructive phase is followed by a pruning phase, in which some basis functions of the over-fitting model are deleted. This leads to a sequence of consecutively smaller MARS models, from which the optimal one is selected in a third step.

Unlike neural networks, MARS can identify 'important' independent variables through the built basis functions (giving the importance percent) when many potential variables are considered, and can model complex relationships among variables. Besides, MARS does not need a long training process and, hence, can save lots of modeling time especially when the data set is huge. Furthermore, MARS points out which variables are important in eliciting an effect on the interested parameters, and gives a specific equation. Due to the advantage of model accuracy, reliability and implementation, the MARS technique has been successfully used in QSAR [126–129], and quantitative structure–retention relationships (QSRR) [130], breast cancer pattern mining [131] and, currently, QSAR in the human gastro-intestinal absorption of drugs [132–134]. For example, Deconinck et al. successfully modeled the gastro-intestinal absorption of 67 drugs using MARS and obtained a quantitative relationship equation [133]. This MARS model with 12 basis functions, had a lower $RMSE_{CV}$ (0.13) and a higher determination coefficient between experimental and predicted values ($R^2 = 0.93$), than the stepwise MLR model ($RMSE_{CV} = 0.17$, $R^2 = 0.74$) and the PLS model ($RMSE_{CV} = 0.16$, $R^2 = 0.82$). Therefore, it seems that MARS may be a powerful approach in drug metabolism prediction.

Graph machines

Goulon et al. have described a novel statistical method, named graph machines, which circumvents the problem of designing, computing and selecting molecular descriptors [135]. As an alternative approach to traditional machine-

learning-based QSAR, the graph machines contained the following steps: (i) molecules were represented as graphs from the SMILES file; (ii) a mathematical function (graph machine), called "node function" (e.g. a feed-forward neural network), was built for each example of the data set; (iii) the parameters of the node functions, shared both within and across the graph machines, were estimated by training from examples (i.e., by minimizing a suitable cost function that expressed the discrepancy between the measurements and the corresponding predictions); (iv) model selection was then performed by traditional cross-validation.

Goulon et al. have evaluated the performance of the graph machine approach through QSAR modeling in predicting the toxicity and anti-HIV activity of phenols simultaneously [135]. In the phenol toxicity model, graph machines were built for each example of the whole set (131 phenol derivatives for the training set and 22 for the test set). Consequently, the RMSE of the training set (test set) for MLR, the radial basis function neural networks (RBFNN), SVM and the graph machine method were 0.30 (0.46), 0.19 (0.29), 0.22 (0.36) and 0.19 (0.27), respectively. The results indicated that graph machines provide the best results on the test set, and did not require the computation and selection of descriptors as opposed to other alternative methods. In the prediction of the anti-HIV activity of HEPT derivatives, graph machines were compared with the performances of 4D-QSAR using PLS and SOM (RMSE were 0.98, 1.41 and 1.39 for test set 1). The latter two models require both the computation of any descriptor and the optimization of the molecular geometry.

With its powerful predictability and no need for the selection and computation of descriptors, the graph machine may be an efficient method in drug metabolism prediction, although no application has yet been reported involving drug metabolism prediction.

Recommendations for developing QSAR models in drug metabolism

As far as the considerations described above are concerned, we would like to present some recommendations for developing QSAR models in drug metabolism. Many of them are common sense or have appeared previously but in different reports. We have collected them here in order to help the reader develop predictable and interpretable QSAR models for drug metabolism.

- Suitable metabolism parameter (such as IC_{50}, V_{max}, CL_h and $CL_{int,\ in\ vitro}$) should be selected first, and should be care of accuracy/consistency of the entire dataset.
- The molecular structures must be drawn from and searched within various databases to confirm the molecular structures, especially structures with stereochemistry. When different structures are found, the structure that occurs most frequently should be selected.
- Rational division of the whole dataset into training and test sets can be done by cluster-based methods, such as the *K*-means clustering algorithm. Concomitantly, evaluation of the dataset diversity by DI, and the representativeness between training and test sets by RI, should be performed.
- Variable selection can be performed by correlation coefficient-based method first, and then through some stochastic methods, such as GA, SA, and ensemble methods. However, different stochastic approaches should be applied simultaneously to obtain essential descriptors influencing the metabolic parameters.
- QSAR models can be developed by various linear (MLR, PLS and linear discriminant analysis (LDA)) and non-linear statistical methods (NBC, *k*-NN, SOM, RP, ANN, SVM, RF, Bagging, Adaboost, MARS and graph machines). However, due to the complexity of drug metabolism, different non-linear methods should be chosen concurrently as a modeling approach.

Recent advances in QSAR models for cytochrome P450-mediated drug metabolism

According to the endpoint selected, QSAR models have been established to predict various levels of drug metabolism, such as in vivo hepatic clearance, drug metabolic stability and drug–CYP interactions. We can briefly summarize the latest advances as follows.

QSAR models for in vivo hepatic clearance and in vitro metabolic stability

Attempts have been made to predict human in vivo hepatic clearances from in vitro hepatocytes and animal in vivo/in vitro data by applying multivariate statistics [136, 137]. However, most of these models are based on experimental data and are limited by time, cost and resources. Recently, Lee and Kim reported a new method for predicting human hepatic clearance using the MLR approach from in vitro experimental data and six molecular descriptors (including molecular weight, number of total atoms, number of aromatic rings, number of single bonds, TPSA, and SKlogP calculated from PreADME/T website) of 19 (human hepatocyte data, dataset 1) and 30 (human microsomal data, dataset 2) drugs based on leave-one-out evaluation [138]. In contrast to the in vitro–in vivo scaling factor method (A) and MLR analysis using only the experimental

data method (B), the new method (C) was the most accurate prediction model. R^2 of dataset 1 (dataset 2) for method A, B and C were 0.795 (0.783), 0.857 (0.809) and 0.921 (0.883), respectively. These results indicate that the information about molecular descriptors is significant in improving the predictive accuracy of human in vivo hepatic clearance.

As described in section "Ensemble methods: Random Forest (RF), Bagging, and Adaboost", Sakiyama et al. has derived QSAR models for predicting human microsomal metabolic stability and obtained satisfactory results [121]. After that, Lee et al. developed highly predictive classification models for human liver microsomal stability using RF and Bayesian methods with 15283 compounds (11646 for the training set, 2911 for the test set and 726 for the validation set) [139]. The best predictive RF model based on MOE and E-state descriptors showed 80% and 75% prediction accuracy for the test and validation sets, respectively. The model indicated that molecular hydrophobicity is an important descriptor to differentiate HLM stable compounds from unstable ones, and this was then used in sub-setting analysis to select representative sets of compounds for actual HLM experiments in Pfizer discovery projects. The initial results showed that 377 of 465 experimentally stable compounds and the 744 of 788 experimentally unstable compounds were correctly predicted, with the prediction accuracy being 90%.

QSAR models for CYP450 inhibition identification

Drug–drug interactions are an important issue in human health care. Many of the major pharmacokinetic interactions between drugs are due to hepatic CYP450 enzymes inhibited by concomitant administration of other drugs. Furthermore, modulation of this enzyme activity by CYPs inhibitors could have important implications for cancer prevention. Thus, the following section will focus on models for predicting CYPs inhibitors.

Since CYP3A4 is one of the most abundant CYP isoforms responsible for the metabolism of almost 50% of known drugs, early identification of CYP3A4 inhibitors is essential to minimize the risk associated with clinically relevant drug–drug interactions. Thus, several QSAR models for the classification of CYP3A4 inhibition have been developed [140–144]. Unfortunately, due to differences in the chemical compound collections and experimental methodologies between research institutions, the published models are of limited use when applied to internal drug discovery projects.

Recently, Arimoto et al. developed QSAR models of rCYP3A4 inhibition with SVM for a large data set of 4000 proprietary compounds [145]. The internal experimental percentage inhibition of 7-benzyloxy-4-trifluoromethylcoumarin was considered as the target variable. The overall accuracy of the three best predictive models was 82, 82, and 81%, and the kappa values were 0.62, 0.61, and 0.62 for the test set. Also, the consensus model yielded a further improvement in the kappa (0.65) and accuracy (83%) values. Interestingly, the examination of the model predictability with an additional 2195 compounds showed a strong correlation between the predictive performance and the structural similarity between the test and the training sets. The predictive power of the model dropped sharply when the similarity index fell below 0.8. Also, they observed that the limit should be defined as 0.7 to sustain the applicability of the models with kappa > 0.4.

Subsequently, Zhou et al. built a successful and fast binary classification model of CYP3A4 inhibition using SVM with a balanced training set (826 inhibitors and 873 non-inhibitors) [146]. The model achieved 83% accuracy for the test set (186 inhibitors and 677 non-inhibitors). Similarly, they showed that when new compounds are evaluated, their Tanimoto similarity index to the training set compounds should be evaluated with a limit of 0.7. In addition, they also proposed the distance to the separating surface in the feature space of the SVM model as another valuable confidence index for prediction. If the absolute distance fell below 0.5, the prediction was less reliable.

Many QSAR models have also focused on the identification of CYP2D6 and CYP1A2 inhibition [147–149]. However, in these models, the analysis of dataset diversity and the different metabolic parameters linked to the inhibition process are insufficient. According to this, Burton et al. constructed CYP1A2 and CYP2D6 inhibition models with RP and sufficiently diverse datasets (498 and 306 compounds respectively) [150]. More importantly, they compared the performance of the global, probe substrates, K_i and IC_{50} based models. They found the most predictive models were K_i based models. For CYP2D6, the accuracy, sensitivity, specificity and positive precision of the external validation set (training set) were 89% (90%), 91% (88%), 81% (92%) and 91% (90%), respectively. For CYP1A2, these measures were 81% (89%), 76% (95%), 86% (83%) and 85% (85%), respectively.

In summary, all of these QSAR models have their own advantages and make a great contribution to CYP inhibition prediction.

QSAR models for CYP450 substrate recognition

QSAR models for early identification of the predominant CYP450 isoforms responsible for drug metabolism and the specific sites of certain metabolic reactions, not only contribute to the elucidation of drug–drug interactions, but also help make drug design more predictable and rational in the early stages of drug discovery.

Manga et al. [151] have established a simple and interpretable model for the determination of CYP450 isoforms predominantly responsible for drug metabolism, involving CYP3A4, CYP2D6 and CYP2C9. FIRM (formal inference-based recursive modeling) was used as model construction method and recursive partitioning was used as a variable selection tool. The final model with three classes of descriptors (lipophilicity, bulk and Broensted acidity) resulted in a correct classification ratio of 94% for the training set and 68% for the test set. Yap and Chen [54] reported a similar study of a different data set using consensus SVM (see section "Diversity of the dataset").

After that, Terfloth et al. [152] investigated the issue of predicting the isoform specificity for cytochrome P450 3A4, 2D6, and 2C9 substrates. Models were constructed with five model-building methods, including *k*-NN, decision tree (C4.5/J48), multilayer perception, RBF (Radial basis function neural networks), logistic regression, and SVM. In this, 146 drugs of the training set were collected from the paper by Manga et al. (149 drugs), while 3 drugs were removed from the test set because they were already present in the training set, and stereochemistry was added for 95 compounds that had one or more steric centers or unsymmetrical double bonds. The model with the best results was established by combining automatic variable selection (best first search) with SVM. In the leave-one-out cross-validation, 89% of all compounds in the training set were correctly classified. The accuracy of the test set (303 compounds) was 83%, a substantial improvement in comparison with the value of 68% achieved by Manga et al. [151].

Knowing where a molecule is preferentially oxidized, that is, the region-selectivity by a particular CYP, would give medicinal chemists insight into where to block the metabolism and make their drug candidates more stable in vivo. Therefore, a computational model for region-selectivity could allow chemists to make rational decisions more quickly. To date, almost all models of CYP region-selectivity have been mechanism-based [153–156]. However, Sheridan et al. [157] have presented QSAR-based region-selectivity models for CYP450 3A4, 2D6, and 2C9. Three substructure descriptors and two physical property descriptors of 557 molecules (532 for training set and 25 for test set) were used, with RF to cover the most commonly observed potential sites of oxidation (sp3 carbons, sp2 carbons, sulfurs, etc.). The descriptor conclusions from the RF model showed that factors other than the immediate chemical environment and the accessibility of the hydrogen atoms affected the region-selectivity in all three isoforms (CYP 3A4, 2D6, and 2C9). The cross-validated predictions were compared with the results obtained from the earlier mechanistic model [20] and MetaSite [154]. For the same set of compounds (n = 316), the oxidation site was in the top two atoms in 77, 51 and 62% of the molecules for the QSAR model, an earlier mechanistic model and MetaSite, respectively. This indicated that the QSAR model for 3A4 was clearly better than the earlier mechanistic model and MetaSite. For 2D6 (n = 124) and 2C9 (n = 92), the QSAR predictions were only slightly better than those of the earlier mechanistic model and MetaSite (The oxidation site was in the top two atoms in 72%/73% (2D6/2C9), 24%/31%, and 65%/69% of the molecules for the QSAR model, the earlier mechanistic model and the MetaSite model, respectively).

Conclusions and prospects

This review has highlighted the statistical aspects which should be emphasized in the QSAR process for drug metabolism, such as the diversity/consistency, representation and diversity of the data sets (training and test sets), variable selection and the potential application of novel statistical methods. However, since drug metabolism is an extremely complex pharmacokinetic process, accurate modeling of the drug–metabolic enzyme interactions, including the metabolic degree (hepatic metabolic clearance), type of metabolic enzymes (CYPs or UGTs), type of interactions (substrates, inducers, or inhibitors), site of the interaction (hydrophilic domain or hydrophobic region), and the stereo-chemical selectivity, is difficult. Also, various approaches should be combined to predict the complex drug metabolism process. For example, QSAR models should be combined with pharmacophore-based approaches or docking methods. As a consequence, the quantitative relationship contained in the QSAR model can be clearly explained and certain mechanism pharmacophore-based or docking methods can also be used to assist in the design of new drugs.

References

1. van de Waterbeemd H, Gifford E (2003) Nat Rev Drug Discov 2:192. doi:10.1038/nrd1032
2. Fostel J (2005) Expert Opin Drug Metab Toxicol 1:565. doi:10.1517/17425255.1.3.565
3. Kola I, Landis J (2004) Nat Rev Drug Discov 3:711. doi:10.1038/nrd1470
4. Ruiz-Garcia A, Bermejo M, Moss A, Casabo VG (2007) J Pharm Sci [Epub ahead of print]
5. Li AP (2001) Drug Discov Today 6:357. doi:10.1016/S1359-6446(01)01712-3
6. Mager DE (2006) Adv Drug Deliv Rev 58:1326. doi:10.1016/j.addr.2006.08.002
7. Kumar GN, Surapaneni S (2001) Med Res Rev 21:397. doi:10.1002/med.1016
8. Fagerholm U (2007) J Pharm Pharmacol 59:803. doi:10.1211/jpp.59.6.0007

9. Yamashita F, Hashida M (2004) Drug Metab Pharmacokinet 19:327. doi:10.2133/dmpk.19.327
10. de Groot MJ, Kirton SB, Sutcliffe MJ (2004) Curr Top Med Chem 4:1803. doi:10.2174/1568026043387061
11. Zamora I, Afzelius L, Cruciani G (2003) J Med Chem 46:2313. doi:10.1021/jm021104i
12. Domanski TL, Halpert JR (2001) Curr Drug Metab 2:117. doi:10.2174/1389200013338612
13. Lewis DF (2002) Drug Metab Rev 34:55. doi:10.1081/DMR-120001390
14. Dai R, Pincus MR, Friedman FK (2000) Cell Mol Life Sci 57:487. doi:10.1007/PL00000709
15. De Rienzo F, Fanelli F, Menziani MC, De Benedetti PG (2000) J Comput Aided Mol Des 14:93. doi:10.1023/A:1008187802746
16. Sali A, Blundell TL (1993) J Mol Biol 234:779. doi:10.1006/jmbi.1993.1626
17. Shoichet BK, Kuntz ID (1993) Protein Eng 6:723. doi:10.1093/protein/6.7.723
18. Williams PA, Cosme J, Ward A, Angove HC, Matak Vinković D, Jhoti H (2003) Nature 424:464. doi:10.1038/nature01862
19. Jones JP, Mysinger M, Korzekwa KR (2002) Drug Metab Dispos 30:7. doi:10.1124/dmd.30.1.7
20. Singh SB, Shen LQ, Walker MJ, Sheridan RP (2003) J Med Chem 46:1330. doi:10.1021/jm020400s
21. Beck ME (2005) J Chem Inf Model 45:273. doi:10.1021/ci049687n
22. Korzekwa KR, Jones JP (1993) Pharmacogenetics 3:1. doi:10.1097/00008571-199302000-00001
23. Korhonen LE, Turpeinen M, Rahnasto M, Wittekindt C, Poso A, Pelkonen O et al (2007) Br J Pharmacol 150:932. doi:10.1038/sj.bjp.0707173
24. Kurogi Y, Güner OF (2001) Curr Med Chem 8:1035
25. Ekins S, Stresser DM, Williams JA (2003) Trends Pharmacol Sci 24:161. doi:10.1016/S0165-6147(03)00049-X
26. Hansch C, Fujita T (1964) J Am Chem Soc 86:1616. doi:10.1021/ja01062a035
27. Eriksson L, Johansson E, Müller M, Wold S (2000) J Chemometr 14:599. doi:10.1002/1099-128X(200009/12)14:5/6<599::AID-CEM619>3.0.CO;2-8
28. Refsgaard HH, Jensen BF, Christensen IT, Hagen N, Brockhoff PB (2006) Drug Dev Res 67:417. doi:10.1002/ddr.20108
29. Cronin MTD (2005) In: Helma C (ed) Predictive toxicology. Taylor & Francis Press, USA, p 93
30. Madden JC, Cronin MT (2006) Expert Opin Drug Toxicol 2:545. doi:10.1517/17425255.2.4.545
31. Di Marco A, Yao D, Laufer R (2003) Eur J Biochem 270:3768. doi:10.1046/j.1432-1033.2003.03763.x
32. McGinnity DF, Tucker J, Trigg S, Riley RJ (2005) Drug Metab Dispos 33:1700. doi:10.1124/dmd.105.005884
33. Zlokarnik G, Grootenhuis PD, Watson JB (2005) Drug Discov Today 10:1443. doi:10.1016/S1359-6446(05)03580-4
34. Cohen LH, Remley MJ, Raunig D, Vaz AD (2003) Drug Metab Dispos 31:1005. doi:10.1124/dmd.31.8.1005
35. Huebert ND, Dasgupta M, Chen Y (2004) Curr Opin Drug Discov Devel 7:69
36. Korfmacher WA (2003) Curr Opin Drug Discov Devel 6:481
37. Andersson TB, Bredberg E, Ericsson H, Sjöberg H (2004) Drug Metab Dispos 32:715. doi:10.1124/dmd.32.7.715
38. Clarke SE, Jeffrey P (2001) Xenobiotica 31:591. doi:10.1080/00498250110057350
39. Masimirembwa CM, Bredberg U, Andersson TB (2003) Clin Pharmacokinet 42:515. doi:10.2165/00003088-200342060-00002
40. Nagilla R, Frank KA, Jolivette LJ, Ward KW (2006) J Pharmacol Toxicol Methods 53:106. doi:10.1016/j.vascn.2005.08.005
41. Eriksson L, Jaworska J, Worth AP, Cronin MT, McDowell RM, Gramatica P (2003) Environ Health Perspect 111:1361
42. Golbraikh A, Tropsha A (2002) J Comput Aided Mol Des 16:357. doi:10.1023/A:1020869118689
43. Yasri A, Hartsough D (2001) J Chem Inf Comput Sci 41:1218. doi:10.1021/ci010291a
44. Kauffman GW, Jurs PC (2001) J Chem Inf Comput Sci 41:1553. doi:10.1021/ci010073h
45. Mattioni BE, Jurs PC (2002) J Chem Inf Comput Sci 42:94. doi:10.1021/ci0100696
46. Leach AR, Gillet VJ (2003) An introduction to chemoinformatics. Kluwer Academic Publisher, Boston, p 123
47. Daszykowski M, Walczak B, Massart DL (2002) Anal Chim Acta 468:91. doi:10.1016/S0003-2670(02)00651-7
48. Wang YH, Li Y, Yang SL, Yang L (2005) J Comput Aided Mol Des 19:137. doi:10.1007/s10822-005-3321-5
49. Leonard JT, Roy K (2006) QSAR Comb Sci 25:235. doi:10.1002/qsar.200510161
50. Schultz TW, Netzeva TI, Cronin MT (2003) SAR QSAR Environ Res 14:59. doi:10.1080/1062936021000058782
51. Rajer-Kanduc K, Zupan J, Majcen N (2003) Chemom Intell Lab Syst 65:221. doi:10.1016/S0169-7439(02)00110-7
52. Perez JJ (2005) Chem Soc Rev 34:143. doi:10.1039/b209064n
53. Maldonado AG, Doucet JP, Petitjean M, Fan BT (2006) Mol Divers 10:39. doi:10.1007/s11030-006-8697-1
54. Yap CW, Chen YZ (2005) J Chem Inf Model 45:982. doi:10.1021/ci0500536
55. Yap CW, Li ZR, Chen YZ (2006) J Mol Graph Model 24:383. doi:10.1016/j.jmgm.2005.10.004
56. PreADME. http://preadmet.bmdrc.org/preadmet/index.php
57. Li ZR, Han LY, Xue Y, Yap CW, Li H, Jiang L et al (2007) Biotechnol Bioeng 97:389. doi:10.1002/bit.21214
58. Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen EL (2006) Curr Pharm Des 12:2111. doi:10.2174/138161206777585274
59. Tetko IV, Gasteiger J, Todeschini R, Mauri A, Livingstone D, Ertl P et al (2005) J Comput Aided Mol Des 19:453. doi:10.1007/s10822-005-8694-y
60. Dudek AZ, Arodz T, Gálvez J (2006) Comb Chem High Throughput Screen 9:213. doi:10.2174/138620706776055539
61. Akamatsu M (2002) Curr Top Med Chem 2:1381. doi:10.2174/1568026023392887
62. Cramer RD, Patterson DE, Bunce JD (1988) J Am Chem Soc 110:5959. doi:10.1021/ja00226a005
63. Klebe G, Abraham U, Mietzner T (1994) J Med Chem 37:4130. doi:10.1021/jm00050a010
64. Silverman BD, Platt DE (1996) J Med Chem 39:2129. doi:10.1021/jm950589q
65. Todeschini R, Lasagni M, Marengo E (1994) J Chemometr 8:263. doi:10.1002/cem.1180080405
66. Cruciani G, Crivori P, Carrupt P-A, Testa B (2000) J Mol Struct THEOCHEM 503:17. doi:10.1016/S0166-1280(99)00360-7
67. Crivori P, Zamora I, Speed B, Orrenius C, Poggesi I (2004) J Comput Aided Mol Des 18:155. doi:10.1023/B:JCAM.0000035184.11906.c2
68. Leardi R, Seasholtz MB, Pell RJ (2002) Anal Chim Acta 461:189. doi:10.1016/S0003-2670(02)00272-6
69. Bi J, Bennet K, Embrechts M, Breneman C, Song M (2003) J Mach Learn Res 3:1229. doi:10.1162/153244303322753643
70. Wegner JK, Fröhlich H, Zell A (2004) J Chem Inf Comput Sci 44:931. doi:10.1021/ci034233w
71. Xue CX, Zhang RS, Liu HX, Yao XJ, Liu MC, Hu ZD et al (2004) J Chem Inf Comput Sci 44:669. doi:10.1021/ci034248u
72. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Mach Learn 46:389. doi:10.1023/A:1012487302797

73. Xue Y, Li ZR, Yap CW, Sun LZ, Chen X, Chen YZ (2004) J Chem Inf Comput Sci 44:1630. doi:10.1021/ci049869h
74. Xue Y, Yap CW, Sun LZ, Cao ZW, Wang JF, Chen YZ (2004) J Chem Inf Comput Sci 44:1497. doi:10.1021/ci049971e
75. Xue CX, Zhang RS, Liu HX, Yao XJ, Liu MC, Hu ZD et al (2004) J Chem Inf Comput Sci 44:1693. doi:10.1021/ci049820b
76. Xue CX, Zhang RS, Liu MC, Hu ZD, Fan BT (2004) J Chem Inf Comput Sci 44:950. doi:10.1021/ci034280o
77. Xue CX, Zhang RS, Liu HX, Liu MC, Hu ZD, Fan BT (2004) J Chem Inf Comput Sci 44:1267. doi:10.1021/ci049934n
78. Terfloth L, Gasteiger J (2001) Drug Discov Today 6:102. doi:10.1016/S1359-6446(01)00173-8
79. Baumann K (2003) Trends Analyt Chem 22:395. doi:10.1016/S0165-9936(03)00607-1
80. Baurin N, Mozziconacci JC, Arnoult E, Chavatte P, Marot C, Morin-Allory L (2004) J Chem Inf Comput Sci 44:276. doi:10.1021/ci0341565
81. Hemmateenejad B, Safarpour MA, Miri R, Nesari N (2005) J Chem Inf Model 45:190. doi:10.1021/ci049766z
82. Barrett SJ, Langdon WB (2005) In: Ashutosh T, Joshua K, Erel A, Keshav D, Rajkumar R (eds) Applications of soft computing: recent trends. Springer Publisher, p 99
83. Itskowitz P, Tropsha A (2005) J Chem Inf Model 45:777. doi:10.1021/ci049628±
84. Luke BT (1994) J Chem Inf Comput Sci 34:1279. doi:10.1021/ci00022a009
85. Izrailev S, Agrafiotis D (2001) J Chem Inf Comput Sci 41:176. doi:10.1021/ci000336s
86. Izrailev S, Agrafiotis DK (2002) SAR QSAR Environ Res 13:417. doi:10.1080/10629360290014296
87. Agrafiotis DK, Cedeño W (2002) J Med Chem 45:1098. doi:10.1021/jm0104668
88. Rogers D, Hopfinger AJ (1994) J Chem Inf Comput Sci 34:854. doi:10.1021/ci00020a020
89. Leardi R, Gonzalez AL (1998) Chemom Intell Lab Syst 41:195. doi:10.1016/S0169-7439(98)00051-3
90. Leardi R (2000) J Chemometr 14:643. doi:10.1002/1099-128X(200009/12)14:5/6<643::AID-CEM621>3.0.CO;2-E
91. Hashida M (2005) Yakugaku Zasshi 125:853. doi:10.1248/yakushi.125.853 Article in Japanese
92. Guangli M, Yiyu C (2006) J Pharm Pharm Sci 9:210
93. González MP, Caballero J, Tundidor-Camba A, Helguera AM, Fernández M (2006) Bioorg Med Chem 14:200. doi:10.1016/j.bmc.2005.08.009
94. Hemmateenejad B, Safarpour MA, Taghavi F (2003) J Mol Struct THEOCHEM 635:183. doi:10.1016/S0166-1280(03)00418-4
95. Hemmateenejad B (2005) Chemom Intell Lab Syst 75:231. doi:10.1016/j.chemolab.2004.09.005
96. Jalali-Heravi M, Kyani A (2007) Eur J Med Chem 42:649. doi:10.1016/j.ejmech.2006.12.020
97. Sutter JM, Dixon SL, Jurs PC (1995) J Chem Inf Comput Sci 35:77. doi:10.1021/ci00023a011
98. Zheng W, Tropsha A (2000) J Chem Inf Comput Sci 40:185. doi:10.1021/ci980033m
99. Blower P, Fligner M, Verducci J, Bjoraker J (2002) J Chem Inf Comput Sci 42:393. doi:10.1021/ci0101049
100. Ng C, Xiao Y, Putnam W, Lum B, Tropsha A (2004) J Pharm Sci 93:2535. doi:10.1002/jps.20117
101. Hasan M, AlKhamis T, Ali J (2000) Comput Ind Eng 38:323. doi:10.1016/S0360-8352(00)00043-7
102. Jung M, Tak J, Lee Y, Jung Y (2007) Bioorg Med Chem Lett 17:1082. doi:10.1016/j.bmcl.2006.11.022
103. Guha R, Jurs PC (2005) J Chem Inf Model 45:800. doi:10.1021/ci050022a
104. Guha R, Stanton DT, Jurs PC (2005) J Chem Inf Model 45:1109. doi:10.1021/ci050110v
105. Stanton DT (2003) J Chem Inf Comput Sci 43:1423. doi:10.1021/ci0340658
106. Dutta D, Guha R, Wild D, Chen T (2007) J Chem Inf Model 47:989. doi:10.1021/ci600563w
107. Guha R, Jurs PC (2004) J Chem Inf Comput Sci 44:2179. doi:10.1021/ci049849f
108. Jensen BF, Sørensen MD, Kissmeyer AM, Björkling F, Sonne K, Engelsen SB et al (2003) J Comput Aided Mol Des 17:849. doi:10.1023/B:JCAM.0000021861.31978.da
109. Duchowicz PR, González MP, Helguera AM, Dias Soeiro Cordeiro MN, Castro EA (2007) Chemom Intell Lab Syst 88:197. doi:10.1016/j.chemolab.2007.05.001
110. Fox T, Kriegl JM (2006) Curr Top Med Chem 6:1579. doi:10.2174/156802606778108915
111. Arimoto R (2006) Curr Top Med Chem 6:1609. doi:10.2174/156802606778108951
112. Yap CW, Xue Y, Li ZR, Chen YZ (2006) Curr Top Med Chem 6:1593. doi:10.2174/156802606778108942
113. Jolivette LJ, Ekins S (2007) Adv Clin Chem 43:131. doi:10.1016/S0065-2423(06)43005-5
114. Hudelson MG, Jones JP (2006) J Med Chem 49:4367. doi:10.1021/jm0601553
115. Jensen BF, Vind C, Padkjaer SB, Brockhoff PB, Refsgaard HH (2007) J Med Chem 50:501. doi:10.1021/jm060333s
116. Breiman L (2001) Mach Learn 45:5. doi:10.1023/A:1010933404324
117. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003) Chem Inf Comput Sci 43:1947. doi:10.1021/ci034160g
118. Lombardo F, Obach RS, Dicapua FM, Bakken GA, Lu J, Potter DM et al (2006) J Med Chem 49:2262. doi:10.1021/jm050200r
119. Zhang QY, Aires-de-Sousa J (2007) J Chem Inf Model 47:1. doi:10.1021/ci050520j
120. Bruce CL, Melville JL, Pickett SD, Hirst JD (2007) J Chem Inf Model 47:219. doi:10.1021/ci600332j
121. Sakiyama Y, Yuki H, Moriya T, Hattori K, Suzuki M, Shimada K, Honma T (2007) J Mol Graph Model 27. doi:10.1016/j.jmgm.2007.06.005
122. Breiman L (1996) Mach Learn 24:123
123. Schapire RE, Freund Y, Bartlett P, Lee WS (1998) Ann Statist 26:1651. doi:10.1214/aos/1024691352
124. Koike A (2006) SAR QSAR Environ Res 17:497. doi:10.1080/10629360600934168
125. Svetnik V, Wang T, Tong C, Liaw A, Sheridan RP, Song Q (2005) J Chem Inf Model 45:786. doi:10.1021/ci0500379
126. Ren S, Kim H (2003) J Chem Inf Comput Sci 43:2106. doi:10.1021/ci034092y
127. Ren S (2003) J Chem Inf Comput Sci 43:1679. doi:10.1021/ci034046y
128. Xu QS, Daszykowski M, Walczak B, Daeyaert F, de Jonge MR, Heeres J et al (2004) Chemom Intell Lab Syst 72:27. doi:10.1016/j.chemolab.2004.02.007
129. Xu QS, Daeyaert F, Lewi PJ, Massart DL (2006) Chemom Intell Lab Syst 82:24. doi:10.1016/j.chemolab.2005.07.005
130. Put R, Xu QS, Massart DL, Vander Heyden Y (2004) J Chromatogr A 1055:11. doi:10.1016/j.chroma.2004.07.112
131. Choua SM, Leeb TS, Shaoc YE, Chen IF (2004) Expert Syst Appl 27:133. doi:10.1016/j.eswa.2003.12.013
132. Deconinck E, Xu QS, Put R, Coomans D, Massart DL, Vander Heyden Y (2005) J Pharm Biomed Anal 39:1021. doi:10.1016/j.jpba.2005.05.034

133. Deconinck E, Ates H, Callebaut N, Van Gyseghem E, Vander Heyden Y (2007) J Chromatogr A 1138:190. doi:10.1016/j.chroma.2006.10.068
134. Deconinck E, Coomans D, Vander Heyden Y (2007) J Pharm Biomed Anal 43:119. doi:10.1016/j.jpba.2006.06.022
135. Goulon A, Picot T, Duprat A, Dreyfus G (2007) SAR QSAR Environ Res 18:141. doi:10.1080/10629360601054313
136. Schneider G, Coassolo P, Lavé T (1999) J Med Chem 42:5072. doi:10.1021/jm991030j
137. Zuegge J, Schneider G, Coassolo P, Lavé T (2001) Clin Pharmacokinet 40:553. doi:10.2165/00003088-200140070-00006
138. Lee S, Kim D (2007) Arch Pharm Res 30:182
139. Lee PH, Cucurull-Sanchez L, Lu J, Du YJ (2007) J Comput Aided Mol Des [Epub ahead of print]
140. Molnar L, Keseru GM (2002) Bioorg Med Chem Lett 12:419. doi:10.1016/S0960-894X(01)00771-5
141. Ekins S, Berbaum J, Harrison RK (2003) Drug Metab Dispos 31:1077. doi:10.1124/dmd.31.9.1077
142. Wanchana S, Yamashita F, Hashida M (2003) Pharm Res 20:1401. doi:10.1023/A:1025702009611
143. Merkwirth C, Mauser H, Schulz-Gasch T, Roche O, Stahl M, Lengauer T (2004) J Chem Inf Comput Sci 44:1971. doi:10.1021/ci049850e
144. Kriegl JM, Arnhold T, Beck B, Fox T (2005) QSAR Comb Sci 24:491. doi:10.1002/qsar.200430925
145. Arimoto R, Prasad MA, Gifford EM (2005) J Biomol Screen 10:97. doi:10.1177/1087057104274091
146. Zhou DS, Liu RF, Otmani SA, Grimm SW, Zauhar RJ, Zamora I (2007) Lett Drug Des Discov 4:192. doi:10.2174/157018007780077462
147. Susnow RG, Dixon SL (2003) J Chem Inf Comput Sci 43:1308. doi:10.1021/ci030283p
148. O'Brien SE, de Groot MJ (2005) J Med Chem 48:1287. doi:10.1021/jm049254b
149. Chohan KK, Paine SW, Mistry J, Barton P, Davis AM (2005) J Med Chem 48:5154. doi:10.1021/jm048959a
150. Burton J, Ijjaali I, Barberan O, Petitet F, Vercauteren DP, Michel A (2006) J Med Chem 49:6231. doi:10.1021/jm060267u
151. Manga N, Duffy JC, Rowe PH, Cronin MT (2005) SAR QSAR Environ Res 16:43. doi:10.1080/10629360412331319871
152. Terfloth L, Bienfait B, Gasteiger J (2007) J Chem Inf Model 47:1688. doi:10.1021/ci700010t
153. Jones JP, Shou M, Korzekwa KR (1996) Adv Exp Med Biol 387:355
154. Cruciani G, Carosati E, De Boeck B, Ethirajulu K, Mackie C, Howe T et al (2005) J Med Chem 48:6970. doi:10.1021/jm050529c
155. Zhou D, Afzelius L, Grimm SW, Andersson TB, Zauhar RJ, Zamora I (2006) Drug Metab Dispos 34:976
156. de Graaf C, Oostenbrink C, Keizers PH, van der Wijst T, Jongejan A, Vermeulen NP (2006) J Med Chem 49:2417. doi:10.1021/jm0508538
157. Sheridan RP, Korzekwa KR, Torres RA, Walker MJ (2007) J Med Chem 50:3173. doi:10.1021/jm0613471