Available online at www.sciencedirect.com

ScienceDirect

Journal of Molecular Graphics and Modelling 26 (2008) 907–915

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

Predicting human liver microsomal stability with machine learning techniques

Yojiro Sakiyama [a,*], Hitomi Yuki [b], Takashi Moriya [a], Kazunari Hattori [b], Misaki Suzuki [c], Kaoru Shimada [c], Teruki Honma [b]

[a] *Research Planning and Coordination, Nagoya Laboratories, Pfizer Global Research and Development, Pfizer Inc., 5-2 Taketoyo, Aichi 470-2393, Japan*
[b] *Medicinal Chemistry Technologies, Nagoya Laboratories, Pfizer Global Research and Development, Pfizer Inc., Japan*
[c] *Pharmacokinetics Dynamics Metabolism, Nagoya Laboratories, Pfizer Global Research and Development, Pfizer Inc., Japan*

Received 23 March 2007; received in revised form 22 June 2007; accepted 22 June 2007
Available online 27 June 2007

Abstract

To ensure a continuing pipeline in pharmaceutical research, lead candidates must possess appropriate metabolic stability in the drug discovery process. *In vitro* ADMET (absorption, distribution, metabolism, elimination, and toxicity) screening provides us with useful information regarding the metabolic stability of compounds. However, before the synthesis stage, an efficient process is required in order to deal with the vast quantity of data from large compound libraries and high-throughput screening. Here we have derived a relationship between the chemical structure and its metabolic stability for a data set of in-house compounds by means of various *in silico* machine learning such as random forest, support vector machine (SVM), logistic regression, and recursive partitioning. For model building, 1952 proprietary compounds comprising two classes (stable/unstable) were used with 193 descriptors calculated by Molecular Operating Environment. The results using test compounds have demonstrated that all classifiers yielded satisfactory results (accuracy > 0.8, sensitivity > 0.9, specificity > 0.6, and precision > 0.8). Above all, classification by random forest as well as SVM yielded kappa values of approximately 0.7 in an independent validation set, slightly higher than other classification tools. These results suggest that nonlinear/ensemble-based classification methods might prove useful in the area of *in silico* ADME modeling.

Keywords: Random forest; *In silico*; ADMET; Metabolic stability; Machine learning

1. Introduction

The aim of pharmaceutical research and development is to ensure a continuing pipeline of new chemical entities displaying high therapeutic efficacy with few or no side effects. Failure to find promising lead candidates in the drug discovery process can lead to a heavy damage in the current business environment [1]. Above all, poor ADMET properties of compounds could be a major cause of attrition in drug development [2,3]. To avoid costly late-stage failure, *in vitro* ADMET screening has been used as an important part of the early stages in the drug discovery process. Specifically, metabolic stability screening is widely conducted so that appropriate metabolic stability of lead candidates is characterized in advance.

On the other hand, recent substantial development in combinatorial chemistry and in the supply of compounds from private ventures has dramatically increased the library of compounds. To maximize the opportunity of identifying as many lead compounds as possible and for optimizing them, pharmaceutical industries have made large investments in high throughput screening (HTS) applicable to biological screening and/or *in vitro* ADMET screening. Many work-flows/procedures, however, are employed in the process, such as managing the compound library, manufacturing high-density plates, constructing assay systems and processing the data, as well as making full use of robotics and information technology support. Although maximizing the efficiency of HTS is the key to fully utilizing these advances, the limitations of reducing cycle time and the number of screening compounds have also

* Corresponding author. Tel.: +81 569 74 4653; fax: +81 569 74 4678.
E-mail address: Yojiro.Sakiyama@pfizer.com (Y. Sakiyama).

1093-3263/$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2007.06.005

been recognized. To cope with such limitations, *in silico* ADMET screening is now being considered an essential paradigm, ideally before the synthesis stage, to extract meaningful computational results from vast quantities of data.

Following these trends, the use of machine learning on *in silico* ADMET has gained considerable interest in drug discovery. Prediction of metabolic stability by machine learning was first described by Bursi et al. in 2001, for a data set of 32 in-house steroidal androgens using the C5.0 decision tree [4]. Prediction of the metabolic turnover rate in human S9 homogenate has also been reported for 631 proprietary compounds using the kNN QSPR method [5]. Although these methods produced satisfactory results, their applicability may be limited to the particular chemical class and the particular metabolic endpoints evaluated. Recent state-of-the-art classifiers such as random forest or support vector machines (SVMs) have also been used to predict cytochrome P450 (CYP) mediated metabolism [6–15]. However, the accurate prediction is still challenging due to the complexity of CYP–ligand interaction (which could partly be due to the overlap of action between CYP isoenzymes), as well as various CYP gene polymorphism [16–18]. Another reason is that most studies applying machine learning to the prediction of drug metabolism dealt with a relatively small compound collection, with the exception of one approach reported by Arimoto et al. [6,19]. Ideally, thousands of experimental data from a wide variety of compounds would be necessary to obtain a satisfactory model.

It would be highly beneficial for medicinal chemists to provide data for overall metabolic stability of compounds in the early stage of drug discovery, rather than a detailed metabolic process such as CYP-mediated pathway. As previously mentioned, it is very difficult to predict the CYP-mediated metabolic process, and even more difficult to interpret the overall metabolic process due to its complexity. However, we believe that it may be possible to predict it if we utilize the most advanced machine learning technique with a large compound set. Based on these perspectives, the present study was undertaken to develop computational techniques to predict the overall human microsomal stability by various machine learning techniques based on *in vitro* experimental data collected from a relatively large number (approximately 2000) of compounds. The classification techniques we applied include the following machine learning algorithms: random forest, support vector machine, logistic regression, and recursive partitioning.

2. Methods

2.1. Metabolic stability assay

An incubation mixture was prepared consisting of liver microsomes (0.78 mg protein/ml), substrate (1.0 μM) and $MgCl_2$ (3.3 mM) in a potassium phosphate buffer (100 mM, pH 7.4). Reactions were initiated with the addition of NADPH (1.0 mM) and kept in a shaking water bath at 37 °C. During the reaction, aliquots were collected and added to termination mixtures containing internal standards (ISTD) at 0, 10, 30 and 60 min. The samples were centrifuged and the supernatant was subjected to HPLC-MS analysis.

In the determination of the *in vitro* half life ($T_{1/2}$), the analyte/ISTD peak height ratios were converted to a percentage of drug remaining, using the $T = 0$ peak height ratio values as 100%. The slope of the linear regression from log percentage remaining versus incubation time relationships ($-k$) was used in the conversion to *in vitro* $T_{1/2}$, values by *in vitro* $T_{1/2} = -0.693/k$. The value of *in vitro* intrinsic clearance (CLint, ml/min/kg) was calculated according to the following formula proposed by Obach et al. [20]:

$$\mathrm{CLint} = \frac{0.693}{in\,vitro T_{1/2}} \times \frac{\text{ml incubation}}{\text{mg microsomes}} \times \frac{45\ \text{mg microsomes}}{\text{gm liver}} \times \frac{21\ \text{gm liver}}{\text{kg b.w.}}$$

2.2. Data sets and descriptors

Overall, 2439 compounds were prepared from the Pfizer proprietary compound library. These compounds included more than 13 chemical series that were synthesized for multiple drug discovery projects. Among these compounds, 487 test compounds were randomly selected, and the remaining 1952 compounds were used for training. The training set was used for model building, and the test set was used for validation, with 193 descriptors calculated by MOE 2005.06 (Molecular Operating Environment, Chemical Computing Group Inc., Montreal, Canada). MOE descriptors include various 2D descriptors such as volume, shape, atom and bonds count, Kier-Hall connectivity, adjacency, partial charges, etc. (see Appendix A for details). These descriptors (independent variables) were standardized so that they have a mean of zero and a variance of 1.0.

The compound sets were classified by two groups, metabolically "stable" and "unstable" compounds (see

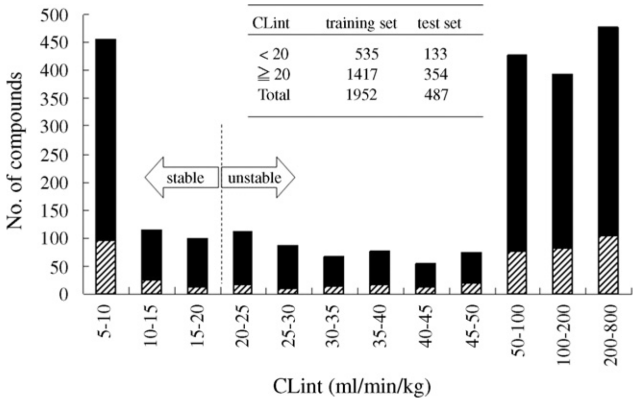

Fig. 1. Distribution of the experimental data of *in vitro* intrinsic clearance (CLint, ml/min/kg) used in the present study. Data were originally obtained from human liver microsomal assay calculated according to the equation described in the methods section. The black area represents the training set of 1952 compounds, and the hatched area represents the test set of 487 compounds. The inset table lists the data set composition. Threshold for classification (stable, unstable) was defined at CLint = 20 ml/min/kg.

Fig. 1 for details). The criterion for their classification was set at CLint = 20 ml/min/kg, because the number of commercially available drugs commonly used have their values less than 20 ml/min/kg [21].

To assess the diversity of the datasets, we compared their chemical space with that of drug-like molecules by principal component analysis (PCA). As representatives of drug-like molecules, we selected compounds that are clinically studied at phase I or later and are compliant with Lipinski's rule of five [22] from the MDL drug data report (MDDR) database [23]. PCA scores were calculated with the use of 50 important MOE 2D descriptors listed in Appendix A. In addition, we assessed the diversity of the datasets using parameters included in Lipinski's rules [22].

2.3. Classification and performance evaluation

Machine learning algorithms were originally developed in fields far from pharmaceutical research. The algorithms are especially useful for data mining in large databases to automatically discover patterns or rules. It is also useful for deriving models for problems where the underlying mechanism is very complex. There are two different types of machine learning. The first is unsupervised learning, where the property to be modeled is not employed during the training process. The second method is supervised learning, which is what we have used here, where a model is created from training data consisting of input objects and desired outputs, where the output of a model can predict a class label of the input object (called classification). The predicted results can be obtained from each classifier, basically as a confusion matrix (Table 1), consisting of true positive, true negative, false positive and false negative classifications. From these results, the following measures were calculated according to the equations listed in Table 1 to evaluate classification performance:

- Accuracy: probability to correctly classify compounds.
- Sensitivity: probability to predict positive (unstable) when true state is positive.
- Specificity: probability to predict negative (stable) when true state is negative.
- Positive precision: probability to correctly classify compounds predicted to be positive.
- Negative precision: probability to correctly classify compounds predicted to be negative.
- Balanced accuracy: average of positive and negative precision.
- Kappa: true accuracy (agreement by chance is corrected). More than 0.4 is desirable [24].
- Matthews correlation coefficient: overall accuracy of the prediction [25,26].

To validate the classification performance of each model, we have used 10-fold cross-validation, in which the data set was split into 10-folds; onefold used for testing, and the rest for training. This procedure was repeated 10 times, so all data were used as test data once, and finally these outputs were averaged.

The classification was conducted in the computing and statistical package R [27], with the use of three packages, randomForest (for random forest), e1071 (for support vector machine) and rpart (for recursive partitioning).

2.4. Random forest

Random forest [28] is one of the ensemble machine learning techniques which include boosting [29,30] and bagging [31]. It is an extension of the recursive partitioning [32,33]. In the random forest procedure, n samples are randomly drawn with replacement from the original data (i.e. take a bootstrap sample). These samples represent n training sets to construct n trees ($n$tree). For each node of the tree, m variables ($m$try) are

Table 1
Measures that can be calculated from a confusion matrix

Measure	Calculation
Accuracy	$\frac{A + D}{A + B + C + D}$
Sensitivity (Recall)	$\frac{D}{C + D}$
Specificity	$\frac{A}{A + B}$
Positive Precision	$\frac{D}{B + D}$
Negative Precision	$\frac{A}{A + C}$
Balanced Accuracy	$\frac{1}{2}\left(\frac{A}{A + C} + \frac{B}{B + D}\right)$
Kappa	$\frac{\text{Accuracy} - E}{1 - E}$ $E = \frac{(A + C)(A + B) + (B + D)(C + D)}{(A + B + C + D)^2}$
Matthews correlation coefficient	$\frac{AD - BC}{\sqrt{(A + B)(A + C)(B + D)(C + D)}}$

		Predicted	
		Stable	Unstable
Observed	Stable	A true negative	B false positive
	Unstable	C false negative	D true positive

randomly chosen from a total of 193 descriptors. The best split is calculated based on these *m* variables in each training set. Each tree is grown to the largest extent to achieve as much node homogeneity as possible. The importance of the variables as listed in Appendix A is measured based on the node homogeneity (mean decrease in Gini index). The predictions from the ensemble of trees are then aggregated to predict new data by the majority of votes. Although random forest is not widely used in the literature, it has several desirable characteristics, such as its accuracy, robustness to noise, simplicity of fine-tuning parameters, etc. In this study we used random forest to classify compounds on the preliminary defined condition (*n*tree = 500, *m*try = 13).

2.5. Support vector machine

Support vector machine (SVM) is another machine learning technique based on the statistical learning theory. A full description for how to use SVM for classification has been described by Vapnik [34]. SVMs create a separating hyperplane in the descriptor space of the training data, and compounds are classified on the basis of what side of the hyperplane they are located. The advantage of the SVM is that, by use of the so-called ‘‘kernel trick’’, the distance between a molecule and the hyperplane can be calculated in a transformed (nonlinear) feature space, without requiring the explicit transformation of the original descriptors. A variety of kernels have been suggested thus far [35]. The radial basis function kernel (Gaussian kernel) which is the most commonly used was applied to this study. The kernel function is expressed as follows:

$$K(\vec{x},\vec{x}_i) = \exp\left(-\frac{\left\|\vec{x}-\vec{x}_i\right\|^2}{2\sigma^2}\right)$$

In the above equation, the kernel width parameter σ controls the amplitude of the Gaussian function reflecting the generalization ability of SVM. The regularization parameter C controls the tradeoff between maximizing the margin and minimizing the training error. In this study, the optimal value of accuracy was found at $\sigma = 5$ when σ varied between 3 and 7. We also found no changes in the results when C was varied within the range of 10–10,000. To ensure the learning process was stable, these parameters were set on the specified condition ($\sigma = 5$, $C = 10{,}000$).

2.6. Other classification methods

Logistic regression can be used to predict dichotomous-dependent variables on the basis of continuous and/or categorical independents [36]. More specifically, the probability of a certain event occurring or not is defined by the following equation:

$$\text{Prob} = \frac{1}{1+\exp(\beta_0+\beta_1x_1+\beta_2x_2+\cdots)}$$

The coefficient is determined with maximum likelihood estimation so that the compounds with Prob of 0.5 or greater are classified as unstable compounds, and the compounds with Prob less than 0.5 are classified as stable compounds.

Recursive partitioning [32,33] is one of the decision tree methods. Here we used CART (classification and regression tree) which is a form of binary recursive partitioning. The term ‘‘binary’’ implies that each group represented by a ‘‘node’’ in a decision tree, can only be split into two groups. To form the classification tree, CART repeatedly partitions or splits the study set into separate nodes. To determine which variable to use for each split, CART examines all possible binary splits of the sample by each candidate predictor. CART then selects the predictor (and its particular dichotomization) and splits the sample into smaller and more homogeneous binary subgroups. The Gini impurity criterion that measures and ranks the extent to which each split departs from complete homogeneity was used for this purpose. For each split, CART selected the variable with the lowest impurity score [37].

3. Results and discussion

Fig. 1 shows the distribution of experimental data used in this study. The black area represents the training set ($n = 1952$) and the hatched area on the bottom represents the test set ($n = 487$). The number of the two classes (unstable/stable) distributed is shown in the inset table. The ratios between the classes in the training set and in the test set were reasonably close, 0.726/0.274 and 0.727/0.273, respectively.

Fig. 2 shows the two-dimensional PCA scores plot of our datasets and MDDR drug-like compounds. More than 52% of cumulative contribution rate could be ensured by the two principal components. Compounds of our datasets shared more than half of the distribution of MDDR drug-like molecules. This indicates that our datasets have acceptable diversity to build predictive models for general drug-like molecules in terms of their descriptors. Fig. 2 also indicates that both the training set and the test set are distributed in almost the same areas, which suggest that these data sets were randomly selected with no bias. Fig. 3 shows the distribution of our dataset against chemical properties; molecular weight, $c\log P$, number of hydrogen bond donors and acceptors, number of rotational bond and polar surface area. The results also demonstrate that the distribution of out dataset, as well as that of general drug-like molecules, is not biased in terms of their chemical properties according to the Lipinski’s rules [22].

The main purpose of this study was to evaluate the four classifiers on their ability to distinguish between ‘‘stable’’ and ‘‘unstable’’ compounds in the validation test set. Fig. 4 clearly shows that the classification of the test set by each classifiers yielded satisfactory results; accuracy > 0.8, sensitivity > 0.9, specificity > 0.6. Recursive partitioning produced relatively lower specificity than the other classifiers. Specificity of each classifier calculated by 10-fold cross-validation was as follows; random forest: 0.64 ± 0.06; SVM: 0.56 ± 0.06; logistic regression: 0.66 ± 0.08; recursive partitioning: 0.61 ± 0.06. Thus the lower value of specificity by recursive partitioning could be due to the variance of validation. Fig. 5 shows the precision (probability to correctly classify compounds

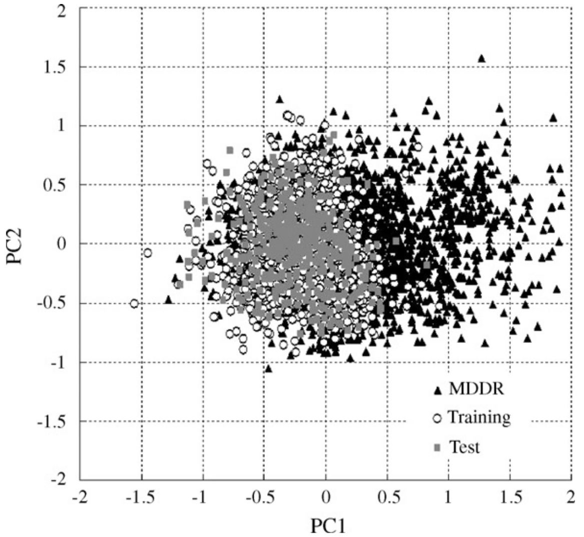

Fig. 2. PCA scores plot to illustrate the diversity of 2439 Pfizer proprietary compound datasets in terms of the MOE 2D descriptors, in comparison with 2003 MDDR drug-like molecules. The full black triangle, the open circle and the full gray square denote MDDR molecules, training set and test set, respectively.

predicted) of positive and negative compounds using the four classifiers. All of the classifiers yielded satisfactory results; positive precision > 0.8, negative precision > 0.7, balanced accuracy (average of positive and negative precision) > 0.8. Logistic regression yielded relatively lower negative precision than the other classifiers. The number of negative (stable) compounds was relatively small (approximately 27% of the total compounds) as shown in Fig. 1. In such case, it is suggested that simple generalized linear models such as logistic regression would be statistically less advantageous than other classifiers thus leading to lower performance of classification.

Table 2 summarizes the overall predictability by comparing two measures between four classifiers, kappa value and Matthews correlation coefficient (MCC). Kappa value is a measure of true (corrected) accuracy which takes into account the agreement that may have occurred by chance. It is considered to be preferable when it is larger than 0.4 [24]. MCC is a measure of how the normalized variables tend to have the same sign and magnitude. MCC = −1 for total disagreement, +1 for total agreement and 0 for completely random prediction [26]. Thus the kappa value and MCC are thought to be appropriate measures to evaluate the overall performance of the classifiers. As shown in Table 2, all classifiers yielded satisfactory results of kappa values and MCC. Above all, classification by random forest as well as SVM yielded predictive capability with the use of an independent validation set (kappa = 0.70–0.71). It is also worth noting that standard deviation on the results of cross-validation was small enough with all of the classifiers. The minimal and maximal values (min, max) of kappa among each of the 10-fold cross-validated results for each classifier were as follows; (0.54, 0.67) in random forest, (0.42, 0.63) in SVM, (0.03, 0.67) in logistic regression and (0.52, 0.63) in recursive partitioning, respectively. Although SVM shows a relatively higher kappa value in the independent validation set than that in the cross-validation, this presumably could have occurred by chance, considering the variation of the cross-validated results described above. Random forest resulted in models showing slightly higher predictive capacity (statistically significant at $p < 0.01$ by unpaired t-test)

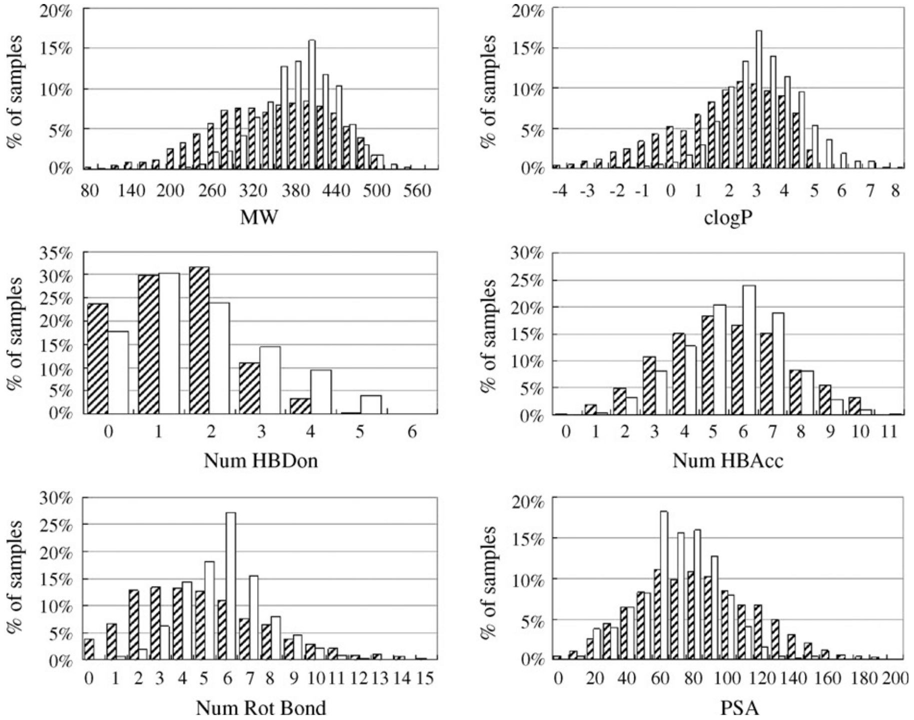

Fig. 3. Distribution of the physicochemical properties of the present data set in terms of Lipinski's rules. The properties include molecular weight (MW), calculated log P (c log P), number of hydrogen bond donors and acceptors (NumHBDon, NumHBAcc), number of rotational bond (NumRotBond) and the polar surface area (PSA). The hatched bar represents MDDR molecules, and the open bar represents the present data set. Values are expressed as a percentage of the number of samples against total number of data set.

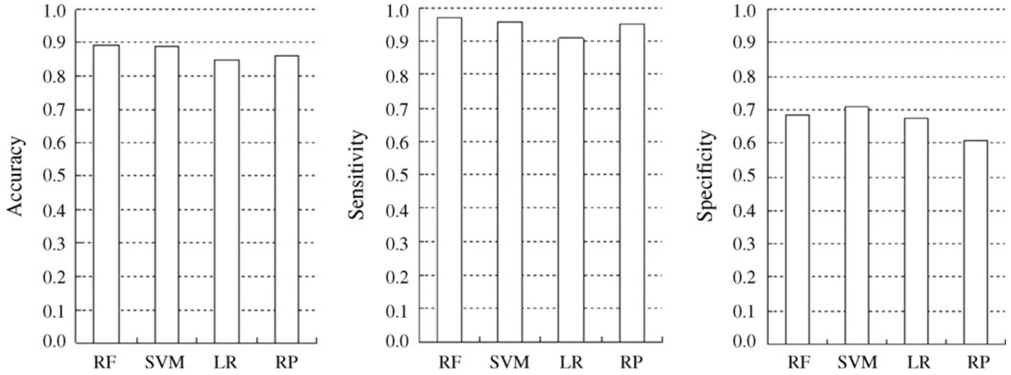

Fig. 4. Comparison of the prediction accuracy, sensitivity, and specificity using the following classifiers: random forest (RF), support vector machine (SVM), logistic regression (LR), and recursive partitioning (RP).

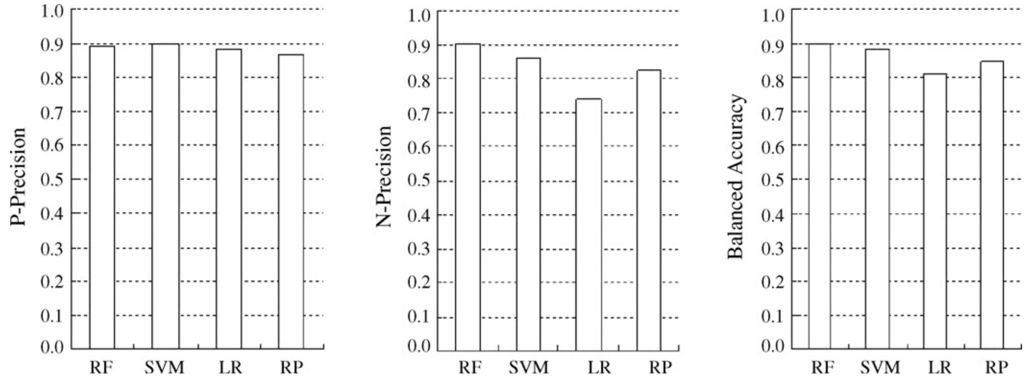

Fig. 5. Comparison of the overall performance of precision, positive precision, negative precision and balanced accuracy (average of posi and nega precision), using the following classifiers: random forest (RF), support vector machine (SVM), logistic regression (LR), and recursive partitioning (RP).

than SVM. These results suggest that nonlinear/ensemble machine learning has a slightly higher performance and reliability than other classifiers. As shown in Appendix A, descriptors that exhibited a significant decrease of Gini index (ΔGini) by random forest, were mostly physicochemical descriptors such as surface areas or $S\log P$, suggesting that physicochemical properties, particularly in relation to solubility or lipophilicity are key factors in explaining the dependent variables. The metabolic fate of a compound depends on a large number of variables which are related to both the chemical itself (chemical structure, physicochemical properties, etc.) and the biological system (enzyme and its environment). In such cases it seems to be better suited to use the most advanced machine learning tools such as random forest or SVM rather than linear methods.

Table 2
Comparison of overall predictability by different statistical learning methods

Method	10-Fold cross-validation			Independent validation set	
		Kappa	MCC	Kappa	MCC
Random forest	Mean	0.62**	0.63**	0.71	0.72
	S.D.	0.05	0.05		
SVM	Mean	0.52	0.53	0.70	0.71
	S.D.	0.06	0.06		
Logistic regression	Mean	0.51	0.52	0.60	0.60
	S.D.	0.19	0.19		
Recursive partitioning	Mean	0.58*	0.58*	0.61	0.62
	S.D.	0.05	0.05		

$*p < 0.05$, $**p < 0.01$.Significantly different compared to SVM group by Student's *t*-test.

Although large data sets with thousands of compounds were available in this study, they are still proprietary. This makes it difficult to fully assess the models. On the other hand, the use of data sets compiled from literature are also valuable, especially to utilize information of compound structures. However, these data sets also have a weakness in that the assays are not comparable, which may decrease the quality of the resulting models.

The application of random forest for compound classification was first described by Svetnik et al. [38], and the method has been gradually applied for prediction in various fields, including CYP2D6 inhibition [14], protein interactions [39], aqueous solubility [40], mutagenicity [41], volume distribution [42], doping in sport [43], disease marker [44], EEG [45], tumor classification [46], etc. This study is the first application of random forest in the prediction of metabolic stability using such a large number of compounds. The models are useful in identifying compounds with the potential risk of metabolically unstable conditions, which could be taken into consideration early in the discovery process, thus avoiding the costs of late-stage failure.

4. Conclusions

The present investigation using various machine learning techniques has demonstrated satisfactory results in predicting metabolic stability. Above all, the results when using an independent validation set indicated that random forest as well as SVM yielded a slightly more predictive capability than other classification methods. This suggests that nonlinear/ensemble-based machine learning techniques have the potential to predict overall metabolic stability, which is currently very difficult to predict due to its complicated mechanisms. These machine learning techniques can be used to assess ADMET properties of large compound sets for which no experimental testing can be performed due to capacity reasons. Moreover, these techniques can predict the properties of compounds prior to synthesis, driving medicinal chemistry into a promising part of chemical space.

Acknowledgements

The authors would like to acknowledge Prof. Yoshimasa Takahashi, Toyohashi University of Technology, for valuable advice with this article. We would also like to mention our appreciation to the members of Pharmacokinetics Dynamics Metabolism for the development and execution of human liver microsomal assays.

Appendix A

Molecular descriptors used in the present study were calculated by MOE. The decrease of Gini index (ΔGini), indicating the importance of descriptors in the random forest model, are listed below in a descending order, in which the top 50 of 193 descriptors were selected here.

Symbol	ΔGini	Description	Symbol	ΔGini	Description
PEOE_VSA-6	21.67	Total negative 6 vdw surface area	logP(o/w)	7.25	Log octanol/water partition coefficient
TPSA	17.82	Topological Polar Surface area (A**2)	PEOE_RPC	7.12	Relative negative partial charge
a_acid	17.14	Number of acidic atoms	SMR_VSA1	7.08	Bin 1 SMR (0.110, 0.260)
SlogP	17.08	Log Octanol/Water partition coefficient	PEOE_VSA_POL	7.00	Total polar 1 vdw surface area
a_aro	16.02	Number of aromatic atoms	PEOE_VSA-3	6.82	Total negative 3 vdw surface area
vsa_acid	15.80	VDW acidic surface area (A**2)	PEOE_VSA_FPP OS	6.80	Fractional polar positive vdw surface area
b_ar	15.17	Number of aromatic bonds	Slog_P_VSA4	6.71	Bin 4 SlogP (0.10, 0.15)
SlogP_VSA0	13.98	Bin 0 SlogP (-10-0.40)	PEOE_VSA_PNE G	6.56	Total polar negative 1 vdw surface area
SlogP_VSA7	13.24	Bin 7 SlogP (0.25, 0.30)	PEOE_VSA_NEG	6.46	Total negative 1 vdw surface area
PEOE_PC	11.98	Total negative partial charge	VDistEq	6.14	Vertex distance equality index
vsa_acc	10.58	VDW acceptor surface area (A**2)	PEOE_VSA_POS	5.98	Total positive vdw surface area
SMR_VSA3	9.32	Bin 3 SMR (0.350, 0.390	Slog_P_VSA8B	5.93	Bin 8 SlogP (0.30, 0.40)
vsa_hyd	8.89	VDW hydrophobe surface area (A**2)	SMR	5.81	Molar refractivity
PEOE_VSA_FPNEG	8.85	Fractional Polar negative vdw surface area	chi0v_C	5.79	Carbon valence connectivity index (order 0)
chi0_c	8.80	Carbon connectivity index (order 0)	a_ICM	5.68	Atom information content (mean)
PEOE_PC +	8.15	Total positive partial charge	balabanJ	5.67	Balaban averaged distance sum connectivity
SMR_VSA0	8.03	Bin 0 SMR (0.000, 0.110)	chil_C	5.46	Carbon connectivity index (order 1)
PEOE_VSA-1	7.90	Total negative 1 vdw surface area	PEOE_VSA-0	5.43	Total negative 0 vdw surface area
SlogP_VSA1	7.86	Bin 1 slogP (-0.40, -0.20)	kS_ssCH$_2$	5.36	Kier Atom Type E-state Sum (ssCH$_2$)
PEOE_VSA_FPOS	7.86	Fractional positive vdw surface area	PEOE_VSA_HYD	5.35	Total vdw hydrophobic surface area
PEOE_VSA_FHYD	7.70	Fractional hydrophobic vdw surface area	kS_aaCH	5.32	Kier Atom Type E-state Sum (aaCH)
PEOE_VSA_FPOL	7.69	Fractional Polar vdw surface area	PEOE_VSA + 0	5.22	Total positive 0 vdw surface area
PEOE_VSA_FNEG	7.53	Fractional negative vdw surface area	Q_VSA_HYD	5.14	Total hydrophobic vdw surface area
density	7.31	Mass density (AMU/A**3)	mr	5.12	Molar refractivity
PEOE_RPC +	7.28	Relative positive partial charge	PEOE_VSA + 1	5.11	Total positive 1 vdw surface area

Appendix B. A full table of results including the exact numbers of true and false positives and negatives in the independent validation set.

Classifier	True Positive	False Positive	True Negative	False Negative
Random Forest	344	42	91	10
SVM	339	39	94	15
Logistic Regression	322	43	90	32
Recursive Partitioning	337	52	81	17

References

[1] J. Langowski, A. Long, Computer systems for the prediction of xenobiotic metabolism, Adv. Drug Deliv. Rev. 54 (2002) 407–415.

[2] I. Kola, J. Landis, Can the pharmaceutical industry reduce attrition rates? Nat. Rev. Drug Discov. 3 (2004) 711–715.

[3] H. van de Waterbeemd, E. Gifford, ADMET *in silico* modelling: towards prediction paradise? Nat. Rev. Drug Discov. 2 (2003) 192–204.

[4] R. Bursi, M.E. de Gooyer, A. Grootenhuis, P.L. Jacobs, J. van der Louw, D. Leysen, (Q) SAR study on the metabolic stability of steroidal androgens, J. Mol. Graph. Model. 19 (552–556) (2001) 558–607.

[5] M. Shen, Y. Xiao, A. Golbraikh, V.K. Gombar, A. Tropsha, Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates, J. Med. Chem. 46 (2003) 3013–3020.

[6] R. Arimoto, M.A. Prasad, E.M. Gifford, Development of CYP3A4 inhibition models: comparisons of machine-learning techniques and molecular descriptors, J. Biomol. Screen. 10 (2005) 197–205.

[7] K.V. Balakin, S. Ekins, A. Bugrim, Y.A. Ivanenkov, D. Korolev, Y.V. Nikolsky, A.V. Skorenko, A.A. Ivashchenko, N.P. Savchuk, T. Nikolskaya, Kohonen maps for prediction of binding to human cytochrome P450 3A4, Drug Metab. Dispos. 32 (2004) 1183–1189.

[8] K.K. Chohan, S.W. Paine, J. Mistry, P. Barton, A.M. Davis, A rapid computational filter for cytochrome P450 1A2 inhibition potential of compound libraries, J. Med. Chem. 48 (2005) 5154–5161.

[9] S. Ekins, J. Berbaum, R.K. Harrison, Generation and validation of rapid computational filters for cyp2d6 and cyp3a4, Drug Metab. Dispos. 31 (2003) 1077–1080.

[10] J.M. Kriegl, T. Arnhold, B. Beck, T. Fox, A support vector machine approach to classify human cytochrome P450 3A4 inhibitors, J. Comput. Aided Mol. Des. 19 (2005) 189–201.

[11] C. Merkwirth, H. Mauser, T. Schulz-Gasch, O. Roche, M. Stahl, T. Lengauer, Ensemble methods for classification in cheminformatics, J. Chem. Inf. Comput. Sci. 44 (2004) 1971–1978.

[12] L. Molnar, G.M. Keseru, A neural network based virtual screening of cytochrome P450 3A4 inhibitors, Bioorg. Med. Chem. Lett. 12 (2002) 419–421.

[13] S.E. O'Brien, M.J. de Groot, Greater than the sum of its parts: combining models for useful ADMET prediction, J. Med. Chem. 48 (2005) 1287–1291.

[14] R.G. Susnow, S.L. Dixon, Use of robust classification techniques for the prediction of human cytochrome P450 2D6 inhibition, J. Chem. Inf. Comput. Sci. 43 (2003) 1308–1315.

[15] C.W. Yap, Y. Xue, Z.R. Li, Y.Z. Chen, Application of support vector machines to *in silico* prediction of cytochrome p450 enzyme substrates and inhibitors, Curr. Top Med. Chem. 6 (2006) 1593–1607.

[16] W.E. Evans, M.V. Relling, Pharmacogenomics: translating functional genomics into rational therapeutics, Science 286 (1999) 487–491.

[17] J.B. Houston, K.E. Kenworthy, *In vitro-in vivo* scaling of CYP kinetic data not consistent with the classical Michaelis–Menten model, Drug Metab. Dispos. 28 (2000) 246–254.

[18] S.A. Wrighton, E.G. Schuetz, K.E. Thummel, D.D. Shen, K.R. Korzekwa, P.B. Watkins, The human CYP3A subfamily: practical considerations, Drug Metab. Rev. 32 (2000) 339–361.

[19] T. Fox, J.M. Kriegl, Machine learning techniques for *in silico* modeling of drug metabolism, Curr. Top. Med. Chem. 6 (2006) 1579–1591.

[20] R.S. Obach, J.G. Baxter, T.E. Liston, B.M. Silber, B.C. Jones, F. MacIntyre, D.J. Rance, P. Wastall, The prediction of human pharmacokinetic parameters from preclinical and *in vitro* metabolism data, J. Pharmacol. Exp. Ther. 283 (1997) 46–58.

[21] R.S. Obach, Prediction of human clearance of twenty-nine drugs from hepatic microsomal intrinsic clearance data: an examination of *in vitro* half-life approach and nonspecific binding to microsomes, Drug Metab. Dispos. 27 (1999) 1350–1359.

[22] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Adv. Drug Deliv. Rev. 46 (2001) 3–26.

[23] The MDL Drug Data Report (MDDR) database 2005.2 is an online version of the Drug Data Report journal by Prous Science Publishers and is distributed by MDL Information Systems, Inc. Coverage: 1988–2005; updated monthly. Focus: Drugs launched or under development, as referenced in the patent literature, conference proceedings, and other sources; descriptions of therapeutic action and biological activity; tracking of compounds through development phases. Size: 164,647 molecules.

[24] M. Stokes, C. Davis, G. Koch, Observer agreement, in: M.E. Stokes, C.S. Davis, G.G. Koch (Eds.), Categorical Data Analysis Using the SAS System, 2nd ed., SAS Institute, Cary, NC, 1995, pp. 98–102.

[25] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta 405 (1975) 442–451.

[26] P. Baldi, S. Brunak, Y. Chauvin, C.A. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics 16 (2000) 412–424.

[27] R. Development Core Team, A Language and Environment for Statistical Computing, Foundation for Statistical Computing, Vienna, Australia, 2005 http://www.R-project.org.

[28] L. Breimann, Random forests, Mach. Learn. 45 (2001) 5–32.

[29] Y. Freund, R.E. Schapire, A decision-theoretic generalization of online learning and an application to boosting, J. Comput. Syst. Sci. 55 (1997) 119–139.

[30] R.E. Schapire, In a brief introduction to boosting, in: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, A Brief Introduction to Boosting, 1999.

[31] L. Breimann, Bagging predictors, Machine Learn. 24 (1996) 123–140.

[32] L. Breimann, Friedman. J.H., Olschen. R.A., Stone. C.J., Classification and Regression Trees, Wadsworth, 1984.

[33] D.M. Hawkins, S.S. Young, A.I. Rusinko, Analysis of large structure-activity data set using recursive partitioning, Quant. Struct. Acta Relat. 16 (1997) 296–302.

[34] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.

[35] N. Christianini, J. Shawe-Taylor, Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, Cambridge, MA, 2000.

[36] D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, Wiley, New York, 1989.

[37] L. Calvocoressi, M. Stolar, S.V. Kasl, E.B. Claus, B.A. Jones, Applying recursive partitioning to a prospective study of factors associated with adherence to mammography screening guidelines, Am J. Epidemiol. 162 (2005) 1215–1224.

[38] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, B.P. Feuston, Random forest: a classification and regression tool for compound classification and QSAR modeling, J. Chem. Inf. Comput. Sci. 43 (2003) 1947–1958.

[39] X.W. Chen, M. Liu, Prediction of protein–protein interactions using random decision forest framework, Bioinformatics 21 (2005) 4394–4400.

[40] D.S. Palmer, N.M. O'Boyle, R.C. Glen, J.B.O. Mitchell, Random forest models to predict aqueous solubility, J. Chem. Inf. Model. 47 (2007) 150–158.

[41] Q.Y. Zhang, J. Aires-de-Sousa, Random forest prediction of mutagenicity from empirical physicochemical descriptors, J. Chem. Inf. Model. 47 (2007) 1–8.

[42] F. Lombardo, R.S. Obach, F.M. Dicapua, G.A. Bakken, J. Lu, D.M. Potter, F. Gao, M.D. Miller, Y. Zhang, A hybrid mixture discriminant

analysis-random forest computational model for the prediction of volume of distribution of drugs in human, J. Med. Chem. 49 (2006) 2262–2267.

[43] E.O. Cannon, A. Bender, D.S. Palmer, J.B.O. Mitchell, Chemoinformatics-based classification of prohibited substances employed for doping in sport, J. Chem. Inf. Model. 46 (2006) 2369–2380.

[44] C. Strobl, A.L. Boulesteix, A. Zeileis, T. Hothorn, Bias in random forest variable importance measures: illustrations, sources and a solution, BMC Bioinformatics 8 (2007) 25.

[45] C. Lehmann, T. Koenig, V. Jelic, L. Prichep, R.E. John, L.O. Wahlund, Y. Dodge, T. Dierks, Application and comparison of classification algorithms for recognition of Alzheimer's disease in electrical brain activity EEG, J. Neurosci. Methods 161 (2007) 342–350.

[46] T. Shi, D. Seligson, A.S. Belldegrun, A. Palotie, S. Horvath, Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma, Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology Inc. 18 (2005) 547–557.