Pharm Res (2015) 32:2360–2371
DOI 10.1007/s11095-015-1629-y

RESEARCH PAPER

In Silico Predictions of Human Skin Permeability using Nonlinear Quantitative Structure–Property Relationship Models

Hiromi Baba • Jun-ichi Takahara • Hiroshi Mamitsuka

Received: 20 August 2014 / Accepted: 13 January 2015 / Published online: 24 January 2015

ABSTRACT

Purpose Predicting human skin permeability of chemical compounds accurately and efficiently is useful for developing dermatological medicines and cosmetics. However, previous work have two problems; 1) quality of databases used, and 2) methods for prediction models. In this paper, we attempt to solve these two problems.

Methods We first compile, by carefully screening from the literature, a novel dataset of chemical compounds with permeability coefficients, measured under consistent experimental conditions. We then apply machine learning techniques such as support vector regression (SVR) and random forest (RF) to our database to develop prediction models. Molecular descriptors are fully computationally obtained, and greedy stepwise selection is employed for descriptor selection. Prediction models are internally and externally validated.

Results We generated an original, new database on human skin permeability of 211 different compounds from aqueous donors. Nonlinear SVR achieved the best performance among linear SVR, nonlinear SVR, and RF. The determination coefficient, root mean square error, and mean absolute error of nonlinear SVR in external validation were 0.910, 0.342, and 0.282, respectively.

Conclusions We provided one of the largest datasets with purely experimental log k_p and developed reliable and accurate prediction models for screening active ingredients and seeking unsynthesized compounds of dermatological medicines and cosmetics.

KEY WORDS *in silico* prediction · quantitative structure–property relationship · random forest · skin permeability · support vector regression

ABBREVIATIONS

ALOGP	Ghose–Crippen octanol–water partition coefficient
ANN	Artificial neural network
C_d	Chemical concentration in dose formulation
J_{ss}	Steady state flux of the solute
K	Skin–vehicle partition coefficient
k_p	Permeability coefficient
L	Thickness of the skin
log P	Octanol–water partition coefficient
MAE	Mean absolute error
MW	Molecular weight
PCA	Principal component analysis
QSPR	Quantitative structure–property relationship
R^2	Determination coefficient
RF	Random forest
RMSE	Root mean square error
SVR	Support vector regression
SVR-G	Support vector regression with Gaussian (radial basis function) kernel
SVR-L	Support vector regression with linear kernel

INTRODUCTION

The skin is the human body's largest organ and vitally protects the body from xenobiotic invasion. Local and systemic drugs may also be administered through the skin. Currently, percutaneous absorption of chemicals is measured by various established *in vivo* (1) and *in vitro* (2) techniques. In particular, diffusion studies of excised human skin (3), animal skin (4), and artificial model membranes (5) have been widely

Electronic supplementary material The online version of this article (doi:10.1007/s11095-015-1629-y) contains supplementary material, which is available to authorized users.

H. Baba (✉) • J.-i. Takahara
Kyoto R&D Center, Maruho Co., Ltd., Shimogyo-ku, Kyoto, Japan
e-mail: baba_dfq@mii.maruho.co.jp

H. Mamitsuka
Bioinformatics Center, Institute for Chemical Research
Kyoto University, Uji, Kyoto, Japan

reported; these experiments have provided excellent indications of the permeability of skin to various chemicals. The skin permeability of a solute depends on several parameters, namely, chemical concentration in dose formulation (C_d), skin–vehicle partition coefficient (K), diffusion coefficient in the skin (D), and thickness of the skin (L). The permeability coefficient (k_p) quantifies the percutaneous absorption of chemicals through the skin defined as follows (6):

$$k_p = \frac{K \cdot D}{L} = \frac{J_{ss}}{C_d} \tag{1}$$

where J_{ss} is the steady state flux of the solute.

Measuring skin permeability of chemicals is generally time-consuming, because optimizing experimental conditions and building analytical methods for each chemical are also required in association with the permeation study. Moreover, unsynthesized compounds certainly can not be evaluated. When developing dermatological medicines and cosmetics for external use, an efficient and accurate *in silico* model of human skin permeability is useful, primarily because early-stage screening for active skin-penetrating ingredients substantially reduces product development costs. Under this background, several databases of permeability coefficients have been established (7–14), from which researchers have developed a lot of quantitative structure–property relationship (QSPR) models. However, there have been two problems in previous work; 1) quality of databases used, and 2) methods for generating prediction models.

The first problem – 1) quality of databases used – can be subdivided into two problems further; i) database size, and ii) uniformity of data.

i) Database size: some databases are particulary focused on specific chemical groups (15,16), including steroids, alcohols, and acids, resulting in around 30 compounds for such databases. Models based on these databases can work for predicting limited types of compounds, but are insufficient for predicting a wider range of chemicals.

ii) Uniformity of data: databases were frequently compiled from data collected under different experimental conditions, such as *in vitro* and *in vivo* conditions (7, 9–12), data using different membranes (human and animal skin or artificial membranes) (13), and coexisting predicted and measured data (9–12). Furthermore, absolute values of skin permeability differ among species (1,17) and under *in vivo* and *in vitro* conditions (18). Permeability can be changed by different experimental conditions, and no databases with consistent experimental conditions currently exist. This point casts a serious doubt whether reliable prediction models for the skin permeability of chemicals were built or not. A permeability dataset must be obtained under consistent experimental conditions.

The second problem is methods for construction of prediction models, *i.e.*, QSPR models. Table I summarises major existing models for predicting human skin permeability using comparatively large databases. Most of existing models for predicting human skin permeability were based on linear algorithms. Linear models are useful for interpreting the contributions of descriptors, but their prediction ability is relatively low. For example, the classic Potts and Guy model, which predicts the permeability coefficient (k_p) from the logarithm of the measured octanol–water partition coefficients (log P) and molecular weights (MW), given as follows:

$$log\ k_p = 0.71\ log\ P - 0.0061\ MW - 6.3 \tag{2}$$

Equation (2) gives 0.67 to the correlation coefficient between the observed and predicted values when the number of compounds is 93 (19). Nonlinear models for predicting skin permeability of chemicals were mainly artificial neural networks (ANNs) (20–22). ANNs are powerful and widely applied (23), while they are likely to overfit givendata and be trapped in local minima, and their network structures cannot be fully determined (24,25). These difficulty and the high computational cost of optimizing parameters of ANNs are disadvantageous for training the current model again, which is commonly required in training QSPR models when new data become available.

On the other hand, various promising nonlinear regression techniques, including support vector regression (SVR) (26) and random forest (RF) (27), have been developed and applied to QSPR models. SVR has at least three advantages (over ANNs): 1) unique global solutions, 2) avoiding the overfitting problem, and 3) lower computational cost. These advantages have rendered SVR an attractive option in various research fields (28,29). The RF algorithm assembles classification or regression trees. This method is generally robust to the overfitting problem and is one of the most high-performance learning algorithms (30–32). The performance of SVR and RF has yet to be evaluated in prediction models of human skin permeability of chemicals.

In this paper, we attempt to provide solutions to the above problems of conventional *in silico* models of human skin permeability. We first compiled a large dataset of 211 structurally diverse compounds through excised

Table I Previous Works

Information		Analysis conditions				All training		Cross-validation		Internal and external validation			
1st Author	Comments	N	3D optimization	The no. of descriptors	Method	Fitting ability		Predictability		Training set		Test set	
						R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Potts RO (19)	a few *in vivo* data	93	–	2	Linear	0.67	–	–	–	–	–	–	–
Cronin MT (63)	many calculated values	107	semi-empirical (AM1)	2	Linear	0.859	–	–	–	–	–	–	–
Buchwald P (64)	a few *in vivo* data	98	unknown	3	Linear	0.723	–	–	–	–	0.637	–	0.716
Patel H (10)	many calculated values	143	semi-empirical (AM1)	4	Linear	0.9	–	–	–	–	–	–	–
Lim CW (20)	a few *in vivo* data	92	semi-empirical (AM1)	4	ANN	–	0.528	–	0.669	–	–	–	–
Abraham MH (49)	–	119	–	5	Linear	0.832	–	–	–	–	–	–	–
Katritzky AR (21)	a few *in vivo* data	143	semi-empirical (AM1)	5	Linear	0.8	–	0.781	–	–	–	–	–
Katritzky AR (21)	a few *in vivo* data	143	semi-empirical (AM1)	41	Linear	0.907	–	0.812	–	–	–	–	–
Katritzky AR (21)	a few *in vivo* data	143	semi-empirical (AM1)	4	ANN	0.812		–	–	0.813	0.519	0.721	0.661
Neumann D (50)	–	110	*ab initio* (3-21G)	3	kNN	–	–	0.73	–	–	–	–	–
Basak SC (65)	a few *in vivo* data	101	*ab initio* (STO-3G)	unknown (PCA)	RR	–	–	0.729	–	–	–	–	–
Chen LJ (22)	many calculated values	<164	–	5	ANN	0.832	0.369	–	–	0.841	0.365	0.792	0.386
Neely BJ (51)	–	160	semi-empirical	10	ANN	0.997	0.036	–	–	0.93 (training + validation + test)			
Chauhan P (11)	many calculated values	208	unknown	12 (PCA)	PLS	–	–	–	–	0.755	0.518	0.936	0.267
Khajeh A (12)	many calculated values	283	semi-empirical (PM3)	3	ANFIS	–	–	–	–	0.899	0.312	0.89	0.317

N in Analysis Conditions means the number of different compounds in the database. AM1 and PM3 in 3D optimization column represent Austin Model 1 and Parameterized Model number 3 in semi-empirical method, respectively. 3-21G and STO-3G in 3D optimization column are among the basis set in quantum chemical calculation. kNN, RR, PLS, and ANFIS in Method column represents k-nearest neighbor algorithm, ridge regression, partial least squares, and adaptive neuro-fuzzy inference system, respectively

human skin with measured permeability coefficients. All data were collected from the literature and rigorously screened for deriving only from *in vitro* diffusion studies of excised human skin in the absence of permeation enhancement technologies or chemicals. We then applied sophisticated machine learning techniques (SVR and RF) to our database to develop reliable *in silico* prediction models.

MATERIALS AND METHODS

Data Collection

We collected the permeability coefficients of 211 compounds from aqueous donors in various literature reports. The dataset of permeability coefficients, along with the data sources, is shown in the Electronic supplementary material. The quality of predictive models largely depends on the quality of the used database. Ideally values should be obtained under the same experimental conditions with the same lab setup, but one laboratory could not amass a sufficiently large dataset of chemically diverse compounds. To minimize the effect of the experimental environment on the database quality, we used the data, which satisfy the following four criteria:

(a) data are obtained by an *in vitro* diffusion system, such as static or flow-through diffusion cells.
(b) the diffusion membrane is excised human skin.
(c) the donor solvent is an aqueous solution containing no organic solvents, which can affect skin permeation.
(d) no permeation enhancement technologies, such as iontophoresis, sonophoresis, or microneedles are used.

Some studies have reported the steady state flux and drug concentration of the formulation instead of the permeability coefficient. In these cases, the permeability coefficient was

obtained as the quotient of the steady state flux and formulated drug concentration, as specified in Eq. (1). When the permeability coefficients were not reported as numerical values but shown in only diagrams, the coefficeitnts were estimated from the diagrams.

Descriptor Generation

We generated 4803 numerical features on the three-dimensional (3D) molecular structure of each compound by using Dragon (v.6.0.32, Talete srl). In this step, the geometrical structures of all chemicals were optimized by two steps: 1) CS ChemBio3D (Ultra 13.0.2, PerkinElmer Inc.) with its molecular mechanics (MM2) feature, and 2) density functional theory optimization (33) using GAMESS (34,35) with the 6-31G(d, p) basis set and exchange potential of Becke and the correlation functional of Lee, Yang and Parr (B3LYP) (36,37).

Descriptors contained various constitutional, topological, molecular properties, functional group descriptors, weighted holistic invariant molecular (WHIM), and geometry, topology, and atom-weights assembly (GETAWAY) descriptors. Constant (and nearly constant) descriptors were eliminated, because they have no discriminative information. Descriptors containing errors or missing values were also removed, finally resulting in 2732 descriptors, which were used for model construction.

Chemical Space (Data Visualization)

The 211 compounds in our database cannnot be easily represented visually from 2732 descriptors. We showed a 3D coordinate system by observed permeability (log k_p) and the first two principal components of principal component analysis (PCA) as a chemical space to examine the diversity of compounds. PCA was implemented by the prcomp function of the stats R package (version 3. 0. 2) (38).

Descriptor Selection

Eliminating redundant descriptors from prediction models can reduce the practical computational cost of training the models. In this work, essential descriptors were selected by the method, which we call *stepwise forward selection* that repeats adding one descriptor which most improves the determination coefficient (R^2; see Eq. (3)) for internal cross-validation of the training set until no descriptors improve R^2 by more than 0.001.

Regression with Machine Learning Algorithms

To construct prediction models that combine SVR (or RF) with *stepwise forward selection*, we employed R (version 3.0.2) (38), which is widely used free software package for statistical and data mining research. We describe SVR and RF briefly below.

Support Vector Regression (SVR)

Support vector machine (SVM) is a classification algorithm, which has been widely used in machine learning and *in silico* prediction because of its remarkable versatility. The theory of SVM is detailed in several excellent sources (39,40). The key point of SVM is kernel transformation that is a projection of the descriptor matrix from the input space onto a high-dimensional feature space. This idea is applicable for solving regression problems (26), and this case is called support vector regression (SVR).

SVR is implemented by the svm function of the e1071 R package (version 1.6–2) (41), which supports linear and non-linear SVR. Available kernel functions are radial basis (Gaussian), polynomial, sigmoid, and linear. In this study, we used the linear and radial basis kernels (because such kernels are standard and well used) and the option called epsilon-type regression (eps-regression). We used default settings for all tunable parameters in the svm function.

Random Forest (RF)

RF can be used for both classification and regression (27). For regression, RF is an ensemble of regression trees built from bootstrapped samples of the training data. RF is a technique to solve high-dimensional non-linear problems and one of the most high-performance learning algorithms (30–32). The theory of RF is discussed in the literature (42).

RF regression was implemented by the randomForest function in the randomForest R package (version 4.6–7) (43). We set all tunable parameters at their default values.

Potts and Guy's model

The most well-used prediction model of skin permeability is the Potts and Guy model (19), which is shown in Eq. (2) and uses the Flynn's dataset (7). The permeability is predicted from only two variables (MW and log P) in this model.

We used the Potts and Guy model as a baseline method of nonlinear QSPR models. Since log P values were not reported for all chemicals in the dataset, all of

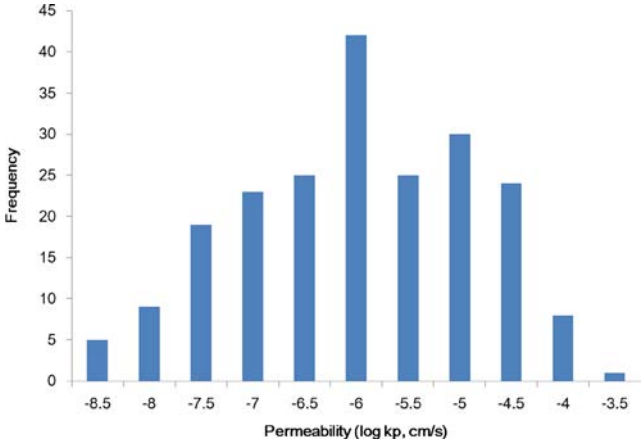

Fig. 1 Distribution of permeability (log k_ps) of compounds in our database.

them were substituted by the Ghose–Crippen octanol–water partition coefficients (ALOGP), which can be computed by Dragon.

Model Validation

The robustness and predictability of a QSPR model is assessed as follows: First, descriptors were selected by internal cross-validation, and then model predictability was evaluated by external validation. Of the 211 chemicals, 80% were randomly selected to be used for internal cross-validation. The remaining 20% were the test set in external validation. We conducted internal cross-validation by repeating 10-fold cross-validation ten times (44–46), in which the statistical parameters of model validity were averaged over ten iterations of the 10-fold cross-validation.

The model performance was evaluated by three measures: determination coefficient (R^2), root mean square error (RMSE), and mean absolute error (MAE). Note that we used R^2 only for our descriptor selection. Three measures are defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i^{obs} - y_i^{prd}\right)^2}{\sum_{i=1}^{n}\left(y_i^{obs} - \hat{y}^{obs}\right)^2},$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i^{obs} - y_i^{prd}\right)^2}{n}}, \tag{3}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}\left|y_i^{obs} - y_i^{prd}\right|}{n},$$

where y_i^{obs} and y_i^{prd} are the observed and predicted log k_p values, respectively, $\hat{y}^{obs}$ is the mean of the observed log k_p values, and n is the number of samples.

Furthermore for our discussion, we computed the following indices (47,48) which are recently used to assess

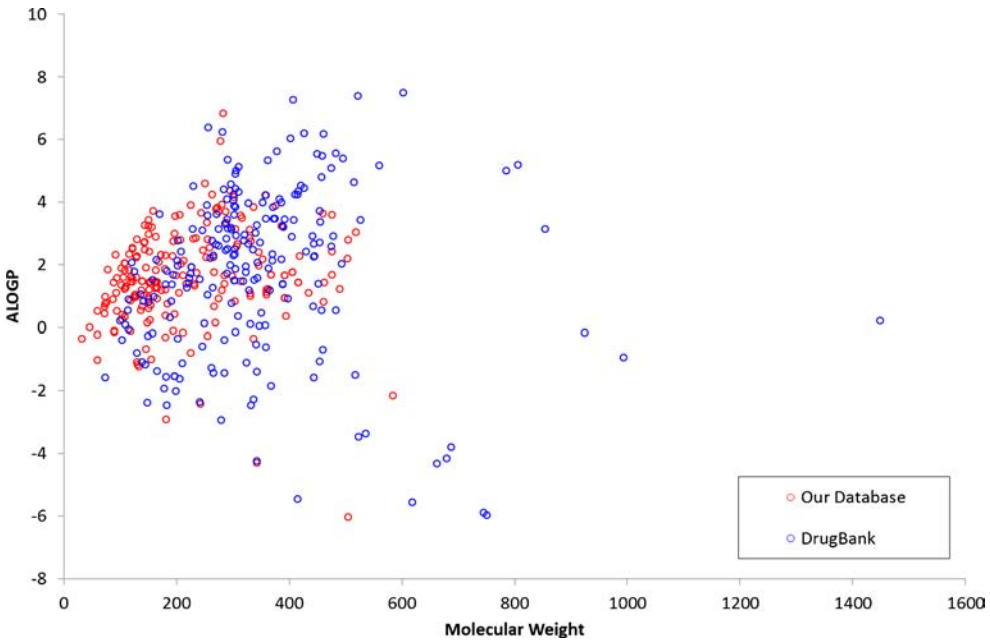

Fig. 2 Comparison of our database with DrugBank over molecular weight and ALOGP.

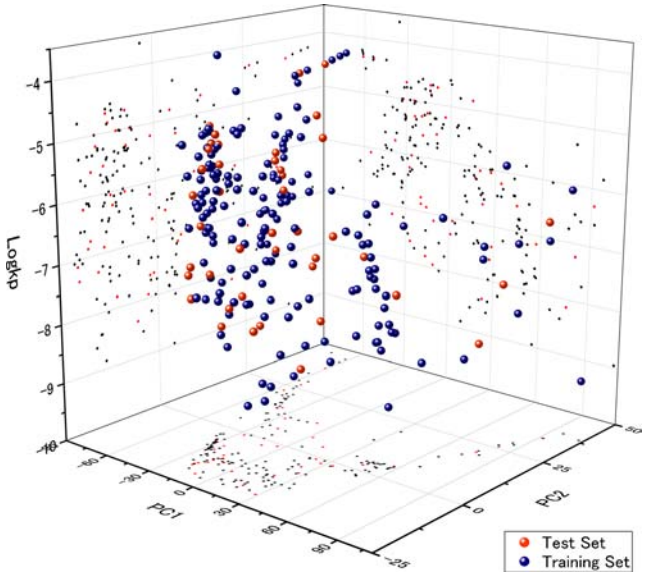

Fig. 3 Projection of the first two principal components and the permeability (observed log k_p) for the training set (*navy*) and the test set (*red*).

the validity and predictive power of the developed QSPR models:

$$k = \frac{\sum_{i=1}^{n}\left(y_i^{obs}y_i^{prd}\right)}{\sum_{i=1}^{n}\left(y_i^{prd}\right)^2},$$

$$k' = \frac{\sum_{i=1}^{n}\left(y_i^{obs}y_i^{prd}\right)}{\sum_{i=1}^{n}\left(y_i^{obs}\right)^2},$$

$$R_0^2 = 1-\frac{\sum_{i=1}^{n}\left(y_i^{prd}-ky_i^{obs}\right)^2}{\sum_{i=1}^{n}\left(y_i^{prd}-\hat{y}^{prd}\right)^2},$$

$$R_0^{2'} = 1-\frac{\sum_{i=1}^{n}\left(y_i^{obs}-k'y_i^{prd}\right)^2}{\sum_{i=1}^{n}\left(y_i^{obs}-\hat{y}^{obs}\right)^2},$$

$$R_m^2 = R^2\left(1-\sqrt{R^2-R_0^2}\right),$$

$$R_m^{2'} = R^2\left(1-\sqrt{R^2-R_0^{2'}}\right).$$

RESULTS AND DISCUSSION

Database Development

Our new database has human skin permeability data of 211 different compounds acquired from aqueous donors in various literature reports. We emphasize that this dataset can be one of the recent largest datasets (8,14,49–51) with purely experimental log k_p. All permeability coefficients meet the quality criteria stated above, and therefore the prediction models based on this new database can properly estimate the influence of molecular structures on human skin permeation. Compounds in this extensive database belong to various chemical classes, including alcohols, amines, amides, aromatics, carbonyls, carboxylic acids, esters, ethers, urea, halides, nitriles, and nitro compounds. Many of the compounds are active ingredients of pharmaceutical products, such as anti-inflammatory, anti-cancer, anti-HIV, local anesthetic, stimulants, and sleep-inducing drugs. Our data with permeability coefficients, molecular weights and data sources (references) are shown in the Electronic supplementary material.

Data Visualization

The logarithm of permeability coefficients (log k_p) in our database ranges from −8.96 to −3.64 (mean=−6.25; standard deviation=1.12), where k_p is expressed in cm/s. Figure 1 summarizes the distribution of log k_p. Figure 1 reveals that log k_p is well distributed like a normal distribution from almost non-permeable values like −8.5 to highly permeable values like −3.5.

Molecular weights range from 32.05 to 584.73 (mean=219.6; standard deviation=115.8), and ALOGPs range from −6.057 to 6.835 (mean=1.603; standard deviation=1.529). Figure 2 shows a comparison of 211 compounds in our database and 211 compounds which were randomly selected from DrugBank (52,53). Note that unlike our database, routes of administration of the drugs in DrugBank are not particularly limited. In Fig. 2 compounds are distributed over a 2D space of molecular weights and ALOGP. Molecular weights of compounds in our database tend to be lower than those of

Table II Coefficients of Determination (R^2), Root Mean Squares Error (RMSE), and Mean Absolute Error for QSPR Models

Algorithm	Type	The number of descriptors	Training - cross validation			Test - external validation		
			R^2	RMSE	MAE	R^2	RMSE	MAE
Potts and Guy	Linear	2	–	–	–	0.740	0.692	0.524
SVM - Linear	Linear	17	0.809	0.494	0.384	0.675	0.658	0.488
SVM - Gaussian	Nonlinear	11	0.867	0.423	0.339	0.910	0.342	0.282
RF	Nonlinear	9	0.856	0.448	0.341	0.884	0.390	0.319

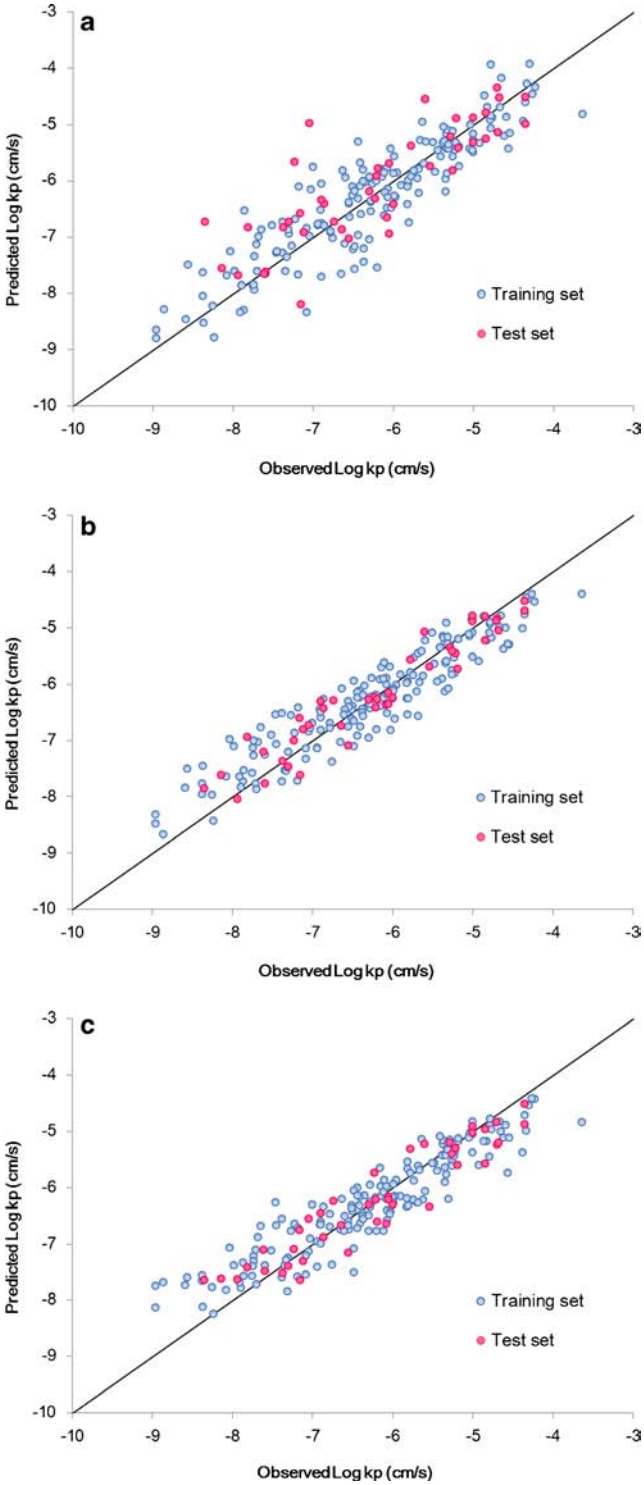

Fig. 4 (**a**). Predicted log k_p by SVM with linear kernel *vs.* experimental data. (**b**) Predicted log k_p by SVM with Gaussian kernel *vs.* experimental data. (**c**). Predicted log k_p by RF *vs.* experimental data.

DrugBank, and this result would be due to that smaller molecules are favorable for percutaneous absorption (54). Regarding ALOGP, both databases incline toward lipophilic compounds (ALOGP > 0), and especially 68% of chemical compounds in our database are within a range from 1 to 4. This result is consistent with the fact that drugs with log P from 1 to 4 are thought to be ideal for external use (55,56).

Figure 3 shows a chemical space, which is a 3D scatter plot of permeability and two principal components of PCA. In this figure, we show training (navy) and test (red) sets. Clearly, the diversity of permeability and structural characteristics in the training and test sets were very similar, indicating that compounds assigned to the test set were almost uniformly selected from the original chemical space. This implies that the test compounds cover most of the structural features in the dataset, enabling the predictability and validity of the prediction models to be statistically assessed by external validation.

Prediction Models and Comparison of Different Models

In our internal cross-validation, out of the original 2732 descriptors, we obtained 17, 11, and 9 descriptors by *stepwise forward selection* for SVR with linear kernel (SVR-L), SVR with Gaussian kernel (SVR-G), and RF, respectively. Table II shows the R^2 values, RMSEs and MAEs of SVR-L, SVR-G, and RF. In Table II, SVR-G and RF showed higher R^2 values and lower RMSEs and MAEs than SVR-L, in both external validation and internal cross-validation. Thus, the predictive ability of SVR-G and RF was higher than that of SVR-L. Figure 4 a, b, and c show the observed *vs.* predicted log k_p values for the training and test set of SVR-L, SVR-G, and RF, respectively. As a reference, the result of the Potts and Guy model for the test data are summarized in Table II and plotted in Fig. 5. The predictive ability ($R^2 = 0.740$) of the Potts and Guy model was lower than those of SVR-G and RF models. The major defect in the Potts and Guy model is that it is composed of only two explanatory variables, which are not completely independent. Moreover, Fig. 3 also reveals that skin permeability (log k_p) is distributed nonlinearly against the first or second principal components. These results show

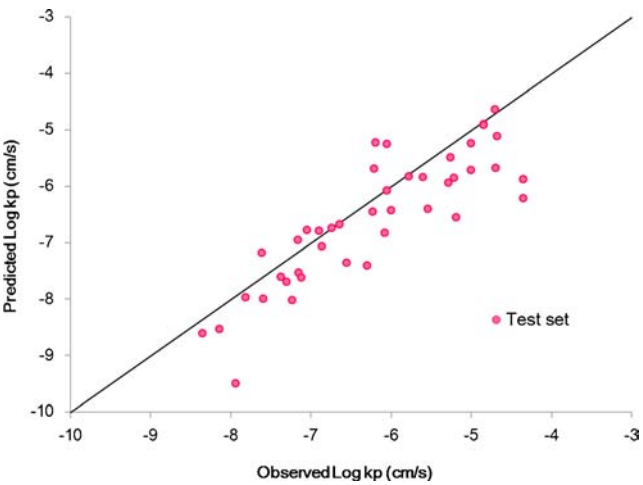

Fig. 5 Predicted log k_p by Potts and Guy's model *vs.* experimental data.

Table III Statistical Parameters of External Validation for QSPR Models

Algorithm	External validation							
	k	k'	R_0^2	R'^2_0	$\frac{R^2-R_0^2}{R^2}$	$\frac{R^2-R'^2_0}{R^2}$	R_m^2	R'^2_m
SVM - Linear	0.966	1.024	0.611	0.665	0.0948	0.0148	0.504	0.606
SVM - Gaussian	0.992	1.005	0.881	0.906	0.0319	0.0049	0.756	0.851
RF	0.999	0.997	0.831	0.876	0.0600	0.0090	0.680	0.807

that permeability cannot be linearly represented by the descriptors used in this study.

The performance and robustness of prediction models are an important points in QSPR studies. Golbraikh and Tropsha (47) and Roy (48) suggested the following criteria to validate QSPR models: A model is highly predictive if its statistical characteristics satisfy the following conditions in external validation:

$$0.85 \le k \le 1.15 \quad and \quad 0.85 \le k' \le 1.15,$$
$$\frac{R^2-R_0^2}{R^2} \le 0.10 \quad and \quad \frac{R^2-R_0^{2'}}{R^2} \le 0.10,$$
$$R_m^2 \ge 0.50 \quad and \quad R_m^{2'} \ge 0.50 .$$

The statistical characteristics of the linear and nonlinear prediction models are summarized in Table III. All statistics of prediction models with SVR-L, SVR-G, and RF satisfy the above three criteria, suggesting that all models in this study can reliably predict the skin permeability of wide range of compounds we used. Overall, SVR-G was most favorable for predicting permeability. SVR and RF are easily constructed and optimized, while the nonlinear regression performances of SVR and RF have not been tested in the human skin permeability predictions of chemicals. Therefore, this study presents a new and important application of SVR and RF.

Our database has experimental skin permeability coefficients determined from excised human skin *in vitro*. This database does not contain calculated permeability, data from other animals and data with chemical or physical penetration enhancement. Therefore, from this database, we can build models that directly detect the influence of molecular structures on human skin permeability. Our high-performance nonlinear prediction models will be reliable and useful tools for developing dermatological ingredients.

An important aspect of QSPR research is selecting descriptors for model construction. Descriptor selection can be performed by numerous methods, such as stepwise selection (57), genetic algorithms (58), ant colony optimization (59), and

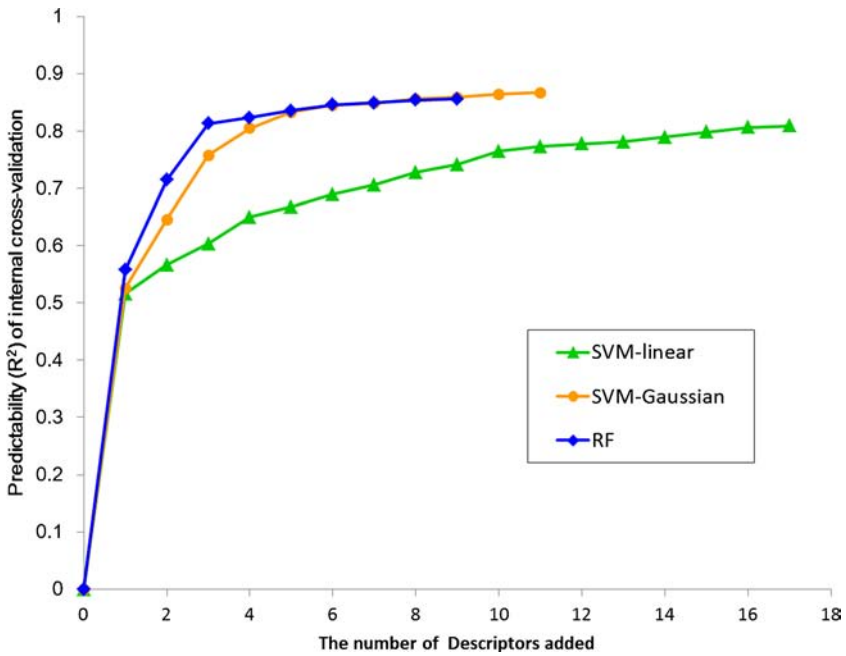

Fig. 6 Relationship between the number of descriptors and predictability (R^2) of internal cross-validation in descriptor selection.

Table IV Selected Descriptors for the SVM Model with Linear Kernel

Name	Description	Block
SM1_Dz(v)	spectral moment of order 1 from Barysz matrix weighted by van der Waals volume	2D matrix-based descriptors
GATS5e	Geary autocorrelation of lag 5 weighted by Sanderson electronegativity	2D autocorrelations
RDF110i	Radial Distribution Function - 110 / weighted by ionization potential	RDF descriptors
Eig10_EA(ed)	eigenvalue n. 10 from edge adjacency mat. weighted by edge degree	Edge adjacency indices
Mor32s	signal 32 / weighted by I-state	3D-MoRSE descriptors
B03[O-O]	Presence/absence of O - O at topological distance 3	2D Atom Pairs
nRCOOR	number of esters (aliphatic)	Functional group counts
MATS1m	Moran autocorrelation of lag 1 weighted by mass	2D autocorrelations
VE1_H2	coefficient sum of the last eigenvector from reciprocal squared distance matrix	2D matrix-based descriptors
RDF070s	Radial Distribution Function - 070 / weighted by I-state	RDF descriptors
RDF090i	Radial Distribution Function - 090 / weighted by ionization potential	RDF descriptors
RDF120s	Radial Distribution Function - 120 / weighted by I-state	RDF descriptors
TDB09e	3D Topological distance based descriptors - lag 9 weighted by Sanderson electronegativity	3D autocorrelations
B06[C-N]	Presence/absence of C - N at topological distance 6	2D Atom Pairs
B04[C-O]	Presence/absence of C - O at topological distance 4	2D Atom Pairs
G2i	2nd component symmetry directional WHIM index / weighted by ionization potential	WHIM descriptors
ChiA_H2	average Randic-like index from reciprocal squared distance matrix	2D matrix-based descriptors

particle swarm optimization (60). In this study, the descriptor selection procedure is a greedy *stepwise forward selection* with repeated 10-fold cross-validation. Our results show that essential and important descriptors for permeability prediction were automatically selected to achieve high-performance prediction models without selecting descriptors in advance (*i.e.*, assuming no *a priori* relationships between the structural features of a chemical and its skin permeability). Figure 6 shows performance improvement by adding descriptors in our greedy *stepwise forward selection*. Although the stepwise procedure needs running algorithms many times, the entire procedure is efficient because the model construction with SVR and RF was computationally fast. Tables IV, V and VI show the selected descritpors of SVR-L, SVR-G, and RF, respectively. The importance of each of the selected descriptors for skin permeability is somewhat unclear due to the "black box" nature of the nonlinear prediction models. However, Tables IV, V and VI reveals that most of the selected descriptors were associated with polarity, including partition coefficients such as log P, and molecular volume. The importance of these descriptors for skin permeability has been pointed out in previous studies (61, 62), suggesting that distribution from the donor aqueous solution to oily stratum corneum is a significant factor in the percutaneous absorption process.

For the practical use of the proposed models, only three steps are required: geometry optimization of permeants of

Table V Selected Descriptors for the SVM Model with Gaussian Kernel

Name	Description	Block
SM1_Dz(v)	spectral moment of order 1 from Barysz matrix weighted by van der Waals volume	2D matrix-based descriptors
H3p	H autocorrelation of lag 3 / weighted by polarizability	GETAWAY descriptors
ALOGP	Ghose-Crippen octanol-water partition coeff. (logP)	Molecular properties
TDB05v	3D Topological distance based descriptors - lag 5 weighted by van der Waals volume	3D autocorrelations
DLS_05	modified drug-like score from Zheng *et al.* (2 rules)	Drug-like indices
MATS1p	Moran autocorrelation of lag 1 weighted by polarizability	2D autocorrelations
R7s	R autocorrelation of lag 7 / weighted by I-state	GETAWAY descriptors
Eig04_AEA(ed)	eigenvalue n. 4 from augmented edge adjacency mat. weighted by edge degree	Edge adjacency indices
CATS2D_06_LL	CATS2D Lipophilic-Lipophilic at lag 06	CATS 2D
B05[O-O]	Presence/absence of O - O at topological distance 5	2D Atom Pairs
Eta_beta_A	eta average VEM count	ETA indices

Table VI Selected Descriptors for the RF Model

Name	Description	Block
nHAcc	number of acceptor atoms for H-bonds (N,O,F)	Functional group counts
ALOGP	Ghose-Crippen octanol-water partition coeff. (logP)	Molecular properties
R4p	R autocorrelation of lag 4 / weighted by polarizability	GETAWAY descriptors
SpMaxA_EA(ri)	normalized leading eigenvalue from edge adjacency mat. weighted by resonance integral	Edge adjacency indices
R1m	R autocorrelation of lag 1 / weighted by mass	GETAWAY descriptors
BLTF96	Verhaar Fish base-line toxicity from MLOGP (mmol/l)	Molecular properties
Mor28i	signal 28 / weighted by ionization potential	3D-MoRSE descriptors
E2s	2nd component accessibility directional WHIM index / weighted by I-state	WHIM descriptors
SpMaxA_B(s)	normalized leading eigenvalue from Burden matrix weighted by I-State	2D matrix-based descriptors

interest (and in our database), generation of selected descriptors (Tables IV, V and VI) based on optimized structures of permeants, and the application of support vector regression without parameter tuning on R to the data matrix composed of experimental permeability coefficients and descriptors. The proposed models require only one additional step (*i.e.*, geometry optimization) compared with earlier prediction models, such as the Potts and Guy model, but requires the same number of steps when compared with the recent prediction models based on the 3D structures of permeants (Table I).

CONCLUSION

We first presented a new large dataset of 211 structurally diverse compounds with human skin permeability coefficients. The dataset was compiled from various literature sources by screening examples which are not obtained by *in vitro* diffusion study. Therefore, this database is highly suitable for evaluating the relationships between molecular characteristics and human skin permeability. It is potentially applicable to related research such as vehicle effects on skin permeability by using it combined with a dataset of skin permeability for solutions other than water. We are working on compiling a skin permeability dataset composed of a wide variety of permeants and solvents to develop the prediction models of vehicle effects on skin permeability in order to optimize topical formulations.

We then applied sophisticated machine learning techniques (SVR-L, SVR-G and RF) to this database to develop high-performance *in silico* prediction models. All SVR-L, SVR-G, and RF were computationally fast. In both external validation and internal cross-validation, SVR-G and RF achieved higher performance than SVR-L. The performance statistics between SVR-G and RF were not significantly different, but SVR-G was slightly better than RF. The descriptors used in this work were all computationally obtained. As with most previous prediction models such as the Potts and Guy relationship, the proposed models are limited to the prediction of skin permeability from aqueous donor solution, which differs from real formulations. Nevertheless, the proposed models are valuable to evaluate the potency of a wide variety of compounds from their chemical structures, particularly in early-stage screening for active skin-penetrating ingredients. Overall, we provide a time- and cost-efficient approach of screening active ingredients, and this approach is applicable to as-yet unsynthesized compounds in dermatological medicines and cosmetics.

REFERENCES

1. Bartek MJ, LaBudde JA, Maibach HI. Skin permeability *in vivo*: comparison in rat, rabbit, pig and man. J Investig Dermatol. 1972;58(3):114–23.
2. Franz TJ. Percutaneous absorption on the relevance of *in vitro* data. J Investig Dermatol. 1975;64(3):190–5.
3. Zhang Q, Grice JE, Li P, Jepps OG, Wang GJ, Roberts MS. Skin solubility determines maximum transepidermal flux for similar size molecules. Pharm Res. 2009;26(8):1974–85.
4. Takeuchi H, Ishida M, Furuya A, Todo H, Urano H, Sugibayashi K. Influence of skin thickness on the *in vitro* permeabilities of drugs through Sprague-Dawley rat or Yucatan micropig skin. Biol Pharm Bull. 2012;35(2):192–202.
5. Karadzovska D, Riviere JE. Assessing vehicle effects on skin absorption using artificial membrane assays. Eur J Pharm Sci. 2013;50(5):569–76.
6. Blank IH, McAuliffe DJ. Penetration of benzene through human skin. J Investig Dermatol. 1985;85(6):522–6.
7. Flynn GL. Physicochemical determinants of skin absorption. In: Gerrity TR, Henry CJ, editors. Principles of route-to-route extrapolation for risk assessment. 1st ed. New York: Elsevier; 1990. p. 93–127.
8. Wilschut A, ten Berge WF, Robinson PJ, McKone TE. Estimating skin permeation. The validation of five mathematical skin permeation models. Chemosphere. 1995;30(7):1275–96.
9. Kirchner LA, Moody RP, Doyle E, Bose R, Jeffery J, Chu I. The prediction of skin permeability by using physicochemical data. ATLA. 1997;25:359–70.
10. Patel H, ten Berge W, Cronin MT. Quantitative structure-activity relationships (QSARs) for the prediction of skin permeation of exogenous chemicals. Chemosphere. 2002;48(6):603–13.

11. Chauhan P, Shakya M. Role of physicochemical properties in the estimation of skin permeability: *in vitro* data assessment by partial least-squares regression. SAR QSAR Environ Res. 2010;21(5–6): 481–94.
12. Khajeh A, Modarress H. Linear and nonlinear quantitative structure-property relationship modelling of skin permeability. SAR QSAR Environ Res. 2014;25(1):35–50.
13. Moss GP, Sun Y, Wilkinson SC, Davey N, Adams R, Martin GP, *et al.* The application and limitations of mathematical modelling in the prediction of permeability across mammalian skin and polydimethylsiloxane membranes. J Pharm Pharmacol. 2011;63(11): 1411–27.
14. Vecchia BE, Bunge AL. Skin absorption databases and predictive equations. In: Guy R, Hadgraft J, editors. Transdermal drug delivery. 2nd ed. New York: Marcel Dekker; 2003. p. 57–141.
15. Roberts MS, Pugh WJ, Hadgraft J, Watkinson AC. Epidermal permeability-penetrant structure relationships: 1. An analysis of methods of predicting penetration of monofunctional solutes from aqueous solutions. Int J Pharm. 1995;126(1–2):219–33.
16. Ghafourian T, Fooladi S. The effect of structural QSAR parameters on skin penetration. Int J Pharm. 2001;217(1–2):1–11.
17. Panchagnula R, Stemmer K, Ritschel WA. Animal models for transdermal drug delivery. Methods Find Exp Clin Pharmacol. 1997;19(5):335–41.
18. Lehman PA, Raney SG, Franz TJ. Percutaneous absorption in man *in vitro-in vivo* correlation. Skin Pharmacol Physiol. 2011;24(4):224–30.
19. Potts RO, Guy RH. Predicting skin permeability. Pharm Res. 1992;9(5):663–9.
20. Lim CW, Fujiwara S, Yamashita F, Hashida M. Prediction of human skin permeability using a combination of molecular orbital calculations and artificial neural network. Biol Pharm Bull. 2002;25(3):361–6.
21. Katritzky AR, Dobchev DA, Fara DC, Hür E, Tämm K, Kurunczi L, *et al.* Skin permeation rate as a function of chemical structure. J Med Chem. 2006;49(11):3305–14.
22. Chen LJ, Lian GP, Han LJ. Prediction of human skin permeability using artificial neural network (ANN) modeling. Acta Pharmacol Sin. 2007;28(4):591–600.
23. Patel J. Science of the science, drug discovery and artificial neural networks. Curr Drug Discov Technol. 2013;10(1):2–7.
24. Castillo E, Fontenla-Romero O, Guijarro-Berdiñas B, Alonso-Betanzos A. A global optimum approach for one-layer neural networks. Neural Comput. 2002;14(6):1429–49.
25. El-Sebakhy EA. Forecasting PVT properties of crude oil systems based on support vector machines modeling scheme. J Petrol Sci Eng. 2009;64(1–4):25–34.
26. Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput. 2004;14(3):199–222.
27. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
28. Wang Y, Zheng M, Xiao J, Lu Y, Wang F, Lu J, *et al.* Using support vector regression coupled with the genetic algorithm for predicting acute toxicity to the fathead minnow. SAR QSAR Environ Res. 2010;21(5–6):559–70.
29. Yap CW, Li ZR, Chen YZ. Quantitative structure-pharmacokinetic relationships for drug clearance by using statistical learning methods. J Mol Graph Model. 2006;24(5):383–95.
30. Chu A, Ahn H, Halwan B, Kalmin B, Artifon EL, Barkun A, *et al.* A decision support system to facilitate management of patients with acute gastrointestinal bleeding. Artif Intell Med. 2008;42(3):247–59.
31. Monte-Moreno E. Non-invasive estimate of blood glucose and blood pressure from a photoplethysmograph by means of machine learning techniques. Artif Intell Med. 2011;53(2):127–38.
32. Hsieh CH, Lu RH, Lee NH, Chiu WT, Hsu MH, Li YC. Novel solutions for an old disease: diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks. Surgery. 2011;149(1):87–93.
33. Robert GP, Yang W. Density-functional theory of atoms and molecules. Oxford: Oxford University Press; 1989.
34. Schmidt MW, Baldridge KK, Boatz JA, Elbert ST, Gordon MS, Jensen JH, *et al.* General atomic and molecular electronic structure system. J Comput Chem. 1993;14(11):1347–63.
35. Gordon MS, Schmidt MW. Advances in electronic structure theory: GAMESS a decade later. In: Dykstra CE, Frenking G, Kim KS, Scuseria GE, editors. Theory and applications of computational chemistry: the first forty years. Amsterdam: Elsevier; 2005. p. 1167–89.
36. Lee C, Yang W, Parr RG. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. Phys Rev B Condens Matter. 1988;37(2):785–9.
37. Beck AD. A new mixing of Hartree-Fock and local density-functional theories. J Chem Phys. 1993;98:1372–7.
38. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. 2013. Available from http://www.R-project.org/.
39. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97.
40. Vapnik V. Statistical learning theory. New York: Wiley; 1998.
41. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6–2. 2014. Available from http://CRAN.R-project.org/package=e1071/.
42. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci. 2003;43(6):1947–58.
43. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2:18–22.
44. Burman P. A Comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. Biometrika. 1989;76(3):503–14.
45. Shao J. Linear model selection by cross-validation. J Am Stat Assoc. 1993;88(442):486–94.
46. Zhang P. Model selection *via* multifold cross validation. Ann Stat. 1993;21(1):486–94.
47. Golbraikh A, Tropsha A. Beware of q2! J Mol Graph Model. 2002;20(4):269–76.
48. Roy PP, Roy K. On some aspects of variable selection for partial least squares regression models. QSAR Comb Sci. 2008;27(3):302–13.
49. Abraham MH, Martins F, Mitchell RC. Algorithms for skin permeability using hydrogen bond descriptors: the problem of steroids. J Pharm Pharmacol. 1997;49(9):858–65.
50. Neumann D, Kohlbacher O, Merkwirth C, Lengauer T. A fully computational model for predicting percutaneous drug absorption. J Chem Inf Model. 2006;46(1):424–9.
51. Neely BJ, Madihally SV, Robinson Jr RL, Gasem KA. Nonlinear quantitative structure-property relationship modeling of skin permeation coefficient. J Pharm Sci. 2009;98(11):4069–84.
52. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, *et al.* DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. 2014;42:D1091–7.
53. We have obtained structures of 6213 compounds in DrugBank: http://www.drugbank.ca/downloads#structures/.
54. Bos JD, Meinardi MM. The 500 Dalton rule for the skin penetration of chemical compounds and drugs. Exp Dermatol. 2000;9(3):165–9.
55. Yano T, Nakagawa A, Tsuji M, Noda K. Skin permeability of various non-steroidal anti-inflammatory drugs in man. Life Sci. 1986;39(12):1043–50.
56. Flynn GL, Yalkowsky SH. Correlation and prediction of mass transport across membranes. I. Influence of alkyl chain length on flux-determining properties of barrier and diffusant. J Pharm Sci. 1972;61(6):838–52.

57. González MP, Terán C, Teijeira M, Helguera AM. Quantitative structure activity relationships as useful tools for the design of new adenosine receptor ligands. 1. Agonist. Curr Med Chem. 2006;13(19):2253–66.
58. Leardi R, Boggia R, Terrile M. Genetic algorithms as a strategy for feature selection. J Chemom. 1992;6(5):267–81.
59. Shamsipur M, Zare-Shahabadi V, Hemmateenejad B, Akhond M. An efficient variable selection method based on the use of external memory in ant colony optimization. Application to QSAR/QSPR studies. Anal Chim Acta. 2009;646(1–2):39–46.
60. Lin WQ, Jiang JH, Shen Q, Shen GL, Yu RQ. Optimized block-wise variable combination by particle swarm optimization for partial least squares modeling in quantitative structure-activity relationship studies. J Chem Inf Model. 2005;45(2):486–93.
61. Barratt MD. Quantitative structure-activity relationships for skin permeability. Toxicol in Vitro. 1995;9(1):27–37.
62. Kasting GB, Smith RL, Cooper ER. Effect of lipid solubility and molecular size on percutaneous absorption. In: Shroot B, Schaefer H, editors. Skin pharmacokinetics. Basel: Kargar; 1987. p. 138–53.
63. Cronin MT, Dearden JC, Moss GP, Murray-Dickson G. Investigation of the mechanism of flux across human skin *in vitro* by quantitative structure-permeability relationships. Eur J Pharm Sci. 1999;7(4):325–30.
64. Buchwald P, Bodor N. A simple, predictive, structure-based skin permeability model. J Pharm Pharmacol. 2001;53(9):1087–98.
65. Basak SC, Mills D, Mumtaz MM. A quantitative structure-activity relationship (QSAR) study of dermal absorption using theoretical molecular descriptors. SAR QSAR Environ Res. 2007;18(1–2):45–55.