Pharm Res (2015) 32:3604–3617
DOI 10.1007/s11095-015-1720-4

RESEARCH PAPER

Modeling and Prediction of Solvent Effect on Human Skin Permeability using Support Vector Regression and Random Forest

Hiromi Baba[1,2] • Jun-ichi Takahara[1] • Fumiyoshi Yamashita[2] • Mitsuru Hashida[2,3]

Received: 18 March 2015 / Accepted: 19 May 2015 / Published online: 2 June 2015

ABSTRACT

Purpose The solvent effect on skin permeability is important for assessing the effectiveness and toxicological risk of new dermatological formulations in pharmaceuticals and cosmetics development. The solvent effect occurs by diverse mechanisms, which could be elucidated by efficient and reliable prediction models. However, such prediction models have been hampered by the small variety of permeants and mixture components archived in databases and by low predictive performance. Here, we propose a solution to both problems.

Methods We first compiled a novel large database of 412 samples from 261 structurally diverse permeants and 31 solvents reported in the literature. The data were carefully screened to ensure their collection under consistent experimental conditions. To construct a high-performance predictive model, we then applied support vector regression (SVR) and random forest (RF) with greedy stepwise descriptor selection to our database. The models were internally and externally validated.

Results The SVR achieved higher performance statistics than RF. The (externally validated) determination coefficient, root mean square error, and mean absolute error of SVR were 0.899, 0.351, and 0.268, respectively. Moreover, because all descriptors are fully computational, our method can predict as-yet unsynthesized compounds.

Conclusion Our high-performance prediction model offers an attractive alternative to permeability experiments for pharmaceutical and cosmetic candidate screening and optimizing skin-permeable topical formulations.

KEY WORDS *in silico* prediction · quantitative structure–property relationship · skin permeability · solvent effect · 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_p	Permeability coefficient
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

INTRODUCTION

The skin is the largest organ in the human body and provides an exquisite interface between the internal biological system

Electronic supplementary material The online version of this article (doi:10.1007/s11095-015-1720-4) contains supplementary material, which is available to authorized users.

✉ Hiromi Baba
baba_dfq@mii.maruho.co.jp

[1] Kyoto R&D Center, Maruho Co., Ltd. 93 Awata-cho, Chudoji Shimogyo-ku 600-8815, Kyoto, Japan

[2] Department of Drug Delivery Research, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29, Yoshida-shimoadachicho, Sakyo-ku, Kyoto 606-8501, Japan

[3] Institute for Integrated Cell-Material Sciences, Kyoto University 46-29, Yoshida-shimoadachicho, Sakyo-ku, Kyoto 606-8501, Japan

and the external environment. Because the skin also serves as a drug administration route (1), the skin permeability of chemicals is of significant interest in pharmaceutical and cosmetic industries. In particular, it is required for optimizing the effectiveness and assessing the toxicological risk of topical formulations.

Currently, the skin permeability of chemicals is measured by *in vivo*, *ex vivo*, and *in vitro* techniques (2, 3). In particular, diffusion studies using human, animal, or artificial skin have been widely reported (4–6). However, these experiments are generally time consuming and expensive; moreover, they cannot evaluate unavailable or as-yet unsynthesized compounds. During the early developmental stages of dermatological medicines or cosmetics, when numerous candidate compounds need to be evaluated, efficiently estimating the skin permeability of these compounds is particularly important. Therefore, various prediction models of skin permeability have been established (7–11), which provide efficient and accurate alternatives to animal experiments. Among the most famous classical models is the Potts and Guy's model (11), which predicts the permeability coefficient (k_p) by linear regression using octanol–water partition coefficients (log P) and molecular weights of permeants. Recently, *in silico* prediction models of skin permeability, based on machine learning algorithms and matrices containing the observed permeabilities and physicochemical features of permeants, have also been reported (12–18).

However, most of the above-mentioned models predict the skin permeability of chemicals in aqueous solutions, which differs from real formulations. Such models are beneficial for estimating the potency of permeants' skin permeability from their chemical structures or physicochemical features, particularly during early screening for active skin-penetrating ingredients, but are insufficient for optimizing the compositions of topical formulations. The latter must relate the skin permeability of the active ingredient to the effectiveness and toxicological risk of the ingredient. In practice, the skin penetration of permeants is strongly affected by the solvent effect, which is known to occur by a wide variety of mechanisms, for example, the adjustment of the permeant's thermodynamic activity, hydration of the stratum corneum, and modifications of the structured lipid domain within the stratum corneum, such as fluidization, altered polarity, phase separation, and lipid extraction (19, 20). In fact, chemical penetration is enhanced by a range of commonly used solvents including alcohols, glycols, amides, esters, fatty acids, sulfoxides, surfactants, and terpenes (21). Some studies have reported the effect of specific permeants or solvents on skin permeability (22–26), wherein the permeability coefficient of a permeant strongly depends on the selected solvent; the permeability coefficients of the same permeant in different solvents may differ by more than one hundred times. However, these models cannot be applied to diverse combinations of chemicals and solvents. The solvents and ingredients used in pharmaceutical and cosmetics cover an extensive range, and their experimental evaluation on a one-by-one basis consumes significant time and cost.

Very few studies have attempted a comprehensive prediction of the solvent effect on skin permeability. Recently, Riviere, Ghafourian, and their coworkers performed an *in vitro* study of porcine skin (27–30) rather than human skin. They evaluated the permeability of 16 permeants in 24 mixtures composed of 5 vehicle ingredients for 384 formulations. As explanatory variables for linear regression, they employed Abraham's parameters (9) or the physicochemical parameters weighted by the fractions of the components in the mixture. Riviere and Ghafourian's representative models are summarized in supplementary material. Although their studies are essentially important, they are unsuitable for industrial applications for two main reasons: the small variety of mixture components and the use of linear regression for generating prediction models.

The first issue—the small variety of mixture components—is the central problem limiting the applicability of prediction models. For example, the molecular weights of the 16 permeants in Riviere and Ghafourian's database are relatively small (ranging from 81.09 to 351.61); moreover, their database includes no carboxylic acids or nitriles, which are common functional groups of drugs. Five of the mixture components include no esters, fatty alcohols, amines, or polyethers. Models based on datasets with comparatively small chemical spaces have limited applicability (31). In particular, they are unsuitable for predicting an extensive range of combinations of industrial ingredients because the structural features of compounds cannot be extrapolated from a small chemical space with high prediction accuracy. For pharmaceutical and cosmetic use, prediction models should be based on a structurally diverse chemical dataset.

The second issue—generating prediction models by linear regression—relates to the suitability of the regression algorithm. Linear models are helpful for understanding the contributions of explanatory variables; however, skin permeability has been reported to be nonlinearly related to the physicochemical features of permeants (32). Moreover, solvents affect skin permeability by interacting with various small and organized biological compounds. Such interactions may not be linearly related to the structural features of the mixture components. In fact, previous models cannot predict the permeabilities of a 384-sample dataset to high accuracy (the determination coefficient between the observed and predicated values in an external validation of 384 samples was 0.61).

Machine learning algorithms such as artificial neural networks (ANNs) (33), support vector regression (SVR) (34), and random forest (RF) (35), which are closely related to artificial

intelligence, are promising techniques for solving high-dimensional nonlinear problems. These algorithms have already been applied to quantitative structure–property relationship (QSPR) models (36, 37). The SVR algorithm is particularly designed to find unique global solutions, avoiding the overfitting problem and providing better generalization than conventional ANNs (34, 38). RF is an ensemble learning method for classification and regression and is typically robust to noise and overfitting. RF also ranks among the most high-performance learning techniques (36, 39). These advantages of SVR and RF have secured their popularity in QSPR studies; however, neither algorithm has been employed in prediction models of the solvent effect on the human skin permeability of chemicals.

In this study, we solve the problems of previous prediction models by creating an original large dataset of human skin permeability that includes chemically diverse permeants and solvents. We also develop unprecedentedly reliable and accurate *in silico* models of the solvent effect on human skin permeability for industrial use, employing SVR and RF algorithms. Variables are selected by the stepwise forward process, which sequentially selects the best variables from a fully computational descriptor set. The generated models are evaluated by internal and external validation.

MATERIALS AND METHODS

Data Collection

From the literature, we collected the permeability coefficients ($\log k_p$) of 261 permeants from various mixtures of 31 solvents reported in *in vitro* diffusion studies on excised human skin. In total, 412 permeability data were accumulated. The dataset, which contains the permeability coefficients, permeants, components, and their fraction in each mixture, and data references, is provided in the electronic supplementary material.

Permeability coefficients can be affected by the experimental conditions and manipulation. Therefore, all data were carefully screened to satisfy the following four requirements for direct evaluation of the solvent effect on skin permeability:

(i) permeability coefficients are obtained by an *in vitro* diffusion system.
(ii) the diffusion membrane is excised human skin.
(iii) the skin is not pretreated with any vehicles or solvents except for aqueous solvents to hydrate the skin.
(iv) no physical or instrumental permeation enhancement technologies are employed.

When the permeability coefficient (k_p) was not reported, it was evaluated from the steady state flux (J_{ss}) and the concentration of the permeant in the mixture (C_d), defined as follows (40):

$$k_p = \frac{J_{ss}}{C_d}. \quad (1)$$

When the permeability coefficients were shown only in the scatter plot, they were estimated from the plot.

Descriptor Generation

A wide set of molecular descriptors was generated from the three-dimensional (3D) molecular structures of the permeants and solvents in their neutral forms. The geometry of all molecules was optimized by the molecular mechanics function (41) in CS ChemBio3D Ultra (ver. 13.0.2, PerkinElmer Inc.) and by subsequent semi-empirical molecular orbital methods using MOPAC2012 (ver. 12, Stewart Computational Chemistry) with a PM7 Hamiltonian (42).

Using MOPAC 2012, we generated 11 descriptors of quantum chemistry. Furthermore, we generated 4870 descriptors in 23 descriptor categories (such as molecular properties and 0–3 dimensional descriptors) using the Dragon (ver. 6.0.32, Talete srl) software. Constant descriptors and those containing errors or missing values were removed. To characterize each mixture composition, the solvent descriptors were average-weighted by the fraction of each component in the mixture, as implemented in Ghafourian *et al.* (29). If two explanatory descriptors were highly intercorrelated (*i.e.*, the correlation coefficient between them exceeded 0.95), the descriptor with the lower determination coefficient in the first iteration of the stepwise descriptor selection (described below) was removed.

Database Visualization

The data are visualized as distributions of permeability coefficients, which are objective variables. The structural diversity of the permeants and solvents in our databases, with their wide range of functional groups and skeletal structures, cannot be easily visualized. Instead, to enable comparison between the chemical features of our database and those of a drug database, we construct a two-dimensional (2D) scatter plot of the calculated octanol–water partition coefficients (ALOGP) *versus* the molecular weight (MW). The drug database comprises 261 drugs (MW < 1000) randomly selected from 6475 compounds in DrugBank (43, 44), because few macromolecules are evaluated as permeants in topical formulations.

Model Construction

The prediction models were constructed by combining machine learning algorithms with greedy stepwise descriptor

selection in an R (ver. 3.0.2) software environment, as implemented in our previous work (12).

Descriptor Selection

QSPR analysis typically generates more descriptors than required for constructing the prediction model, which increases the computational cost of prediction. To eliminate redundant descriptors, we employed our *greedy stepwise selection* algorithm, a greedy algorithm that iteratively adds the descriptor that most improves the determination coefficient (r^2; see Eq. (2)) in an internal cross-validation of the training set.

Support Vector Regression (SVR)

SVR is an extension of a pioneering classification algorithm called the support vector machine. SVR projects the descriptor matrix from the input space into a high-dimensional feature space *via* kernel functions. In this feature space linear regressions of nonlinear problems can be conducted. The theoretical background of SVR is detailed in an excellent source (34).

In the present study, SVR was implemented by the svm function in the e1071 R package (ver. 1.6–2) (45), which supports several kernel functions (radial basis (or Gaussian), polynomial, sigmoid, and linear). We employed Gaussian kernels, which are standard and widely used. All tunable parameters in the svm function were set to their default values.

Random Forest (RF)

RF is an ensemble of classification and regression trees constructed from bootstrapped samples of the training data using random feature selection in tree induction. In each bootstrap sample, each node is split by the best from a random set of selected descriptors, with each unpruned tree grown to its maximum extent. In RF regression, prediction is conducted by averaging the individual tree predictons. An excellent description of RF theory can be found in the literature (46).

RF regression was implemented by the randomForest function in the randomForest R package (ver. 4.6.7) (47). All tunable parameters in the randomForest function were set to their default values.

Model Validation

The model was validated by internal cross-validation and external validation. Our prediction model was constructed by randomly selecting 80% of the samples in our database (330 out of 412 samples) and assigning them to the training set. The remaining 20% (82 samples) were reserved as the test set for external validation. To confirm the adequate assignment of samples into two groups, the chemical space was represented as a 3D scatter plot of permeability and the first two components in the principal component analysis (PCA) of the training and test sets. PCA was implemented by the prcomp function of R (ver. 3.0.2).

The effect of unfavorable sampling was minimized by repeating 10-fold cross-validation ten times (48–50), which averages the statistical characteristics of the model over ten iterations of the 10-fold cross-validation. In this step, we evaluated the determination coefficient (r^2), root mean square error (RMSE), and mean absolute error (MAE). The determination coefficient (r^2) is an index of the above-described descriptor selection. The three statistics are defined as follows:

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

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

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}\left|y_i^{obs} - y_i^{prd}\right|}{n}, \tag{4}$$

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

Having selected the descriptors, we assessed the predictive performance of our prediction models by external validation. The predictive performance of QSPR models has been evaluated and discussed in several recent studies (51–57). Although these studies have proposed statistical parameters for estimating the external predictive performance, the parameters best suited to any datasets remain unconsolidated. In this step, we achieved a detailed evaluation by the following statistics (Eqs. (5)–(14)), which have been appreciated and employed in other QSPR studies (13, 58), as well as the standard statistics r^2, RMSE, and MAE:

$$k = \frac{\sum_{i=1}^{n}\left(y_i^{obs} y_i^{prd}\right)}{\sum_{i=1}^{n}\left(y_i^{prd}\right)^2}, \tag{5}$$

$$k' = \frac{\sum_{i=1}^{n}\left(y_i^{obs} y_i^{prd}\right)}{\sum_{i=1}^{n}\left(y_i^{obs}\right)^2}, \tag{6}$$

$$r_0^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i^{prd} - k y_i^{obs}\right)^2}{\sum_{i=1}^{n}\left(y_i^{prd} - \bar{y}^{prd}\right)^2}, \tag{7}$$

$$r_0'^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i^{obs} - k' y_i^{prd}\right)^2}{\sum_{i=1}^{n}\left(y_i^{obs} - \bar{y}^{obs}\right)^2}, \tag{8}$$

$$r_m^2 = r^2\left(1-\sqrt{r^2-r_0^2}\right), \tag{9}$$

$$r_m'^2 = r^2\left(1-\sqrt{r^2-r_0'^2}\right), \tag{10}$$

$$\overline{r_m^2} = \frac{r_m^2 + r_m'^2}{2}, \tag{11}$$

$$\Delta r_m^2 = \left|r_m^2 - r_m'^2\right|, \tag{12}$$

$$Q_{F3}^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i^{obs}-y_i^{prd}\right)^2/n}{\sum_{i=1}^{n_{TR}}\left(y_i^{obs}-\bar{y}_{TR}\right)^2/n_{TR}}, \tag{13}$$

$$CCC = \frac{2\sum_{i=1}^{n}\left(y_i^{obs}-\bar{y}^{obs}\right)\left(y_i^{prd}-\bar{y}^{prd}\right)}{\sum_{i=1}^{n}\left(y_i^{obs}-\bar{y}^{obs}\right)^2 + \sum_{i=1}^{n}\left(y_i^{prd}-\bar{y}^{prd}\right)^2 + n\left(\bar{y}^{obs}-\bar{y}^{prd}\right)^2}, \tag{14}$$

where $\bar{y}^{prd}$ and $\bar{y}_{TR}$ are the means of the predicted log k_p values of the test set and the observed log k_p values of the training set, respectively. n_{TR} in Eq. (13) is the number of samples in the training set. The applied statistical metrics and proposed criteria for accepting a prediction model as externally predictive (along with their references) are summarized in Table I.

Sensitivity Analysis

We evaluated the impact of each descriptor of the prediction model on skin permeability *via* sensitivity analysis. Considering that the impact of descriptors in the nonlinear model cannot be uniquely described, partial derivatives of a machine learning model at each data point were used as sensitivity measures. Here, normalized input data were used to compensate for the difference in scale among the descriptors. Partial differentiation was performed by using the grad function in the numDeriv R package (ver. 2012.9.1) that supports numerical differentiation using Richardson's extrapolation (59).

RESULTS AND DISCUSSION

Database Development and Visualization

Our newly compiled database for evaluating the solvent effect on human skin permeability is among the largest compiled to date, comprising 412 samples constructed from 261 permeants and 31 solvents. Among these, 130 samples are mixtures of more than two solvents; the remaining 282 samples are single solvents. The database contents are listed in the electronic supplementary material. All data were rigorously screened to ensure that they satisfied the above-mentioned quality criteria of datasets. Therefore, the analysis is not complicated by interspecies differences or by the differences between the *in vivo* and *in vitro* studies. The permeability coefficient (k_p) comprehensively quantifies percutaneous absorption, which embodies the effects of the vehicle–skin partition coefficient, the diffusion coefficient through skin, and the skin thickness. Figure 1 shows the distribution of log k_p in our database, where k_p is expressed in cm/s. According to this figure, log k_p is well-distributed from low permeability (log $k_p < -8.5$) to remarkably high permeability (log $k_p > -4.5$).

The permeants in our database exhibit a vast array of chemical features and include almost all relevant functional groups (alcohols, ethers, thioethers, halides, amines, amides, esters, carbonyls, fatty acids, ureas, sulfoxides, nitriles, nitros, and various aromatic compounds). We emphasize that many of the permeants are biologically active ingredients such as anti-cancer, anti-HIV, and anti-inflammatory drugs. Others can function as agricultural chemicals and toxins.

Table I Metrics for Model Validation

Metrics	References	Equation	Criteria	References of criteria
Determination coefficient (r^2)	–	Eq. (2)	$r^2 > 0.6$	51
Root mean absolute error (RMSE)	–	Eq. (3)	–	–
Mean absolote error (MAE)	–	Eq. (4)	–	–
Slopes (k, k')	51	Eqs. (5), (6)	$0.85 \le k \le 1.15$ *and* $0.85 \le k' \le 1.15$	51
r^2 without intercept $(r_0^2, r_0'^2)$	51	Eqs. (7), (8)	$\frac{r^2-r_0^2}{r^2} \le 0.10$ *and* $\frac{r^2-r_0'^2}{r^2} \le 0.10$	51
Modified r^2 parameter $(r_m^2, r_m'^2)$	52	Eqs. (9), (10)	$r_m^2 \ge 0.50$ *and* $r_m'^2 \ge 0.50$	52
Average of r_m^2 and $r_m'^2$ $(\overline{r_m^2})$	54	Eq. (11)	$\overline{r_m^2} > 0.65$	57
Difference of r_m^2 and $r_m'^2$ (Δr_m^2)	54	Eq. (12)	$\Delta r_m^2 \le 0.2$	57
External predictive ability (Q_{F3}^2)	53	Eq. (13)	$Q_{F3}^2 > 0.70$	56
Concordance correlation coefficient (CCC)	55	Eq. (14)	$CCC > 0.85$	56

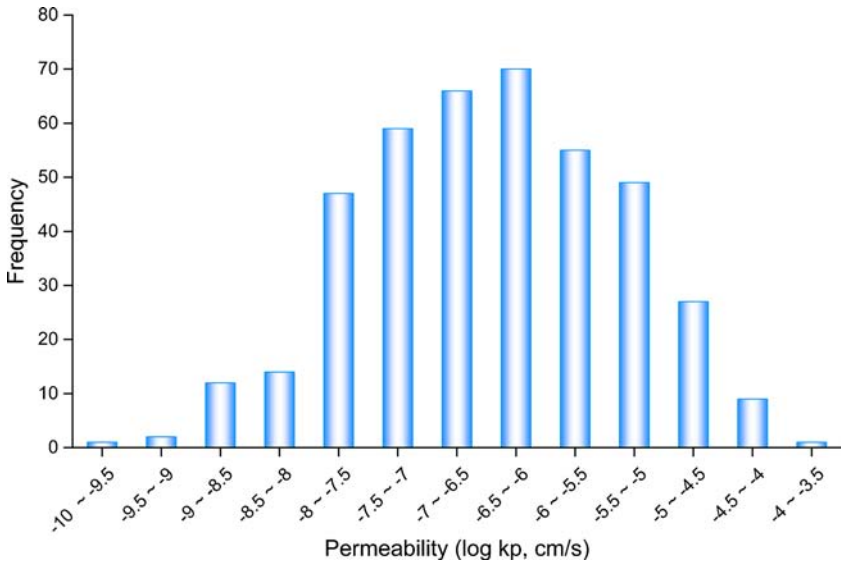

Fig. 1 Distribution of permeability (log k_p) in our database.

The MW of the permeants ranges from 32.05 to 765.0 (mean = 231.08; standard deviation = 117.15), and the ALOGP ranges from −6.06 to 6.84 (mean = 1.70; standard deviation = 1.54). Clearly, the permeants in our database are widely diverse in their MW and octanol–water partition coefficients for skin permeants (60, 61). The importance of both parameters in skin permeability has been demonstrated in previous studies (11, 62). The 2D scatter plot in Fig. 2 compares the ALOGP–MW dependencies of the permeants in our database with those of drugs in DrugBank (whose MW < 1000). Compounds in our database and DrugBank are indicated in red and blue, respectively. Clearly, the structural characteristics of compounds in our database almost encompass the diversity of structures in DrugBank. This indicates that our collected permeants are sufficiently diverse to evaluate the pharmaceutical candidates for dermatological formulations.

Our database also contains chemically diverse solvents, including alcohols, esters, fatty acids, sulfoxides, amines, aromatic compounds, and simple hydrocarbons. The MW of the solvents ranges from 18.02 to 885.61 (mean = 164.45; standard deviation = 176.05), and the ALOGP ranges from −1.95 to 20.93 (mean = 2.54; standard deviation = 4.34). As mentioned above, some of these compounds are used as skin penetration enhancers. The solvents and their categories are listed in Table II. Nine solvents (*i.e.*, ethanol, isopropyl alcohol, propylene glycol, lactic acid, oleic acid, isopropyl myristate, polyethylene glycol, propylene carbonate, and water) in our database are included in the recently reported list of topical excipients used frequently in dermal drugs (63); furthermore, four solvents in our database rank among the top five most frequently used excipients. Although medium chain triglycerides, acetic acid, and triethanol amine, which are frequently used excipients,

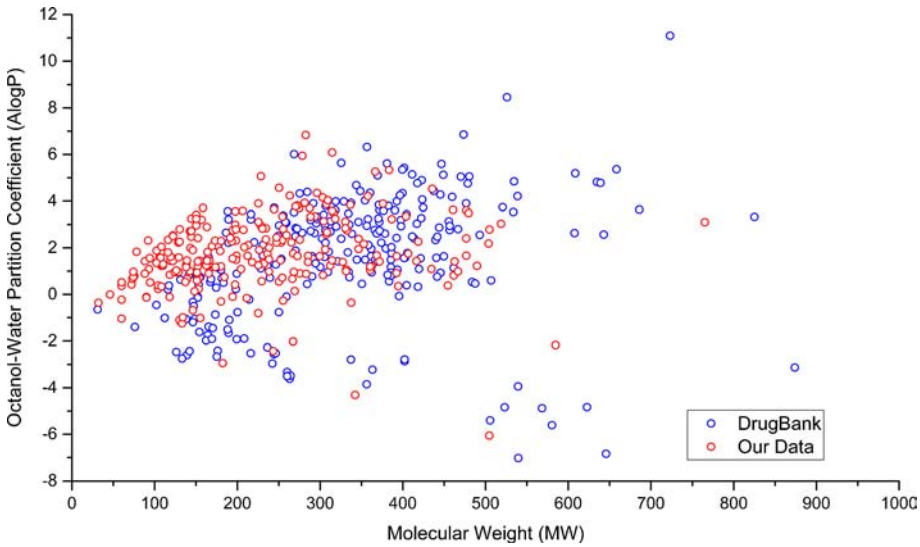

Fig. 2 Comparison of permeants in our database with DrugBank over molecular weight and ALOGP.

Table II Solvents in Our Database

Alcohols	Fatty acids	Esters	Carbonyls	Hydrocarbons	Polyethers	Others
ethanol	hexanoic acid	ethyl decanoate	acetone	hexadecane	PEG200 dilaurate	chloroform
isopropyl alcohol	lactic acid	glycerol trioleate	dimethyl formamide	hexane	PEG400	dimethyl isosorbide
methanol	lauric acid	isopropyl myristate	isophorone	isooctane		dimethyl sulfoxide
n-butanol	octanoic acid			toluene		propylene carbonate
n-octanol	oleic acid					triethylamine
n-propanol	propionic acid					water
propylene glycol						

are not in our database, their chemical features may be compensated by excipients similar in chemical structure, including glycerol trioleate, propionic acid, and trimethylamine, respectively. The above indicates that our database includes data regarding solvent effects on human skin permeability resulting from solvents frequently considered at the design stage of dermatological formulations, and therefore *in silico* models that learn our database as a training set are beneficial for screening or predicting the skin permeability of test dermatological formulations.

In summary, the permeability coefficients of our collected compounds through human skin are widely variable, and the mixture components are chemically diverse. Moreover, by excluding data not satisfying strict criteria, we ensure that the underlying data are collected under the same experimental conditions. Therefore, this novel and large dataset is highly suitable for evaluating the solvent effect on human skin permeability.

Chemical Space Comparison of Training and Test Sets

When internally and externally validating a model, appropriately selecting the training and test sets is important (52). To evaluate whether our original dataset was properly divided into training and test sets, we plotted the chemical space as a 3D scatter plot of objective permeability and the first two principal components of PCA (see Fig. 3). In Fig. 3, the training and test sets are depicted in navy and red, respectively. Note that the distributions of samples in the training and test sets cover the entire feature space and are very similar (overlapping each other). This implies that the test samples cover most of the structural characteristics in the original dataset, and that the predictive performance of the model is properly assessed by the internal and external validations.

Predictive Ability of the Developed Models

Next, we constructed prediction models using SVR and RF. In our stepwise descriptor selection, 25 and 14 descriptors were obtained for SVR and RF, respectively. Table III shows the r^2 values, RMSEs, and MAEs of our models. In both internal and external validations, the strong performers were SVR and RF with greedy stepwise selection of the descriptors. However, according to Table III, SVR yielded more favorable statistics than RF. The superior performance of SVR is also revealed in Figs. 4a and b, which plots the observed *versus* the predicted log k_p values of the training and test sets using SVR and RF, respectively. The compatibility of SVR and RF depend on the quality of the data processed by the algorithms. SVR is better suited to model development in this study than RF.

Predictablity evaluation is an essential component of QSPR research. To determine whether our models are strongly predictive, we used the statistics presented in Table I, which have been proposed as external validation criteria in recent detailed studies (51–57), as well as conventional metrics such as r^2, RMSE, and MAE. The statistical results of our SVR and RF models are summarized in Table IV. Clearly, all performance statistics of both models satisfied the above criteria (Table I), even though SVR yielded better results. Although no models exist for comparison under the same conditions (including Reviere and Ghafourian's models (electronic supplementary material)), the high reliability and predictive ability of our models was confirmed in the statistical evaluation. High-performance QSPR models that comprehensively predict the abstruse solvent effect of structurally diverse compounds on human skin permeability are currently unavailable. Therefore, this study provides an important alternative to permeability experiments.

As an example, Fig. 5 shows the predicted solvent effects of eight samples on skin permeability. These results were obtained in an external validation of the SVR. The model accurately predicted the sample affected by solvent composition (physostigmin), and that unaffected by solvent composition (estradiol). The model also strongly predicted the permeability of compounds with neutral, basic, and acidic functional groups in their solutions (DEET, adenosine, and genistein).

To confirm more comprehensively that there are no specific solvents in which our SVR model does not strongly predict skin permeability, we investigated the relationships between the prediction errors and solvent-related descriptors. Figure 6 shows parallel coordinates (64) for the characterization of prediction

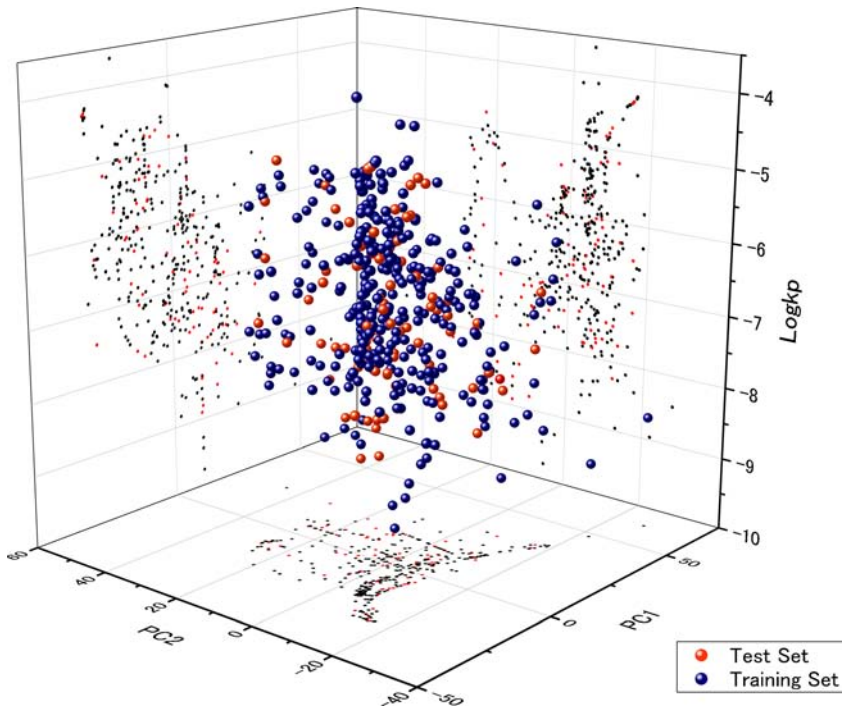

Fig. 3 3D scatter plots of the permeability (observed log k_p) against the first two principal components. Navy and red data points represent the training set and test set, respectively.

errors in the external validation of our model using five descriptors—MAXDN, DISPv, E1u, P_VSA_MR_3, and MATS2i—related to solvent, observed, and predicted skin permeabilities (log kps). Each line in the figure, which corresponds to an entry of the test set, is colored according to the difference between the observed and predicted permeability values. The red- and blue-colored lines, which correspond to entries with comparatively large errors, do not have distinct differences in distribution against the entire entries in each descriptor, which indicates that specific solvent characteristics do not cause any degradation of predictive performance. Hence, our model does not have a defect in prediction against specific solvents and can, therefore, be used to predict a wide range of solvent effects on skin permeability.

Descriptor Selection

Descriptor selection is an important step in constructing an efficient prediction model. Typically, descriptor selection aims to improve the calculation efficiency of the model by excluding redundant and irrelevant descriptors (65). Descriptor selection can be assisted by various techniques such as stepwise methods, swarm intelligence, genetic algorithms, and *k*-nearest neighbor methods. In this study, essential descriptors for predicting the solvent effect were selected by the greedy stepwise forward selection. The descriptors selected for the SVR and RF models are presented in Tables V and VI, respectively. The descriptors related to mixture solvents were averaged by the fractions of each solvent in the mixture (29). Although our models require more descriptors than previous models (up to five descriptors), this is of no consequence as our models are not overtrained (Table III). According to Tables V and VI, the selected descriptors include variables related to permeant lipophilicity (ALOGP, ALOGP2, and BLTA96), and variables associated with electronegativity (RDF035e, and R5e), nucleophilicity, and ionization potential (HOMO). Similarly, solvent descriptors involve variables related to nucleophilicity and ionization potential (Eion,

Table III Determination Coefficients (r^2), Root Mean Squares Error (RMSE), and Mean Absolute Error (MAE) for Our Models

Regression algorithms	The number of descriptors	Training - cross validation			Test - external validation		
		r^2	RMSE	MAE	r^2	RMSE	MAE
SVR	25	0.846	0.436	0.342	0.899	0.351	0.268
RF	14	0.794	0.507	0.383	0.825	0.459	0.327

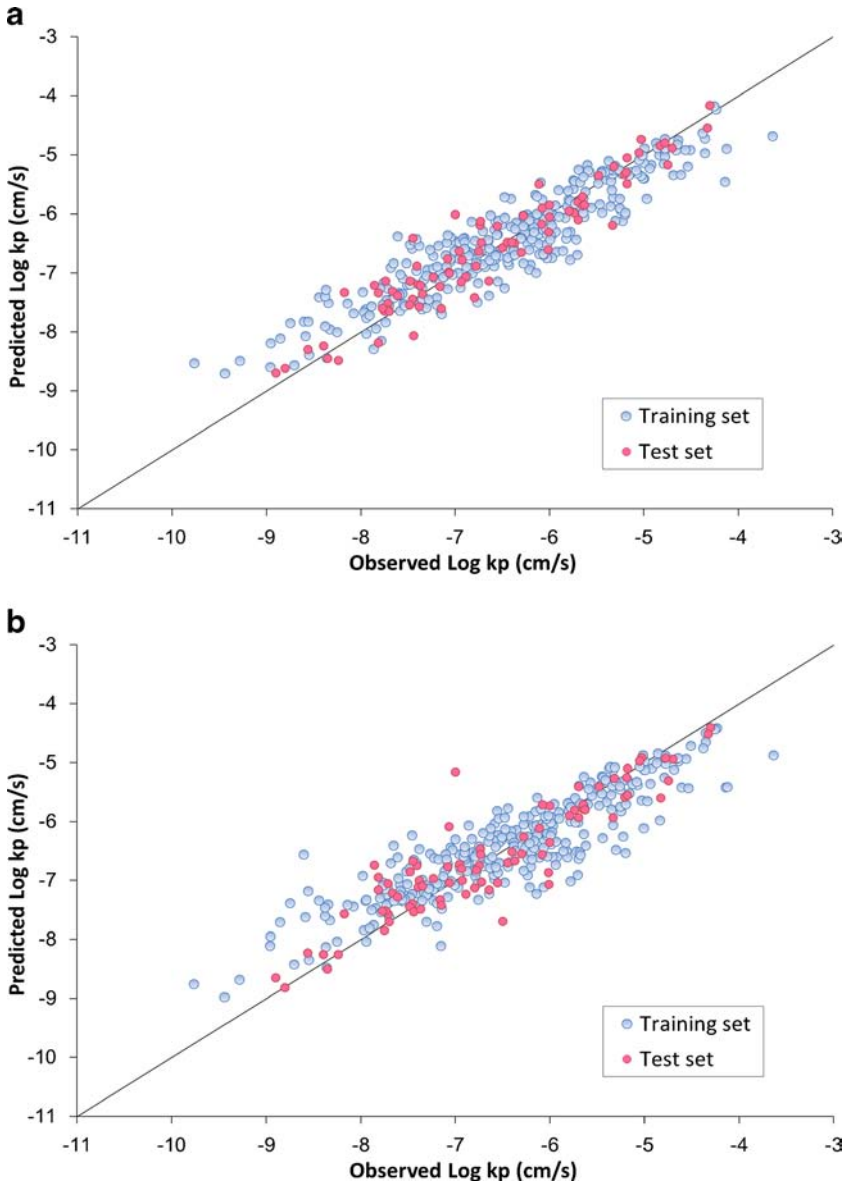

Fig. 4 (**a**) Predicted log k_p by SVR *vs.* experimental data. (**b**) Predicted log k_p by RF *vs.* experimental data.

MAXDN, MATS1e, and MATS2i). In general, the skin permeability of a permeant is significantly affected by its lipophilicity (such as log P), polarity and molecular form (ionic or neutral) (66,67). Therefore, we may reasonably expect that descriptors associated with these properties are selected as essential for prediction.

Unlike linear regression, it is generally difficult to elucidate the direct effects of explanatory variables (*i.e.*, molecular descriptors in this study) on objective variables (*e.g.*, log k_p) due to the nonlinearity of SVR models. Hence, we evaluated the sensitivities of each of the 25 selected descriptors. Figure 7 shows individual and median values of partial derivatives of

Table IV Statistical Parameters of External Validation for QSPR Models

Algorithm	Statistical parameters									
	k	k'	$\frac{r^2-r_0^2}{r^2}$	$\frac{r^2-r_0'^2}{r^2}$	r_m^2	r'^2_m	$\overline{r_m^2}$	Δr_m^2	Q_{F3}^2	CCC
SVR	0.991	1.007	0.012	0.000	0.804	0.898	0.851	0.094	0.895	0.946
RF	0.991	1.005	0.047	0.000	0.662	0.820	0.741	0.158	0.821	0.903

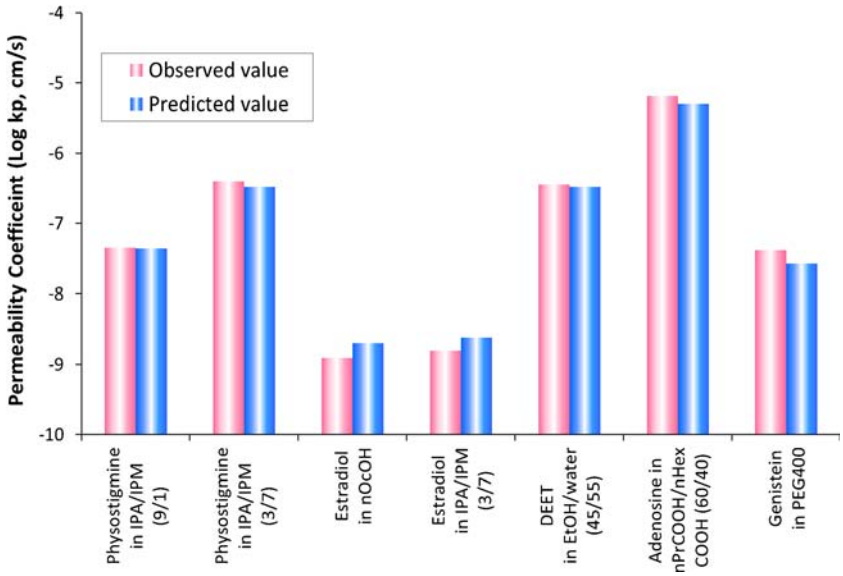

Fig. 5 Examples of solvent effects predicted by SVR. IPA: isopropyl alcohol, IPM: isopropyl myristate, nOcOH: n-octyl alcohol, DEET: *N*,*N*-diethyl m-toluamide, EtOH: ethanol, nPrCOOH: propioic acid, nHexCOOH: caproic acid, and PEG400: polyethylene glycol 400.

the SVR model for each of all 412 entries in our database with respect to each of the descriptors. Regarding permeant descriptors, Fig. 7 shows that CATS2D_04_AA and ALOGP have comparatively large positive effects on skin permeability, and conversely, BLTA96 (which is composed of −MLOGP (68)), SpMax_Bh(m), R5e, and RDF035e tend to negatively affect skin permeability. Regarding solvent descriptors, no descriptor has a clear positive or negative impact on permeability; however, DISPv, E1u, and MATS2i tend to be rather positive, whereas MAXDN tends to be negative. Although we provided a summary of these descriptors in Table V, details of the mechanistic effects such descriptors are somewhat obscure.

The descriptors in the present study are fully computational. That is, the human skin permeability of the permeant is solely predicted from the structural formula and the fraction of each component in the mixture. This property enables the evaluation of as-yet unsynthesized compounds or unavailable ingredients, without requiring experiments. Each step of the prediction, namely, geometry optimization, descriptor calculation, and prediction by SVR and RF models, is considerably more efficient than the procedures used in permeability experiments. In practice, the predictive model can screen numerous candidate compounds for their skin permeability at an early stage and optimize the compositions of topical formulations.

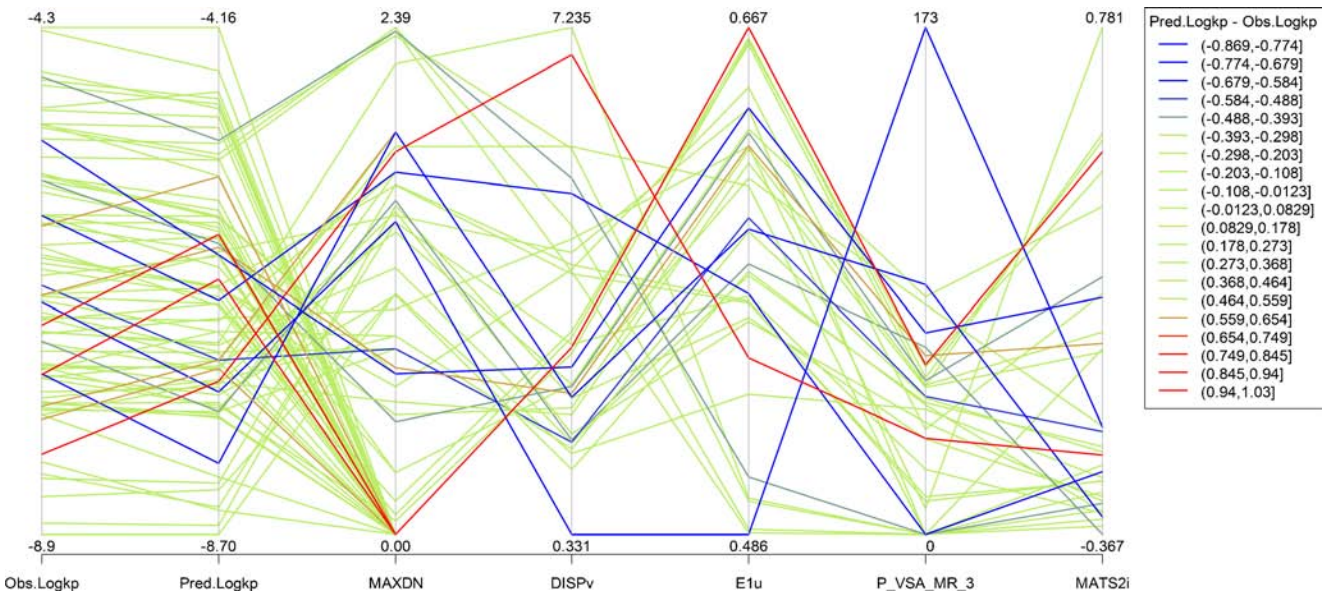

Fig. 6 Parallel coordinates for the characterization of prediction results in external validation of SVR model using five descriptors related to solvent and observed and predicted skin permeabilities (log kps).

Table V Selected Descriptors for the SVR with Gaussian Kernel

Name	Description	Permeants or Solvents	Block
R5e	R autocorrelation of lag 5 / weighted by Sanderson electronegativity	Permeants	GETAWAY descriptors
MAXDN	maximal electrotopological negative variation	Solvents	Topological indices
ALOGP	Ghose-Crippen octanol-water partition coeff. (logP)	Permeants	Molecular properties
RDF035e	Radial Distribution Function - 035 / weighted by Sanderson electronegativity	Permeants	RDF descriptors
DISPv	displacement value / weighted by van der Waals volume	Solvents	Geometrical descriptors
SpMax4_Bh(m)	largest eigenvalue n. 4 of Burden matrix weighted by mass	Permeants	Burden eigenvalues
PDI	packing density index	Permeants	Molecular properties
E1u	1st component accessibility directional WHIM index / unweighted	Solvents	WHIM descriptors
HOMO	energy of highest occupied molecular orbital	Permeants	Quantum chemistry
H3m	H autocorrelation of lag 3 / weighted by mass	Permeants	GETAWAY descriptors
CATS2D_04_AA	CATS2D Acceptor-Acceptor at lag 04	Permeants	CATS 2D
H-047	H attached to C1(sp3)/C0(sp2)	Permeants	Atom-centred fragments
F06[C-N]	Frequency of C - N at topological distance 6	Permeants	2D Atom Pairs
O%	percentage of O atoms	Permeants	Constitutional indices
BLTA96	Verhaar Algae base-line toxicity from MLOGP (mmol/l)	Permeants	Molecular properties
nCIR	number of circuits	Permeants	Ring descriptors
P_VSA_MR_3	P_VSA-like on Molar Refractivity, bin 3	Solvents	P_VSA-like descriptors
MWC05	molecular walk count of order 5	Permeants	Walk and path counts
ALOGP2	squared Ghose-Crippen octanol-water partition coeff. (logP ^ 2)	Permeants	Molecular properties
Psychotic-80	Ghose-Viswanadhan-Wendoloski antipsychotic-like index at 80%	Permeants	Drug-like indices
TDB06v	3D Topological distance based descriptors - lag 6 weighted by van der Waals volume	Permeants	3D autocorrelations
F10[C-Cl]	Frequency of C - Cl at topological distance 10	Permeants	2D Atom Pairs
CATS2D_06_DP	CATS2D Donor-Positive at lag 06	Permeants	CATS 2D
MATS2i	Moran autocorrelation of lag 2 weighted by ionization potential	Solvents	2D autocorrelations
R1m+	R maximal autocorrelation of lag 1 / weighted by mass	Permeants	GETAWAY descriptors

Table VI Selected Descriptors for the RF

Name	Description	Permeants or Solvents	Block
GGI5	topological charge index of order 5	Permeants	2D autocorrelations
nHAcc	number of acceptor atoms for H-bonds (N,O,F)	Permeants	Functional group counts
Eion	ionization potential	Solvents	Quantum chemistry
ALOGP	Ghose-Crippen octanol-water partition coeff. (logP)	Permeants	Molecular properties
HOMT	HOMA (Harmonic Oscillator Model of Aromaticity index) total	Permeants	Geometrical descriptors
MATS1e	Moran autocorrelation of lag 1 weighted by Sanderson electronegativity	Solvents	2D autocorrelations
B07[N-O]	Presence/absence of N - O at topological distance 7	Permeants	2D Atom Pairs
P_VSA_LogP_4	P_VSA-like on LogP, bin 4	Permeants	P_VSA-like descriptors
B07[C-N]	Presence/absence of C - N at topological distance 7	Permeants	2D Atom Pairs
P_VSA_MR_6	P_VSA-like on Molar Refractivity, bin 6	Solvents	P_VSA-like descriptors
CATS2D_07_DD	CATS2D Donor-Donor at lag 07	Permeants	CATS 2D
E2v	2nd component accessibility directional WHIM index / weighted by van der Waals volume	Permeants	WHIM descriptors
nRCN	number of nitrilces (aliphatic)	Permeants	Functional group counts
TDB02	3D Topological distance based descriptors - lag 2 weighted by polarizability	Solvents	3D autocorrelations

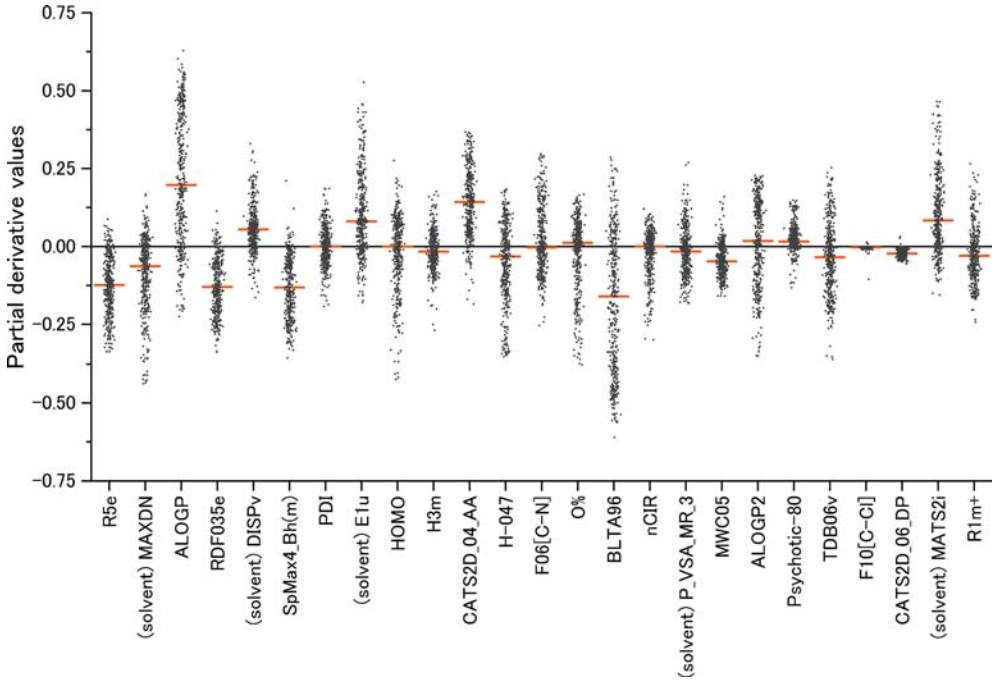

Fig. 7 Distribution and median values of partial derivatives of the SVR model for all 412 entries in our database with respect to each of the selected 25 descriptors.

CONCLUSION

We have compiled a novel database for evaluating the solvent effect on skin permeability. The database includes 412 samples comprising 261 permeants and 31 solvents collected from the literature. This is the first large dataset of human skin permeability coefficients for elucidating complicated solvent effects.

The accumulated data were carefully screened against strict quality criteria; the permeants and solvents cover a wide range of permeabilities and structural features. In combination with QSPR models or other theoretical considerations, our database provides a comprehensive means of evaluating the effects of solvents on skin permeability.

To realize a high-performance predictive model of solvent effect on skin permeability, we applied SVR and RF with greedy stepwise descriptor selection to our database. SVR and RF were computationally fast. In the model validation, both models satisfied all of the statistical criteria proposed in recent QSPR validation studies, and are accepted as truly predictive. However, the SVR achieved higher performance statistics than RF. As all descriptors were calculated based on the geometrical structures of their constituents, the model can also predict the solvent effects of as-yet unsynthesized or unavailable compounds. In summary, we have developed high-performance predictive models based on chemically diverse permeants and solvents, which effectively screen pharmaceutical and cosmetic candidates and optimize the skin permeability of topical formulations. Because they analyze a wide range of chemical compounds, the models are suitable for industrial use. They also offer a fast, inexpensive alternative to permeability experiments.

REFERENCES

1. Prausnitz MR, Mitragotri S, Langer R. Current status and future potential of transdermal drug delivery. Nat Rev Drug Discov. 2004;3(2):115–24.
2. Bartek MJ, LaBudde JA, Maibach HI. Skin permeability in vivo: comparison in rat, rabbit, pig and man. J Invest Dermatol. 1972;58(3):114–23.
3. Franz TJ. Percutaneous absorption on the relevance of in vitro data. J Invest Dermatol. 1975;64(3):190–5.
4. 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.
5. 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.
6. Karadzovska D, Riviere JE. Assessing vehicle effects on skin absorption using artificial membrane assays. Eur J Pharm Sci. 2013;50(5):569–76.
7. 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.
8. Abraham MH, Martins F. Human skin permeation and partition: general linear free-energy relationship analyses. J Pharm Sci. 2004;93(6):1508–23.

9. Buchwald P, Bodor N. A simple, predictive, structure-based skin permeability model. J Pharm Pharmacol. 2001;53(8):1087–98.
10. 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.
11. Potts RO, Guy RH. Predicting skin permeability. Pharm Res. 1992;9(5):663–9.
12. Baba H, Takahara JI, Mamitsuka H. In silico predictions of human skin permeability using nonlinear quantitative structure–property relationship models. Pharm Res. 2015. doi:10.1007/s11095-015-1629-y.
13. Khajeh A, Modarress H. Linear and nonlinear quantitative structure–property relationship modelling of skin permeability. SAR QSAR Environ Res. 2014;25(1):35–50.
14. 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.
15. Baert B, Deconinck E, Van Gele M, Slodicka M, Stoppie P, Bodé S, *et al.* Transdermal penetration behaviour of drugs: CART-clustering, QSPR and selection of model compounds. Bioorg Med Chem. 2007;15(22):6943–55.
16. 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.
17. 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.
18. 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.
19. Smith EW, Maibach HI, editors. Percutaneous penetration enhancers, second ed. New York: CRC Press; 2005.
20. Barry BW. Breaching the skin's barrier to drugs. Nat Biotechnol. 2004;22(2):165–7.
21. Lane ME. Skin penetration enhancers. Int J Pharm. 2013;447(1–2):12–21.
22. Watkinson RM, Guy RH, Hadgraft J, Lane ME. Optimisation of cosolvent concentration for topical drug delivery - II: influence of propylene glycol on ibuprofen permeation. Skin Pharmacol Physiol. 2009;22(4):225–30.
23. Roy SD, Manoukian E. Transdermal delivery of ketorolac tromethamine: permeation enhancement, device design, and pharmacokinetics in healthy humans. J Pharm Sci. 1995;84(10):1190–6.
24. Sathyan G, Ritschel WA, Hussain AS. Transdermal delivery of tacrine: I: Identification of a suitable delivery vehicle. Int J Pharm. 1995;114(1):75–83.
25. Goldberg-Cettina M, Liu P, Nightingale J, Kurihara-Bergstrom T. Enhanced transdermal delivery of estradiol in vitro using binary vehicles of isopropyl myristate and short-chain alkanols. Int J Pharm. 1995;114(2):237–45.
26. Pardo A, Shiri Y, Cohen S. Percutaneous absorption of physostigmine: optimization of delivery from a binary solvent by thermodynamic control. J Pharm Sci. 1990;79(7):573–8.
27. Riviere JE, Brooks JD. Prediction of dermal absorption from complex chemical mixtures: incorporation of vehicle effects and interactions into a QSPR framework. SAR QSAR Environ Res. 2007;18(1–2):31–44.
28. Riviere JE, Brooks JD. Predicting skin permeability from complex chemical mixtures: dependency of quantitative structure permeation relationships on biology of skin model used. Toxicol Sci. 2011;119(1):224–32.
29. Ghafourian T, Samaras EG, Brooks JD, Riviere JE. Modelling the effect of mixture components on permeation through skin. Int J Pharm. 2010;398(1–2):28–32.
30. Ghafourian T, Samaras EG, Brooks JD, Riviere JE. Validated models for predicting skin penetration from different vehicles. Eur J Pharm Sci. 2010;41(5):612–6.
31. Netzeva TI, Worth A, Aldenberg T, Benigni R, Cronin MT, Gramatica P, *et al.* Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. Altern Lab Anim. 2005;33(2):155–73.
32. Moss GP, Sun Y, Prapopoulou M, Davey N, Adams R, Pugh WJ, *et al.* The application of Gaussian processes in the prediction of percutaneous absorption. J Pharm Pharmacol. 2009;61(9):1147–53.
33. Patel J. Science of the science, drug discovery and artificial neural networks. Curr Drug Discov Technol. 2013;10(1):2–7.
34. Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput. 2004;14(3):199–222.
35. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
36. 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.
37. 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.
38. 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.
39. Bharathidason S, Venkataeswaran CJ. Improving classification accuracy based on random forest model with uncorrelated high performing trees. Int J Comput Appl. 2014;101(13):26–30.
40. Blank IH, McAuliffe DJ. Penetration of benzene through human skin. J Investig Dermatol. 1985;85(6):522–6.
41. Burkert U, Norman LA. Molecular mechanics, ACS monograph 177. Washington, DC: American Chemical Society; 1982.
42. Stewart JJ. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. J Mol Model. 2013;19(1):1–32.
43. 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.
44. We have obtained structures of 6475 compounds as approved drugs in DrugBank: http://www.drugbank.ca/downloads#structures/.
45. 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/.
46. 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.
47. Liaw A, Wiener M. Classification and regression by random Forest. R News. 2002;2:18–22.
48. Shao J. Linear model selection by cross-validation. J Am Stat Assoc. 1993;88(442):486–94.
49. Zhang P. Model selection via multifold cross validation. Ann Stat. 1993;21(1):299–313.
50. 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.
51. Golbraikh A, Tropsha A. Beware of q2! J Mol Graph Model. 2002;20(4):269–76.
52. Roy PP, Roy K. On some aspects of variable selection for partial least squares regression models. QSAR Comb Sci. 2008;27(3):302–13.
53. Consonni V, Ballabio D, Todeschini R. Comments on the definition of the Q2 parameter for QSAR validation. J Chem Inf Model. 2009;49(7):1669–78.

54. Ojha PK, Mitra I, Das RN, Roy K. Further exploring rm2 metrics for validation of QSPR models. Chemom Intell Lab Syst. 2011;107(1):194–205.
55. Chirico N, Gramatica P. Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. J Chem Inf Model. 2011;51(9):2320–35.
56. Chirico N, Gramatica P. Real external predictivity of QSAR models.Part 2.New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. J Chem Inf Model. 2012;52(8):2044–58.
57. Roy K, Mitra I, Kar S, Ojha PK, Das RN, Kabir H. Comparative studies on some metrics for external validation of QSPR models. J Chem Inf Model. 2012;52(2):396–408.
58. Funar-Timofei S, Iliescu S, Suzuki T. Correlations of limiting oxygen index with structural polyphosphoester features by QSPR approaches. Struct Chem. 2014;25(6):1847–63.
59. Fornberg B, Sloan DM. A review of pseudospectral methods for solving partial differential equations. Acta Numer. 1994;3:203–67.
60. Bos JD, Meinardi MM. The 500 Dalton rule for the skin penetration of chemical compounds and drugs. Exp Dermatol. 2000;9(3):165–9.
61. 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.
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. Boonen J, Veryser L, Taevernier L, Roche N, Peremans K, Burvenich C, *et al.* Risk evaluation of impurities in topical excipients: the acetol case. J Pharm Anal. 2014;4(5):303–15.
64. Inselberg A. Parallel coordinates: a tool for visualizing multidimensional geometry. New York: Springer; 2009.
65. Shahlaei M. Descriptor selection methods in quantitative structure-activity relationship studies: a review study. Chem Rev. 2013;113(10):8093–103.
66. Geinoz S, Guy RH, Testa B, Carrupt PA. Quantitative structure-permeation relationships (QSPeRs) to predict skin permeation: A critical evaluation. Pharm Res. 2004;21(1):83–92.
67. Michaels AS, Chandrasekaran SK, Shaw JE. Drug permeation through human skin: Theory and invitro experimental measurement. AIChE J. 1975;21(5):985–96.
68. Verhaar HJM, van Leeuwen CJ, Hermens JLM. Classifying environmental pollutants 1: Structure-activity relationships for prediction of aquatic toxicity. Chemosphere. 1992;25(4):471–91.