ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Bioorganic & Medicinal Chemistry 15 (2007) 6943–6955

Bioorganic &
Medicinal
Chemistry

Transdermal penetration behaviour of drugs: CART-clustering, QSPR and selection of model compounds

Bram Baert,[a] Eric Deconinck,[b] Mireille Van Gele,[c] Marian Slodicka,[d] Paul Stoppie,[e] Samuel Bodé,[a] Guido Slegers,[a] Yvan Vander Heyden,[b] Jo Lambert,[c] Johan Beetens[e] and Bart De Spiegeleer[a,*]

[a]*Drug Quality and Registration (DruQuaR) Group, Department of Pharmaceutical Analysis, Faculty of Pharmaceutical Sciences, Ghent University, Harelbekestraat 72, B-9000 Ghent, Belgium*

[b]*Department of Analytical Chemistry and Pharmaceutical Technology, Pharmaceutical Institute, Vrije Universiteit Brussel (VUB), Laarbeeklaan 103, B-1090 Brussels, Belgium*

[c]*Department of Dermatology, Ghent University Hospital, De Pintelaan 185, B-9000 Ghent, Belgium*

[d]*Department of Mathematical Analysis, Ghent University, Galglaan 2, B-9000 Ghent, Belgium*

[e]*Barrier Therapeutics nv., Cipalstraat 3, B-2440 Geel, Belgium*

Received 18 April 2007; revised 4 July 2007; accepted 27 July 2007
Available online 22 August 2007

Abstract—A set of 116 structurally very diverse compounds, mainly drugs, was characterized by 1630 molecular descriptors. The biological property modelled in this study was the transdermal permeability coefficient $\log K_p$. The main objective was to find a limited set of suitable model compounds for skin penetration studies. The classification and regression trees (CART) approach was applied and the resulting groups were discussed in terms of their role as possible model compounds and their determining descriptors. A second objective was to model transdermal penetration as a function of selected descriptors in quantitative structure–property relationships (QSPR) using a boosted CART (BRT) approach and multiple linear regression (MLR) analysis, where regression models were obtained by stepwise selection of the best descriptors. Evaluation of the standard statistical, as well as descriptor-number dependent, regression quality attributes yielded a maximal 10-dimensional MLR model. The CART and MLR models were subjected to an external validation with a test set of 12 compounds, not included in the original learning set of 104 compounds, to assess the predictive power of the models.

1. Introduction

The skin, one of the largest human organs, is an important route to chemicals from different origin: pharmaceuticals, cosmetics, household, agriculture and industrial products. The assessment of absorption/delivery into or across the skin of those chemicals is therefore important in toxicology, pharmacology, drug delivery, environmental-, dermatological- and cosmetic research.

Topical administration of chemicals to the human skin serves different functional purposes.[1] Firstly, it may be the aim of the formulation to keep the active compound on the surface of the skin, for example, skin disinfectants or dermal insect repellents. A second purpose is to design the formulation for dermal drug delivery. These preparations allow the drug to penetrate to the deeper layers of the epidermis and the dermis without (or minimal) absorption into the systemic circulation. As such, high effective concentrations of the drugs can be delivered at the (skin) site of action, without inducing systemic adverse effects. This approach has been successfully applied for the treatment of different skin disorders by topical administration of corticosteroids[2,3] and immunomodulators, such as pimecrolimus and tacrolimus.[4,5] Finally, transdermal delivery systems may aim to provide high and/or appropriately timed plasma concentrations of the drug without inducing local adverse reactions. Traditionally, ointments, gels and creams were used, but in the last two decades several new delivery systems such as transdermal patches[6] were devel-

Keywords: Transdermal permeability; Drugs; CART; QSPR.

* Corresponding author. Tel.: +32 9 264 81 00; fax: +32 9 264 81 93; e-mail: bart.despiegeleer@UGent.be

0968-0896/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2007.07.050

oped and boosted interest. Obviously, there are also chemicals for which the absorption and/or penetration is clearly undesirable, like pesticides.

All types of administration are controlled by the properties of the skin, the formulation characteristics and features of the chemical itself. The methods for measuring the skin permeability characteristics of a compound can be divided into in vivo and in vitro methods. In vivo tests in animals have traditionally been used to assess skin absorption for regulatory purposes and national guidelines are available.[7–9] Nevertheless, this approach has some disadvantages such as the use of living animals, the need for radio-labelled material to generate reliable results, difficulties in determining the early absorption phase and the differences in permeability between the tested species and human skin. The advantages of the in vitro over the in vivo approach are that it can be used equally well with skin from humans or animal species, replicate measurements can reproducibly be made, living animals are not used, the impact of skin differences including damage can be assessed in a more controlled way, a wide range of chemical and physical vehicles can be investigated and non-radio-labelled test substances can be used. In vitro experiments have delivered useful data for a wide range of chemicals, including the evaluation of novel formulations and devices for drug delivery, for example, transdermal penetration enhancers[10] or iontophoresis.[11] When developing new delivery systems, suitable model compounds spanning the wide range of transdermal drug behaviour are thus searched for. This is also required for basic research purposes, investigating the mechanisms of permeability in human and animal species, for example, for pour-on or spot-on formulations. Last, a number of guidelines on the conduct of in vitro skin penetration studies have been recently described.[12] To demonstrate the performance and reliability of a test system in the performing laboratory, system suitability test (SST) controls are to be included for which the results should be in agreement with the set specifications, which are based on inter-laboratory validation studies.[13] The choice of SST-model compounds will greatly influence the adequacy of the control validation. Therefore, the choice of model compounds used in dermatological investigations is important, and has to be rationalized and justified.

The internationally accepted guidelines 427 and 428 from the Organization for Economic Co-operation and Development (OECD) recommend the use of reference substances as SST-model compounds, based only upon their different lipophilicity, given as $\log P_{\mathrm{o/w}}$. However, no strict scientific justification or rationale is given. Typical examples recommended by OECD for in vitro and in vivo studies are caffeine ($\log P_{\mathrm{o/w}} = 0.01$), benzoic acid ($\log P_{\mathrm{o/w}} = 1.83$) and testosterone ($\log P_{\mathrm{o/w}} = 3.32$). The same physico-chemical descriptor is most often also applied in toxicological risk evaluations, currently prominent within the new 'Registration, Evaluation and Authorization of Chemicals' (REACH) regulations.[14] A chemical acting as a skin sensitizer has to be absorbed by the skin, traverse the stratum corneum barrier, complex with immunogenic proteins and pass to responding immune-competent cells in the epithelial layers. If a chemical is either unable to penetrate the skin, or lacks the ability to react with such proteins, then it cannot act as a skin sensitizer. Most expert systems for the prediction of toxicity like DEREK establish the skin sensitization relevance by using the $\log P_{\mathrm{o/w}}$ value, and then applying a simple quantitative structure–activity relationship (QSAR) like the modified Potts and Guy equation.[15,16] This estimated skin permeability value then enables the expert system to indicate whether a skin sensitization hazard alert given from the protein reactivity is likely to be expressed.[17] Most models developed and used are limited to a selected set of compounds and/or use a $\log P$ parameter, which is sometimes combined with a variable describing the permeant size, like molecular volume (MV) or molecular weight (MW).[15,16,18–20] However, other molecular parameters may have a significant influence on the penetration of compounds into or through the skin. Therefore, it would be interesting to link molecular properties, quantitatively expressed in descriptors, to the (trans)dermal behaviour of a certain compound. Attempts have been made,[21,22] but there is still a need for an improved understanding, including more sophisticated models.[23]

Correlations between structural descriptors and the penetration properties of molecules through other non-skin membranes, like the blood–brain barrier,[24–26] the intestinal membrane[27–29] and even artificial membranes like polymer matrices[30,31] have been explored as well.

In this article, we propose a new approach for selecting transdermal SST-model compounds on a rational basis. The classification and regression trees (CART) technique is used to obtain different permeability classes, ranging from low to high absorption, for a large set of selected compounds. While most other techniques require a variable (a descriptor in our case) to be selected before the model makes a split, this is not so in CART where variable selection is part of the methodology. This means that modelling can be started with an extended set of molecular descriptors. Recently, this CART approach was already successfully used for the classification of drugs in oral absorption classes[32,33] and for the prediction of blood–brain barrier passage of drugs.[34] Moreover, quantitative structure–property relationship (QSPR) modelling with predictive power was also performed, the most important descriptors identified and related towards the physico-chemical characteristics of the permeant.

2. Results

2.1. The response variable

The permeability coefficient K_{p} of a compound, expressed in units of distance time^{-1}, is a fundamental biological parameter describing the intrinsic transdermal penetration behaviour of a compound. It describes the rate at which a compound permeates the skin. Due to its low and wide range of values, the log-transformed value was taken as the skin penetration response in our study.

Human-skin permeability data have been measured from aqueous or aqueous extrapolated solutions, and were collected in this study from the literature.[35–37] However, in the Magnusson publication, a related skin permeability descriptor, the logarithm of the maximal flux ($\log J_{max}$), was given instead of the $\log K_p$. The $\log J_{max}$ describes the maximal flux of the compound delivery under steady-state conditions. This flux J, that is the amount of drug traversing through a unit area of skin in unit time, in fact determines the efficiency of topical and transdermal delivery. It consists of the product of the permeability coefficient and the solubilised drug concentration used in the formulation. The given $\log J_{max}$ values were converted to the original $\log K_p$ values for the compounds listed in the Magnusson publication using the relation $\log K_p = \log J_{max} - \log S_{aq(T)}$, with $S_{aq(T)}$ the aqueous solubility of the compound at temperature T. As for some of the Magnusson compounds, published K_p values were also available, we additionally evaluated the consistency of our calculated $\log K_p$ values. The absolute difference between the $\log K_p$ calculated from J_{max} and $S_{aq(T)}$, and the published $\log K_p$, expressed as a percentage relative to the mean of both values, was therefore calculated. The obtained percentage deviation showed a median value of 3.73% ($n = 50$) and most values were almost identical, with 14 values even below 1% relative difference. As a consequence, the $\log K_p$ values calculated from $\log J_{max}$ are included in the dataset.

Since the skin permeability data are derived from more than one laboratory and no exact, consistent procedure to obtain these biological data has been used, inter-laboratory variations were expected. For 70 compounds of the learning set, more than one $\log K_p$ value was available. The average standard deviation was 0.40, with a maximum of 1.70. This was considered acceptable and therefore, the average $\log K_p$ value was used as the final response.

2.2. The diversity of compounds in the learning set

The diversity of the samples used in the training set for modelling strongly influences the interpretation possibilities of the modelling outcome. Moreover, for our purposes of selecting model compounds, a very diverse dataset is required. A commonly used similarity index between 2 compounds is the Tanimoto coefficient, from which a diversity index (DI) of a set of compounds can be calculated.[55–57] The lower the TC and DI values, the more diverse the dataset is from a descriptor's point of view. Using Z-scaled descriptors, congeneric datasets have DI values as high as 0.58, while datasets considered as very diverse have DI values as low as 0.07. Datasets for previous QSPR studies have DI values between 0.27 and 0.58.[57] Our dataset of 104 compounds gave a DI-value of 0.022, confirming that our learning set is extremely diverse.

2.3. The CART-tree models and selection of model compounds

Models are built using the $\log K_p$ values of the 104 molecules of the learning set (Table 1). During the building process the maximal tree is built and then pruned. In the next step, 10-fold cross-validation is carried out, resulting in a graph of the model error, represented by the root mean squared error of cross validation (RMSECV), as a function of the tree complexity (Fig. 1). This graph allows selection of the most suited tree. When a minimal RMSECV is used as selection criterion, a tree with 2 leaves is suggested as optimal, with the Ghose–Crippen octanol–water partition coefficient (ALOGP) as split criterion. The use of a $\log P$ parameter to significantly describe the transdermal behaviour can be expected. Indeed, in the literature, $\log P$ is considered as one of the most important properties of molecules for their passage through biomembranes.[38,39] The selection of $\log P$ out of more than 600 non-correlated descriptors indicates the ability of CART to relate transdermal permeability with molecular descriptors and its use for feature selection in QSPR.

A levelling-off at a complexity of 10 is observed: adding further descriptors is no longer justified. This tree, with 10 terminal leaves, is given in Figure 2. As the transdermal penetration of a compound is intrinsically a multi-route process via sweat ducts, through hair follicles and across the continuous stratum corneum,[40] with each of these consisting of multiple possible subroutes and mechanisms, this highest-complexity tree with 10 classes was finally chosen for an evaluation of the relevant descriptors.

Looking for an appropriate set of model compounds, the 3 currently recommended chemicals by the OECD are belonging to different leaves in this tree: caffeine is in leave 6, benzoic acid in leaf 2 and testosterone in leaf 10. However, at this complexity level, ideally 10 model compounds are required to cover the wide range of drugs, which may not always be feasible. The question 'when to stop' the CART process can thus also be viewed from the number of model compounds. If only 3 model compounds can be used, a tree with 3 terminal leaves is to be constructed (see Supplementary data file 2). However, at this complexity level, the OECD reference compounds caffeine and benzoic acid are belonging to the same leaf 1, while testosterone is in leaf 2. Separation of these 3 OECD compounds into different leaves occurs only from a tree complexity of 6. If 3 model compounds with a maximal diversity are to be chosen, either caffeine or benzoic acid can better be replaced by another compound belonging to leaf 3, for example, ibuprofen.

2.4. The selected CART descriptors

ALOGP is the logarithm of the octanol/water partition coefficient $P_{o/w}$ as calculated according to Ghose–Crippen.[41] This method is based on hydrophobic atomic constants, measuring the lipophilic contributions of atoms in the molecule, each described by its neighbouring atoms. ALOGP is thus one of the measures for the hydrophobicity/lipophilicity of a molecule. According to our CART and QSPR results, the ALOGP is a more important descriptor than MLOGP for the transdermal behaviour. The Morigushi MLOGP is another octanol/

Table 1. log K_p values of 104 learning compounds used in the CART and QSPR study

Substance name	log K_p (cm s^{-1})	CART leaf[a]	Ref.
Acetaminophen	−7.22	6	41
Acetylsalicylic acid	−6.96	6	41
Aldosterone	−8.61	1	41–43
Aminopyrine	−6.62	9	41
Amphetamine	−8.21	6	41
Amylobarbital	−6.22	6	41–43
Antipyrine	−7.72	8	41
Aspartic acid	−7.73	6	41
Atenolol	−8.00	1	41
Atropine	−8.05	6	41–43
Baclofen	−8.14	6	41
Barbital	−7.55	1	41–43
Benzoic acid	−5.16	2	60
Betamethasone	−7.35	6	41
Betamethasone-17-valerate	−6.12	10	41
Butobarbital	−7.30	6	41–43
Caffeine	−7.30	6	41,43
Chloropheniramine	−6.22	3	41–43
Chloroxylenol	−4.79	4	41,42
Clonidine	−6.71	10	41
Codeine	−7.68	1	41–43
Cortexolone	−7.70	8	41,42
Cortexone	−6.73	10	41,42
Corticosterone	−7.35	8	41–43
Cortisone	−8.61	1	41,42
Coumarin	−5.79	3	41
Cyclobarbital	−6.55	6	41
Dexamethasone	−7.28	6	12,41
Dextromethorphan	−4.55	3	41
Diazepam	−4.74	3	41
Diclofenac	−5.92	3	41–43
Diethylcarbamazine	−7.45	6	41–43
Digitoxin	−8.34	8	41–43
Dihydromorphine	−8.38	1	42
Ephedrine	−5.78	7	41–43
Estradiol	−5.84	10	41–43
Estriol	−7.86	8	41,42
Estrone	−6.70	10	41,42
Ethacrynic acid	−2.99	5	41
Ethanol	−6.78	7	41–43
Etorphine	−6.00	3	41–43
Fentanyl	−5.94	9	41–43
Fluocinonide	−6.44	9	41–43
Fluorouracil, 5-	−7.96	6	41
Flurbiprofen	−3.53	5	41
Formaldehyde	−6.35	7	41,42
Furosemide	−6.23	2	41
Glycolic acid	−6.97	6	41
Griseofulvin	−5.98	7	41
Histidine	−8.03	6	41
Hydrocortisone	−8.21	6	41–43
Hydrocortisone 21-hemipimelate	−6.31	3	42,43
Hydrocortisone 21-hemisuccinate	−6.76	6	42,43
Hydrocortisone 21-hexanoate	−5.31	3	42,43
Hydrocortisone 21-octanoate	−4.77	3	42,43
Hydrocortisone 21-propionate	−6.03	3	42,43
Hydromorphone	−8.13	1	41–43
Hydroxyprogesterone	−6.70	10	41,42
Hyoscine	−7.86	6	43
Ibuprofen	−3.01	5	41
Indomethacin	−5.23	3	41–43
Isoprenaline	−8.25	1	41
Isosorbide dinitrate	−5.43	10	41
Ketoprofen	−4.40	5	41
Ketorolac	−5.38	10	41
Levodopa	−7.75	1	41
Lidocaine	−5.59	3	41,43
Lysine, L-	−6.86	6	41
Mannitol	−7.79	7	41
Meperidine	−6.46	3	41–43
Metoprolol	−6.76	7	41
Minoxidil	−7.19	6	41
Morphine	−8.57	1	41–43
Naphthol, 2-	−5.12	4	41–43
Naproxen	−5.27	4	41–43
Nicorandil	−7.22	9	41
Nicotine	−5.35	7	41–43
Nicotinic acid (Vit B_3)	−8.02	6	41
Nifedipine	−6.63	4	41
Nitroglycerin	−5.53	3	41–43
Ouabain	−9.52	1	41–43
Oxprenolol	−6.37	7	41
Pentazocine	−5.76	3	41
Phenobarbital	−6.60	6	41–43
Piroxicam	−6.11	9	41–43
Prednisolone	−7.91	6	41
Pregnenolone	−6.11	10	41,42
Progesterone	−6.00	10	41–43
Propranolol	−6.55	9	41
Resorcinol	−6.91	6	41–43
Salicylamide	−5.31	2	41
Salicylic acid	−5.23	2	41–43
Scopolamine	−7.86	1	41,42
Styrene	−3.75	5	42,43
Sucrose	−8.87	1	41–43
Sufentanil	−5.80	5	41–43
Sulindac	−8.73	8	41
Testosterone	−6.02	10	41–43
Thymol	−4.83	4	41–43
Thyrotropin-releasing hormone	−7.83	6	41
Triamcinolone	−8.95	1	41
Triamcinolone acetonide	−6.99	9	41
Urea	−7.39	6	41–43
Water	−6.52	6	41–43

[a] Tree complexity = 10.

water partition coefficient, calculated from 13 whole-molecular topological parameters, including the summation of hydrophobic atoms and hydrophilic atoms, unsaturated bonds and other specific functionalities. In contrast, the Ghose–Crippen ALOGP is a purely atom-contribution based approach, where more than 100 atom classifications are used with different atom contributions to the ALOGP. This indicates that the fine substructure of the molecule dominates the lipophilic-based transdermal behaviour above the global molecular properties. A split value of 2.0145 separates the compounds into hydrophilic compounds to the left and rather hydrophobic compounds to the right with higher transdermal penetration properties.

Jhetv is the Balaban type index obtained from the van der Waals volume weighed distance matrix. It is a topological index, describing the two-dimensional atom–atom connectivity in a molecule and is used as a measure for molecule size, shape, complexity and branching.[42]

This descriptor with a rather low split value of 2.204 indicates that in the lipophilic group (i.e., with ALOGP $\geqslant$ 2.0145) the more complex, branched drugs show increased transdermal penetration.

Mor13v, Mor26v, Mor11m and Mor09u are all members of the 3D-MoRSE (3D-Molecule Representation of Structures based on Electron diffraction) descriptors' family. These descriptors are based on the equations used for obtaining information from the three-dimensional structure in electron diffraction experiments $I(s) = f(s \cdot r_{ij})$, with s being the scattering angle and r_{ij} the inter-atomic distance between atoms i and j. The signals are unweighed (Mor09u), weighed by atomic masses (Mor11m) or weighed by van der Waals volumes (Mor13v and Mor26v).[43] For a compound and given weighing, the set of all 3D-MoRSE descriptors (32 signals corresponding to the different s) describes the diffraction pattern for the molecule, where the influence of the inter-atomic distances reflects largely the 3D-structure, as recently demonstrated by substituent-position effects within a congeneric series.[44] The abundance of 4 Mor descriptors at relatively high classification levels clearly points to the importance of inter-atomic distances.

P2v is one of the directional WHIM (Weighted Holistic Invariant Molecular) descriptors, capturing molecular 3D information related to the shape of the molecule. P2v represents the 2nd component shape directional WHIM index, weighed by the atomic van der Waals volumes. WHIM descriptors are built by performing a PCA on the centred molecular coordinates, using different weighing schemes for the atoms. While each of the 3 eigenvalues δ of the weighed covariance matrix represents a dispersion measure of the atoms projected on the considered principal axis, and hence is a measure of the axial dimension of the molecule, their fractional relationship gives an indication of the axial shape of the molecule.[45] A split value of <0.214 for P2v indicates that an elongated molecule in one of the principal axis directions, weighed for the atomic van der Waals volumes, will lead to a decreased transdermal permeability, while higher values above 0.214 for P2v, indicative for more symmetric spherical molecules, will give increased permeability.

MATS2e and GATS4e are spatial autocorrelation coefficients, respectively, two-dimensional (Moran autocorrelation lag 2) and three-dimensional (Geary autocorrelation lag 4),[58] reflecting the spatial distribution of molecular properties, both with the atoms weighed for their electronegativity.

2.5. Quantitative structure–permeability relationship (QSPR) models and their validations

2.5.1. CART. To use the CART tree as a predictor, the CART descriptors of the validation compounds are dropped down the tree until a terminal leaf. Then the predicted value of $\log K_{\mathrm{p}}$ for that validation compound is the average of the learning $\log K_{\mathrm{p}}$'s for that terminal leaf-node as visualised in Figure 2. While the RMSECV value of 1.47 is only an internal parameter to assess the quality of the model, estimation of the average prediction error was obtained through calculation of the root mean squared error of prediction (RMSEP) on the validation set, giving a value of 1.76. As additional parameters, the mean and median of the absolute values of the deviations (Δ) from the literature obtained $\log K_{\mathrm{p}}$ values are given in Table 2.

2.5.2. BRT. The boosted regression tree algorithm was applied to the dataset. The descriptors selected by the BRT-model, together with their mean importance, definition and class, are given in Supplementary data file 3. This model shows a data-fitting error, calculated from the bootstrap resampling, of 0.79% or 12.1% calculated over the range of $\log K_{\mathrm{p}}$ values for the learning set. The RMSEP for the test set is 1.14, which corresponds to 17.5% of the $\log K_{\mathrm{p}}$ value range of the learning set.

2.5.3. MLR models. The $\log K_{\mathrm{p}}$ was modelled using a multiple linear regression with the 9 CART descriptors obtained at a tree complexity of 10. This resulted in the following equation (MLR1):

$$\begin{aligned}\log K_{\mathrm{p}} = {} & -8.01(\pm 6.01)\\ & + 4.06\times 10^{-1}(\pm 5.98\times 10^{-2})\ \mathrm{ALOGP}\\ & + 6.06\times 10^{-1}(\pm 4.30\times 10^{-1})\ \mathrm{Mor13v}\\ & + 5.13\times 10^{-1}(\pm 2.21\times 10^{-1})\ \mathrm{Jhetv}\\ & - 1.40(\pm 6.41)\ \mathrm{Mor26v}\\ & + 9.71\times 10^{-1}(\pm 1.17)\ \mathrm{P2v}\\ & - 7.03\times 10^{-2}(\pm 2.86\times 10^{-1})\ \mathrm{Mor11m}\\ & - 4.84\times 10^{-1}(\pm 6.81\times 10^{-1})\ \mathrm{MATS2e}\\ & + 8.09\times 10^{-2}(\pm 1.55\times 10^{-1})\ \mathrm{Mor09u}\\ & - 1.07\times 10^{-1}(\pm 3.53\times 10^{-1})\ \mathrm{GATS4e}\end{aligned}$$

$$\text{with } N = 104,\ R = 0.636,\ S = 1.0656,\ F = 7.088,$$
$$p < 10^{-4},\ \rho = 10.4,\ \mathrm{AIC} = 1.377,\ \mathrm{FIT} = 0.344$$

where N is the number of compounds included in the model, R the correlation coefficient, S the standard deviation of the regression, F the Fisher ratio, p the significance of the model and ρ is the ratio of the number of compounds (N) to the number of adjustable parameters (p'). The statistical parameters (R, S, F and ρ) can always be improved artificially by adding more descriptors to the model. Therefore, two other parameters are introduced as well. AIC is the Akaike's Information Criterion, calculated using the formula:

$$\mathrm{AIC} = \mathrm{RSS}\cdot\frac{(N+p')}{(N-p')^2}$$

where RSS is the sum of the squared residuals or differences between the observed (y) and estimated response ($\hat{y}$) obtained from the analysis of variance (ANOVA) table.[46] The most useful model is considered to have the minimal AIC value. FIT is the Kubinyi function, which is closely related to the F-value but less sensitive towards changes in k:

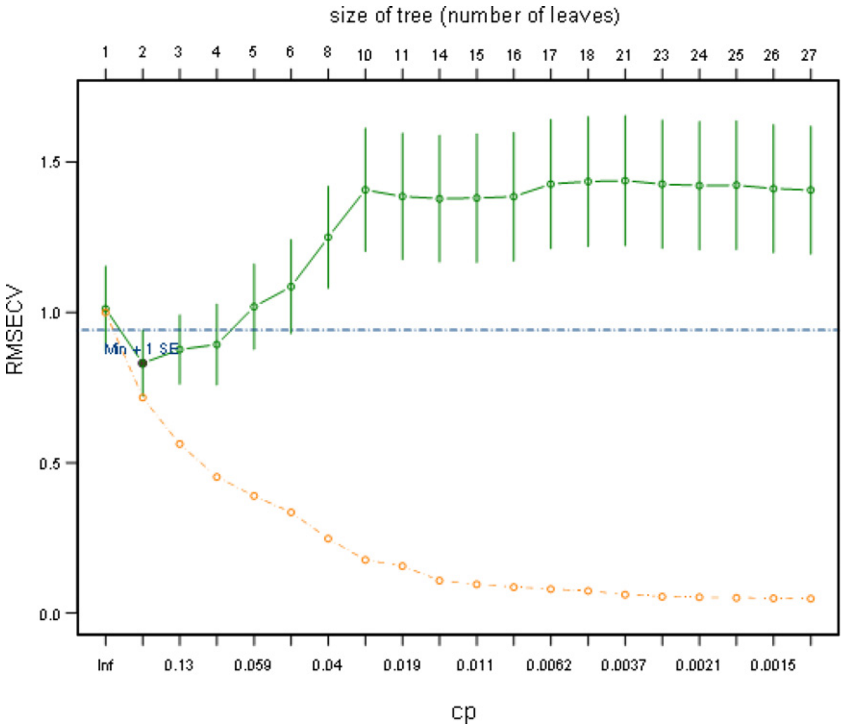

Figure 1. CART-tree complexity, represented by the size of tree, as a function of the misclassification rate estimate, represented by the root mean squared error of cross-validation (RMSECV). The horizontal line (·–·–) represents the minimal RMSECV-value plus one time its standard error. Vertical lines represent the standard error around the RMSECV values from the 10-fold cross-validation steps for the different tree complexities. cp is the complexity parameter or penalty per terminal node and is represented by the (○·–·–○)-line.

$$\mathrm{FIT} = \frac{R^2 \cdot (N - k - 1)}{(N + k^2) \cdot (1 - R^2)}$$

where k is the number of variables in the equation that describe the model. The best model will have the highest FIT value.[47]

In order to find better QSPR results, that is, increases in the values of FIT, F, R or R^2 and a decrease in the values of AIC, S or RSS, we performed a stepwise multiple linear regression analysis (MLR) as implemented in SPSS 12.0 (P-to-enter $\leqslant 0.05$ and P-to-remove $\geqslant 0.10$) using all 649 descriptors. The analysis stopped at model 25, which includes 23 variables. The results are given in Supplementary data file 4.

To avoid that the model is over parameterized, the FIT and AIC values were evaluated (Fig. 3). The FIT was clearly maximized for the models 8 up to 11 and decreased afterwards. The AIC values showed a biphasic asymptotic decrease, where the bending point intersecting both linear parts was between model 8 and 10. These results were in good agreement with the curve for R^2 and S, where bending points were located in the same regions. Therefore, a 10-dimensional MLR model (MLR2) was the best compromise between model fitness and model complexity: additional variables do not lead to such an increase in model information that justifies the increased complexity. This MLR model thus obtained starting from the 649 descriptors is given below. Statistical parameters of both the eight- and 10-dimensional models are included in Figure 3.

$$\begin{aligned}
\log K_{\mathrm{p}} = & -6.243(\pm 2.12 \times 10^{-1}) \\
& -3.14(\pm 6.17 \times 10^{-2})\ \mathrm{H.050} \\
& -1.03(\pm 2.09 \times 10^{-1})\ \mathrm{Hypertens.50} \\
& +1.04 \times 10^{-1}(\pm 5.73 \times 10^{-2})\ \mathrm{ALOGP} \\
& -4.84 \times 10^{-4}(\pm 1.05 \times 10^{-4})\ \mathrm{SRW09} \\
& +1.50 \times 10^{-1}(\pm 3.09 \times 10^{-2})\ \mathrm{RDF075m} \\
& -1.39 \times 10^{-1}(\pm 2.99 \times 10^{-2})\ \mathrm{H.052} \\
& -4.84 \times 10^{-1}(\pm 8.65 \times 10^{-2})\ \mathrm{T.(S..F)} \\
& +4.77 \times 10^{-1}(\pm 1.10 \times 10^{-1})\ \mathrm{C.025} \\
& -10.60(\pm 2.73)\ \mathrm{R1m+} \\
& -6.15(\pm 2.00)\ \mathrm{RTm+}
\end{aligned}$$

New descriptors in addition to the CART descriptors were found in this QSPR modelling to be important. H.050 (atom-centred fragment) represents the number of hydrogen atoms attached to a heteroatom, Hypertens.50 (molecular property class) is the Ghose–Viswanadhan–Wendoloski 50%-antihypertensive drug-like index,[48] SRW09 is the self-returning walk count of order 09, RDF075m is the radial distribution function 7.5 weighted by atomic masses (RDF descriptors can be interpreted as the probability distribution to find an atom in a spherical volume with radius r,[49] which is indicated by the postfix), H.052 is the number of hydrogen's attached to C(sp3) with 1 halogen attached to the next C, T.(S..F) is the sum of topological distances between S and F atoms, and C.025 is the atom-centred fragment R-CR-R. Finally, two descriptors are belonging to the

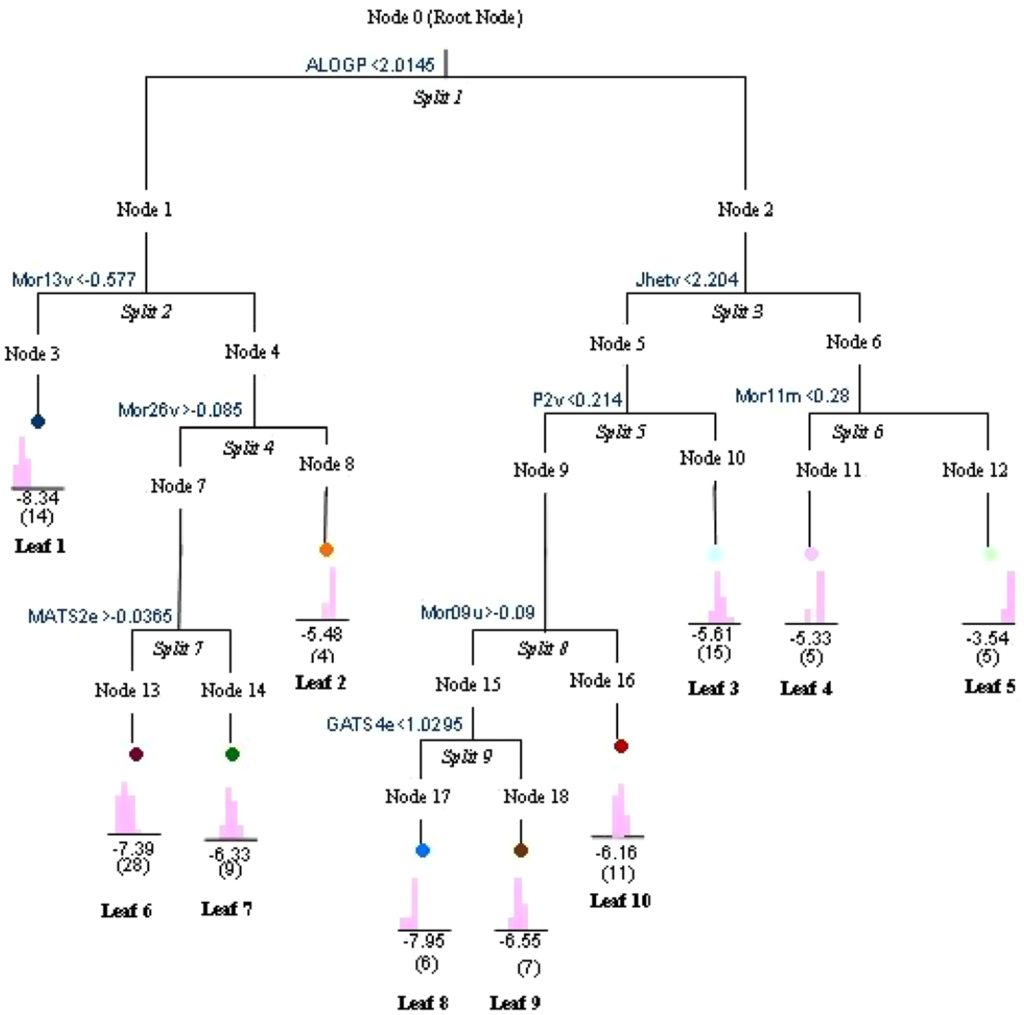

Figure 2. CART tree with complexity 10. Leaf numbers correspond with the numbers given in Table 1. For each leaf, the mean $\log K_p$ value is given, as well as the distribution and number of molecules in that leaf (between brackets). For each split, the criterion that defines the left part is indicated.

GETAWAY class: R1m+ is R maximal autocorrelation of lag 1 and RTm+ is R maximal index, both weighted by atomic masses.

For the validation of both MLR models, the descriptor values for the validation compounds were used in the models to calculate the $\log K_p$ value.

3. Discussion

The skin is a most variable organ, where not only species differences are prominent,[50] but also a multitude of possible penetration routes. Each route is biologically influenced by its anatomical place, the physio-pathological condition of the patient and has a different biological rhythm.[1,51] The 3 reference compounds suggested by the OECD, that is caffeine, benzoic acid and testosterone, fall into 3 different leaves only starting from a tree complexity of 6. At lower tree complexities, caffeine and benzoic acid fall in the same class. This indicates that if only 3 compounds are taken as models for penetration studies, alternative compounds may be more representative to cover the full spectrum of transdermal behaviour as expressed by flux parameters. The 3 OECD reference compounds were selected only according to lipophilicity, as given by their $\log P$ value. The 10 most important descriptive parameters in each of our models are given in Table 3. The most important split value selected by CART was the ALOGP descriptor. Moreover, the variable ranking CART list shows that the 4 most important parameters (i.e. Hy, ALOGP, MLOGP and BLTD48) are related to hydrophobicity. ALOGP and MLOGP are calculated $\log P$ values and BLTD48 is the Verhaar model of Daphnia baseline toxicity calculated from MLOGP. The hydrophilic factor (Hy), which is an empirical index related to the hydrophilicity of compounds, is directly, but in a negative way, correlated with $\log P$. All the other models also contain descriptors related to lipophilicity. The lipophilicity was already shown in previously reported crude models to be important: in fact, most transdermal QSAR models are including only, or as major variable, a $\log P$ descriptor, sometimes modified by a parameter indicative for the permeant size. Our modelling confirms lipophilicity/hydrophobicity as being one of the main parameters determining the transdermal behaviour of chemicals.

Next to the lipophilicity-related parameters, descriptors describing molecular complexity like Mor descriptors,

Table 2. Predictive power results with external validation set

Model			CART		BRT		MLR1		MLR2 (model 8)		MLR2 (model 10)	
Descriptors used			9		55		9		8		10	
Compound	Exp. $\log K_p$[a]	Ref.	$\log K_p$[a]	Δ	$\log K_p$[a]	Δ	$\log K_p$[a]	Δ	$\log K_p$[a]	Δ	$\log K_p$[a]	Δ
Anisole	**−4.47**	41–43	**−6.33**	1.86	**−6.47**	2.00	**−6.32**	1.85	**−6.08**	1.61	**−5.92**	1.45
Benzyl alcohol	**−5.49**	41–43	**−6.33**	0.84	**−6.58**	1.09	**−6.02**	0.53	**−5.92**	0.43	**−6.10**	0.61
Benzyl nicotinate	**−5.07**	41	**−8.34**	3.27	**−6.38**	1.31	**−6.95**	1.88	**−5.04**	0.03	**−4.88**	0.19
Butoxyethanol	**−6.41**	42	**−6.33**	0.08	**−6.62**	0.21	**−6.67**	0.26	**−6.75**	0.34	**−6.87**	0.46
Catechol	**−6.33**	42	**−7.39**	1.06	**−6.69**	0.36	**−6.06**	0.27	**−6.72**	0.39	**−6.36**	0.03
Isoquinoline	**−5.34**	41–43	**−8.34**	3.00	**−6.13**	0.79	**−5.69**	0.35	**−5.11**	0.23	**−5.02**	0.32
Morpholine	**−7.42**	42	**−6.33**	1.09	**−6.72**	0.70	**−7.20**	0.22	**−6.64**	0.78	**−6.71**	0.71
4-Nitrophenol	**−5.81**	41–43	**−5.48**	0.33	**−6.58**	0.77	**−6.12**	0.31	**−6.38**	0.57	**−6.73**	0.92
N-Nitrosodiethanolamine	**−8.81**	41,42	**−6.33**	2.48	**−6.79**	2.02	**−6.48**	2.33	**−6.76**	2.05	**−7.38**	1.43
4-Phenylenediamine	**−7.22**	41–43	**−5.48**	1.74	**−6.69**	0.54	**−6.43**	0.79	**−7.45**	0.23	**−7.09**	0.13
2,4,6-Trichlorophenol	**−4.79**	41–43	**−3.54**	1.25	**−5.62**	0.83	**−4.56**	0.23	**−6.21**	1.42	**−5.22**	0.43
3,4-Xylenol	**−5.00**	41–43	**−5.33**	0.33	**−6.34**	1.34	**−5.41**	0.41	**−5.32**	0.32	**−5.17**	0.17
RMSEP				1.76		1.14		1.08		0.93		0.73
Mean				1.44		0.99		0.79		0.70		0.57
Median				1.17		0.81		0.38		0.41		0.45

[a] K_p is expressed as cm s^{-1}.

Jhetv and radial distribution functions (RDF), are found to be important. GETAWAY descriptors, calculated from the leverage matrix obtained by the centred atomic coordinates (molecular influence matrix), are also important.

So, according to our models, lipophilicity/hydrophobicity-, 3D- and 2D-descriptors reflecting the stereochemistry of the drugs overall explain better the transdermal flux than other descriptors. This is also in accordance with previously published models, demonstrating the ability of the CART methodology to understand the transdermal behaviour of chemicals, but with more fine-tuning available, as well as to have confidence in its selection of model compounds.

Moreover, the CART clustering indicates that, for more lipophilic compounds, the extra-dimensional information encoded in a three-dimensional molecular representation is becoming less significant, while the reverse is true for the more hydrophilic compounds.

Several models were calculated and compared. The MLR1 model, having 9 CART descriptors, gave only moderate goodness-of-fit: only 40% (R^2) of the $\log K_p$ variability could be explained by the MLR1 model. As CART in fact was not meant as feature selection technique for QSPR modelling, a stepwise MLR was performed with all 649 descriptors, yielding a 23-dimensional model with satisfactory accuracy. Further evaluation of several statistical regression parameters gave a final 10-dimensional model, which explained more than 70% of the variability. Considering the wide diversity of compounds included in this study, with no single compound excluded as an outlier, and the inherent variability in literature cited experimental $\log K_p$ values, the quality of the statistical parameters for this model was excellent. As our ρ value was equal to 9.45, and thus well above the minimal value of 4 for the development of a linear model,[52] additional descriptor variables could be introduced. However, the Kubinyi FIT function clearly gave decreased values when adding more variables to the model, indicating overfitting of the model. Also the other statistics levelled off.

The prediction ability of our models was furthermore estimated by external validation, which is considered the most demanding way for predictive validation.[53] While cross-validation and response permutation techniques use the same set of compounds as for the cluster modelling, the external validation uses an independent set of data not used in the model calibration. An overview of the validation results is given in Table 2. The smallest RMSEP (0.73) was obtained with the 10-dimensional MLR2 model, while the CART model gave the worst RMSEP (1.76). Comparison with previously published models indicates that our MLR2 model scores as good as the existing models investigated (see Supplementary data file 5). Scatter plots of experimental $\log K_p$ values versus predicted $\log K_p$ values for the 12 compounds of our test set using the MLR2 and the modified Potts model[15] are given in Figure 4.

In conclusion, we provided here an analysis of some computable physico-chemical properties of known drug molecules with published skin permeability coefficients. Our main goal was to classify the drugs into a distinct number of permeability classes using the CART methodology in order to obtain a selected number of transdermal model compounds. It was shown that the OECD reference compounds caffeine, benzoic acid and testosterone, are in different groups starting from a tree complexity of 6. If only 3 reference compounds are to be chosen, benzoic acid and caffeine are in the same class according to this study. As a second objective, our different approaches show that good statistical models were obtained using parameters related to lipophilicity/hydrophobicity and molecular stereochemical complexity.

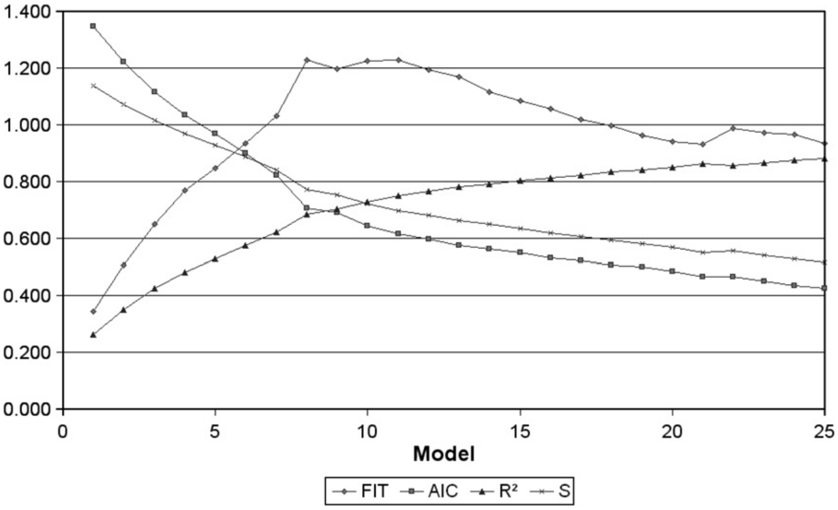

Figure 3. FIT, AIC, R^2 and S values for the MLR2 model.

4. Materials and methods

4.1. Compounds

The dataset consisted of 116 molecules extracted from literature related to the transdermal characterization of chemicals (Table 1). Most are drugs, but a few dermatologically relevant solvents are also retained. These molecules were selected because the logarithm of the K_p was given in the literature or could be calculated from the available flux data. A learning set of 104 compounds was used for constructing the CART and MLR analyses. All models were validated by external validation, using 12 compounds as prediction set, based upon structurally drug-like compounds with human transdermal flux data available in the literature. For each model, the RMSEP was calculated according to:

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}{n}}$$

with y_i the literature obtained $\log K_p$ value for object i and $\bar{y}_i$ the model-predicted $\log K_p$ value for object i.

Three-dimensional molecular structures were obtained as follows. For each of the molecules used, canonical SMILES strings were taken from the PubChem Compound database. Using ACD-Labs® 8.0 software (Advanced Chemistry Development, Toronto, Canada), these SMILES strings were converted into topological representations. The 3D-structures of the molecules were calculated using MARVIN 4.1 software (ChemAxon Ltd, Budapest, Hungary). After input of the molecule, explicit H-atoms were added and bonds were aromatized. Geometry optimization was performed and the conformer with the lowest energy value was selected. This computational optimization of the molecular structure resulted in a data matrix consisting of the Cartesian coordinates of the atoms. This data matrix was then used to calculate the molecular descriptors.

The diversity of our set of compounds was quantified using the Tanimoto coefficient (TC),[55–57] calculated as:

$$\mathrm{TC}_{kl} = \frac{\sum_{j=1}^{m}(x_{kj} \cdot x_{lj})}{\sum_{j=1}^{m}(x_{kj}^2 + x_{lj}^2 - x_{kj} \cdot x_{lj})}$$

with x_{ij} corresponding to the jth feature in the ith object, belonging to the $n \times m$ matrix for a dataset of n (here = 104) compounds characterized by a total of m (here = 649) different types of descriptors. Over the n compounds, the diversity of a dataset can be quantified by a diversity index (DI_n) which is the average value of the TC coefficients between all of the pairs of compounds in the dataset[56,57]:

$$\mathrm{DI}_n = \frac{\sum_{k=1}^{n}\sum_{l=1,l\neq k}^{n}\mathrm{TC}_{kl}}{n \cdot (n-1)}$$

The representativity of the validation set for the learning set was shown by principal component analysis (PCA) and is available as Supplementary data file 1.

4.2. Molecular descriptors

A molecular descriptor is the final result of a logical and mathematical procedure, which transforms chemical information from a symbolic representation of the molecule into a useful numeric value (theoretical descriptor) or is the result of a standardized experiment (experimental descriptor).[58] The theoretical descriptors can be classified depending on the molecular representation they are derived from. The simplest representation is the molecular formula. Descriptors derived from it are called zero-dimensional (0D). The information considered here is, for instance, the number and type of atoms, the molecular mass any functions of atomic properties (e.g. sum of atomic van der Waals volumes). One-dimensional (1D) descriptors, such as count descriptors of functional groups, rings and bonds, are derived from a substructure list representation of the molecule which consists of a list of molecular fragments (e.g. functional groups, substituents). A molecular graph contains topo-

Table 3. The 10 most important descriptors according to each model

CART		BRT	MLR2	Variable ranking CART
Primary split[a]	Surrogate split[a]			
ALOGP, BLTD48, MLOGP, ALOGP2, Hy, Mor24m	ALOGP2, MLOGP, BLTD48, MLOGP2, BEL1e	ALOGP	H.050	Hy
Mor13v, R4u, JGI4, nCrHR, Mor08v, R3u	Mor13m, R1e, Mor19v, Mor19m, Mor08v	Hy	Hypertens.50	ALOGP
Jhetv, MATS1m, HATS8u, RDF020v, Mor30m, Xindex	MATS1m, MATS6e, RDF035m, RDF040m, L1s	RDF020v	ALOGP	MLOGP
Mor26v, Mor20v, GATS1v, Mor31v, Mor29v, MLOGP2	GATS1v, Mor31m, Mor29v, Mor31v, Mv	Mor13m	SRW09	BLTD48
P2v, MATS4v, P1u, GATS6e, P2m, L.Bw	P1u, P2m, GATS7m, ASP, L.Bw	Mor03u	RDF075m	Mor20v
Mor11m, MLOGP, MLOGP2, BLTD48, Mor12u, Mor02u	MLOGP, MLOGP2, BLTD48, TI2, MATS2v	EEig04x	H.052	Mor26v
MATS2e, GATS2e, MATS1v, Hy, MATS7v, X0Av	GATS2e, GATS2m, X0Av, MATS1v, Mor18v	EEig08d	T.(S..F)	Mor24v
Mor09u, RDF090m, MATS4v, IC3, IC2, MATS5e	AAC, IC0, ASP, Mor07m, Mor18m	MLOGP	C.025	RDF020v
GATS4e, MATS4m, MATS5v, GATS5m, GATS5e, Mor26m	Mor15u, Mor15v, E1p, RBF, Rww	MATS8v	R1m+	Mor24m
—	—	Jhetv	RTm+	R3u

[a] Split criteria are arranged from left to right according to their importance.

logical or two-dimensional (2D) information. It describes how the atoms are bonded in a molecule, both the type of bonding and the interaction of particular atoms. The derived molecular properties are called 2D descriptors (e.g. total path count). Three-dimensional (3D)-descriptors are calculated from a geometrical or 3D representation of a molecule. Finally, the descriptors derived from a stereo-electronic or lattice representation are called four-dimensional (4D). The different descriptors within these classes are often further subdivided.

All molecular descriptors were calculated with Dragon® Professional, version 5.0 software. It allows calculating 1630 molecular descriptors that are divided into 20 groups: 48 constitutional descriptors, 119 topological descriptors, 47 walk and path counts, 33 connectivity indices, 47 information indices, 96 2D autocorrelations, 107 edge adjacency indices, 64 Burden eigenvalue descriptors, 21 topological charge indices, 44 eigenvalue-based indices, 41 Randic molecular profiles, 74 geometrical descriptors, 150 RDF descriptors, 160 3D-MoRSE descriptors, 99 WHIM descriptors, 197 GETAWAY descriptors, 121 functional group counts, 120 atom-centred fragments, 14 charge descriptors and 28 molecular properties. Constant descriptors in the given dataset were eliminated as they do not add discriminative values. For descriptors with a correlation higher than 0.98, only one was retained and the other eliminated. This resulted in a final set of 649 descriptors.

4.3. CART

The CART approach was developed by Breiman et al.[59] in order to build a decision tree that describes one response variable (univariate CART), for example, transdermal penetration, as a function of a number of explanatory molecular descriptors. It is a non-parametric technique that can select those variables that are most important in determining the dependent variable. If an outcome variable is continuous, CART produces regression trees; if the variable is categorical, CART produces classification trees. In both cases, CART's major goal is to produce trees justified by an accurate set of data classifiers, and uncovering the predictive structure of the problem under consideration. That means that CART helps in understanding the variables that are responsible for a given phenomenon. The CART steps are briefly summarized hereafter.

4.3.1. Tree building. The tree building process starts by partitioning the root node, consisting of all objects or molecules, using a binary split procedure based upon a very simple question of the form: is $X \leqslant d?$, where X is a variable or descriptor in the dataset and d is a real number. The root node is impure or heterogeneous, since it contains observations of mixed classes. The goal is to define a rule that will initially split these observations and create nodes that are more homogeneous than the root node. The method that CART uses for growing trees is technically known as 'binary recursive partitioning'. The splits of the root node are generated in the following fashion. Starting with the first variable, CART splits the data at all possible split points. Cases with a 'yes' response to the question posed are sent to the left node and the 'no' responses are sent to the right node. CART then applies its *goodness of split criteria* to each split point and evaluates the reduction in impurity that is achieved using the formula:

$$\Delta i(s,t) = i(t) - p_{\mathrm{L}}[i(t_{\mathrm{L}})] - p_{\mathrm{R}}[i(t_{\mathrm{R}})]$$

where s is a particular split, p_{L} and p_{R} are the proportions of cases at node t that go into the left (t_{L}) and the right (t_{R}) daughter node, respectively, and i is the impurity. For regression trees, the impurity i is usually defined as some variance impurity, that is, the total

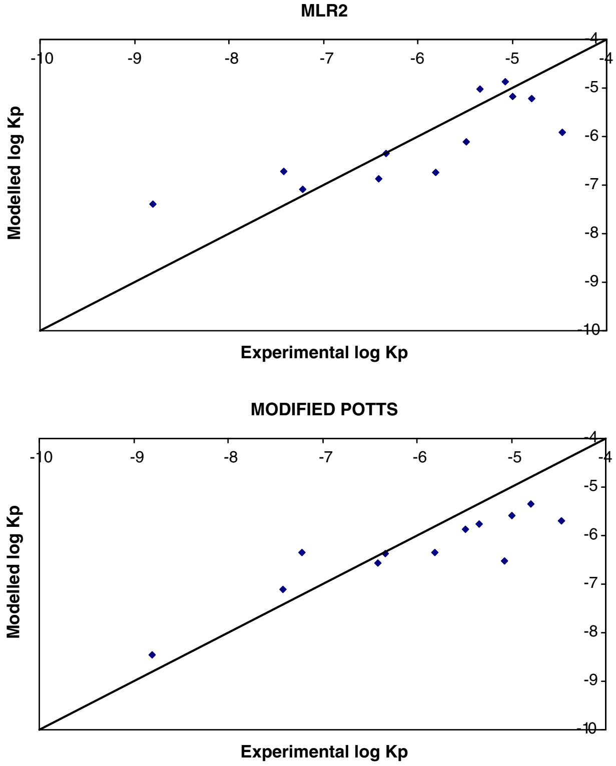

Figure 4. Scatter plots of observed $\log K_p$ values versus predicted $\log K_p$ values for the test set of 12 compounds using the MLR2 model (above) and the modified Potts model (below).

sum of squares of the deviations of the individual responses from the mean response of the group in which the considered molecule is classified:

$$i(t) = \sum_{n=1}^{n} (y_n - \bar{y}(t))^2$$

with $i(t)$ the impurity of group t, y_n the value of the response variable for object x_n and $\bar{y}(t)$ the mean of the response variable in group t. CART selects the best split of the variable as that split which maximizes the reduction in impurity $\Delta i(s,t)$. The way to find this split is as follows: impurity reduction calculations are repeated for each of the possible split values of the explanatory variables. CART then ranks all best splits according to the reduction in impurity. It selects the variable and its split point that most reduced the impurity of the root or parent node.

Because CART is recursive, this procedure is repeated to each non-terminal daughter node. In the extreme case, the maximal tree is obtained if the splitting process continues until every observation constitutes a terminal or leaf node. Obviously, such a tree shows overfitting. As with other modelling techniques, it is necessary to find a compromise between the tree complexity and accuracy, in other words, to find a properly sized tree for the problem at hand.

4.3.2. Tree pruning. Instead of stopping the splitting process at a certain number of objects in the end nodes or leaves or at a certain level of homogeneity, the tree is grown maximally and then the different branches are considered for elimination to reduce the tree size. The pruning selection methodology uses the cost-complexity measure, which is defined as $R(T) + \text{cp} \cdot n$, where $R(T)$ is the resubstitution misclassification rate or cost, cp is the complexity parameter or penalty per terminal node, and n is the number of terminal nodes. If cp = 0, then the cost complexity attains its minimum for the largest possible tree. On the other hand, as cp increases to a sufficiently large value, a tree with one terminal node (the root node) will have the lowest cost complexity. By gradually increasing cp, a series of subtrees with decreasing complexity is obtained.

4.3.3. Selection of the optimal tree. From the obtained sequence of subtrees, an optimal tree has to be selected. The selection is based on the evaluation of the predictive error, which was calculated using cross-validation. A number of objects are randomly removed from the dataset, and used as a test set to evaluate the predictive power of the tree build with the remaining learning data. In this study, a 10-fold cross-validation (CV) procedure was performed, using subsets of 90% (i.e. 94 objects). The most accurate tree is the one with the smallest cross-validation error, defined as the RMSECV:

$$\text{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$

with y_i the response value of object i, $\hat{y}_i$ the predicted response value for object i obtained from the CV subset and n the number of objects in the CV subset. The RMSECV was thus calculated by the same formula as for RMSEP, the difference being that RMSECV was calculated from CV on the learning set and not on an independent validation set. The uncertainties in the estimates of the RMSECV are given in their standard errors (SE), and the 1 SE rule states that the optimal tree is located within the ±1 SE range around the minimum in RMSECV. However, it is well recognized that this 1 SE rule is not a fixed criterion, and the analyst has the liberty of using in addition alternative selections depending on the context of the problem and prior knowledge of the variables.[59] Since our objective is to obtain a small number of representative model compound groups, the optimal tree will be chosen manually based upon the number of terminal nodes and the content of these groups.

4.3.4. Building tree models. The regression tree models were built using the Treeplus® module in the Splus® software (Mathsoft, Cambridge, MA, USA). The skin permeability data were used as continuous response variables and the different molecular descriptors as explanatory variables.

4.3.5. Variable ranking, primary and surrogate splits. It is sometimes observed that a given variable x_2 does not occur in the final tree structure, while it prominently does when another tree, which is almost as accurate as the first one, is grown after removing a so-called masking variable x_1 from the dataset. However, the variables x_1 and x_2 do not necessarily cause a similar split in the dataset: they both cause a considerable decrease in impurity. Such variables are called primary variables and the splits they cause are the so-called primary splits. The importance of the explanatory variables to introduce a split in the tree is detected by the variable ranking method in CART. The most relevant properties to describe the response variable can then be identified, so that CART can be used for feature selection.[59] On the other hand, so-called surrogate splits are defined as splits causing a similar distribution of the objects in the groups obtained after splitting. The variables responsible for these similar distributions are called surrogate variables. When for an object the value of the splitting variable is missing, the value of a surrogate variable is then used to decide to which node the object is awarded.

4.3.6. Boosted regression trees. BRT-modelling is an advanced algorithm which successively fits the classifier to the learning data, each time giving more weight to misclassified learning points. Then, regressors are combined using the weighted median, whereby those predictions that are more 'confident' about their predictions are weighted more heavily.[60] The BRT-model was built using algorithms written in Mathlab 7.0 (The Mathworks, Natic, MA, USA). Programming was done according to the original CART algorithm proposed by Breiman et al.[59] and the boosting algorithm described by Drucker.[60]

4.4. Multiple linear regression

MLR estimates the coefficients of the linear equation (involving several independent variables) that best predict the value of the dependent variable. A variety of different models from the same dataset can be obtained by using different variable selection methods. For the MLR1 model, all selected variables used were entered in a single step. We also carried out a pairwise correlation analysis of the descriptors in this MLR1 model. The highest correlation was found between Mor11m and Mor13v, with a correlation of 0.466, followed by the Mor11m–Mor26v and Mor26v–Jhetv pairs with correlations of, respectively, 0.438 and 0.338. As expected, the descriptors showed low intercorrelation, confirming the adequacy of the initial reduction process of descriptors (from 1630 to 649) for correlation.

For the MLR2 model, a stepwise procedure for variable selection was used. At each step, the independent variable not in the equation which has the smallest probability of F is entered. Variables already in the regression equation are removed if their probability of F becomes sufficiently large. The stepwise method terminates when no more variables are eligible for inclusion or removal. All MLR statistical analyses and hierarchical clustering were carried out using SPSS 12.0 (SPSS Inc. Chicago, IL, USA), with $\log K_p$ being the dependent property in the modelling.

Acknowledgements

The Ph.D. studies of Eric Deconick are funded by a specialization grant from the Institute for the Promovation and Innovation by Science and Technology in Flanders (IWT). Mireille Van Gele is a postdoctoral researcher with the Fund for Scientific Research, Flanders (FWO-Flanders).

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2007.07.050.

References and notes

1. Trommer, H.; Neubert, R. H. H. *Skin Pharmacol. Physiol.* **2006**, *19*, 106.
2. Kwinter, J.; Pelletier, J.; Khambalia, A.; Pope, E. *J. Am. Acad. Dermatol.* **2007**, *56*, 236.
3. Gorgievska Sukarovska, B.; Lipozencic, J. *Acta Dermatovenerol. Croat.* **2006**, *14*, 188.
4. Grassberger, M.; Steinhoff, M.; Schneider, D.; Luger, T. A. *Exp. Dermatol.* **2004**, *13*, 721.
5. Bos, J. D. *Eur. J. Dermatol.* **2003**, *13*, 455.
6. Thomas, B. J.; Finnin, B. C. *Drug Discovery Today* **2004**, *9*, 697.
7. European Commision. Commision Directive 94/79/EC amending Council Directive 91/414/EEC concerning the placing of plant protection products on the market, **1994**.
8. US-EPA. Subchapter E—pesticide programs. Data requirements for registration. 40 CFR 158 (revised), **1993**.
9. Zendzian, R. P. Pesticide assessment guidelines. Subdivision F: Hazard evaluation: Humans and domestic animals: US Environmental Protection Agency, **1994**.
10. Leichtnam, M.-L.; Rolland, H.; Wuthrich, P.; Guy, R. H. *J. Controlled Release* **2006**, *113*, 57.
11. Alvarez-Figueroa, M. J.; Araya-Silva, I.; Diaz-Tobar, C. *Pharm. Dev. Technol.* **2006**, *11*, 371.
12. OECD In *OECD Series on Testing and Assessment*; OECD: Paris, 2004; p 31.
13. van de Sandt, J. J.; van Burgsteden, J. A.; Cage, S.; Carmichael, P. L.; Dick, I.; Kenyon, S.; Korinth, G.; Larese, F.; Limasset, J. C.; Maas, W. J.; Montomoli, L.; Nielsen, J. B.; Payan, J. P.; Robinson, E.; Sartorelli, P.; Schaller, K. H.; Wilkinson, S. C.; Williams, F. M. *Regul. Toxicol. Pharmacol.* **2004**, *39*, 271.
14. REACH http://ec.europa.eu/environment/chemicals/reach/.
15. Wilschut, A.; Tenberge, W. F.; Robinson, P. J.; McKone, T. E. *Chemosphere* **1995**, *30*, 1275.
16. Frasch, H. F. *Risk Anal.* **2002**, *22*, 265.
17. Combes, R.; Rodford, R. In *Predicting Chemical Toxicity and Fate*; Cronin, M., Livingstone, D., Eds.; CRC Press: FL, USA, 2004; p 193.
18. Bos, J. D.; Meinardi, M. M. *Exp. Dermatol.* **2000**, *9*, 165.
19. Flynn, G. L. In *Principles of Route to Route Extrapolation for Risk Assessment*; Gerrity, T. R., Henry, C. J., Eds.; Elsevier: New York, 1990; p 93.
20. Morimoto, Y.; Hatanaka, T.; Sugibayashi, K.; Omiya, H. *J. Pharm. Pharmacol.* **1992**, *44*, 634.
21. Eltayar, N.; Tsai, R. S.; Testa, B.; Carrupt, P. A.; Leo, A. *J. Pharm. Sci.* **1991**, *80*, 590.

22. Abraham, M. H.; Chadha, H. S.; Mitchell, R. C. *J. Pharm. Pharmacol.* **1995**, *47*, 8.
23. Degim, I. T. *Drug Discovery Today* **2006**, *11*, 517.
24. Clark, D. E. *Drug Discovery Today* **2003**, *8*, 927.
25. Katritzky, A. R.; Kuanar, M.; Slavov, S.; Dobchev, D. A.; Fara, D. C.; Karelson, M.; Acree, W. E., Jr.; Solov'ev, V. P.; Varnek, A. *Bioorg. Med. Chem.* **2006**, *14*, 4888.
26. Platts, J. A.; Abraham, M. H.; Zhao, Y. H.; Hersey, A.; Ijaz, L.; Butina, D. *Eur. J. Med. Chem.* **2001**, *36*, 719.
27. Zhao, Y. H.; Le, J.; Abraham, M. H.; Hersey, A.; Eddershaw, P. J.; Luscombe, C. N.; Butina, D.; Beck, G.; Sherborne, B.; Cooper, I.; Platts, J. A. *J. Pharm. Sci.* **2001**, *90*, 749.
28. Zhao, Y. H.; Abraham, M. H.; Le, J.; Hersey, A.; Luscombe, C. N.; Beck, G.; Sherborne, B.; Cooper, I. *Pharm. Res.* **2002**, *19*, 1446.
29. Linnankoski, J.; Makela, J. M.; Ranta, V. P.; Urtti, A.; Yliperttula, M. *J. Med. Chem.* **2006**, *49*, 3674.
30. Gonzalez, M. P.; Helguera, A. M. *J. Comput. Aided Mol. Des.* **2003**, *17*, 665.
31. Gonzalez, M. P.; Helguera, A. M.; Diaz, H. G. *Polymer* **2004**, *45*, 2073.
32. Bai, J. P.; Utis, A.; Crippen, G.; He, H. D.; Fischer, V.; Tullman, R.; Yin, H. Q.; Hsu, C. P.; Jiang, L.; Hwang, K. K. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2061.
33. Deconinck, E.; Hancock, T.; Coomans, D.; Massart, D. L.; Heyden, Y. V. *J. Pharm. Biomed. Anal.* **2005**, *39*, 91.
34. Deconinck, E.; Zhang, M. H.; Coomans, D.; Vander Heyden, Y. *J. Chem. Inf. Model* **2006**, *46*, 1410.
35. Magnusson, B. M.; Anissimov, Y. G.; Cross, S. E.; Roberts, M. S. *J. Invest. Dermatol.* **2004**, *122*, 993.
36. Patel, H.; ten Berge, W.; Cronin, M. T. D. *Chemosphere* **2002**, *48*, 603.
37. Buchwald, P.; Bodor, N. *J. Pharm. Pharmacol.* **2001**, *53*, 1087.
38. Ghose, A. K.; Crippen, G. M. *J. Comput. Chem.* **1986**, *7*, 565.
39. Viswanadhan, V. N.; Reddy, M. R.; Bacquet, R. J.; Erion, M. D. *J. Comput. Chem.* **1993**, *14*, 1019.
40. Barry, B. In *Percutaneous Penetration Enhancers*; Smith, E., Maibach, H., Eds.; Taylor & Francis: FL, USA, 2006; p 3.
41. Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163.
42. Balaban, A. T.; Khadikar, P. V.; Supuran, C. T.; Thakur, A.; Thakur, M. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3966.
43. Schuur, J. H.; Selzer, P.; Gasteiger, J. *J. Chem. Inf. Model.* **1996**, *36*, 334.
44. Saiz-Urra, L.; Gonzalez, M. P.; Teijeira, M. *Bioorg. Med. Chem.* **2006**, *14*, 7347.
45. Todeschini, R.; Lasagni, M. *J. Chemom.* **1994**, *8*, 263.
46. Akaike, H. *IEEE Trans Automatic Control* **1974**, *AC19*, 716.
47. Kubinyi, H. *Quant. Struct.–Act. Rel.* **1994**, *13*, 285.
48. Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. *J. Comb. Chem.* **1999**, *1*, 55.
49. Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. *Vib. Spectrosc.* **1999**, *19*, 151.
50. Mills, P. C.; Cross, S. E. *Vet. J.* **2006**, *172*, 218.
51. Mehling, A.; Fluhr, J. W. *Skin Pharmacol. Physiol.* **2006**, *19*, 182.
52. Garcia-Domenech, R.; de Julian-Ortiz, J. V. *J. Chem. Inf. Comp. Sci.* **1998**, *38*, 445.
53. Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T.; McDowell, R. M.; Gramatica, P. *Environ. Health Perspect.* **2003**, *111*, 1361.
54. The EDETOX Database: http://edetox.ncl.ac.uk/searchinvitro.aspx.
55. Willet, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press Ltd, Letchworth, Herts. SG6 BE, England, **1987**.
56. Perez, J. J. *Chem. Soc. Rev.* **2005**, *34*, 143.
57. Yap, C. W.; Li, Z. R.; Chen, Y. Z. *J. Mol. Graphics Modell.* **2006**, *24*, 383.
58. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2001.
59. Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. *Classification and Regression Trees*; Chapman & Hall/CRC: New York, 1998.
60. Drucker, H. *Improving Regressors using Boosting Techniques*; Morgan Kaufmann: San Francisco, USA, 1997.