European Journal of Pharmaceutics and Biopharmaceutics 94 (2015) 152–159

Contents lists available at ScienceDirect

European Journal of Pharmaceutics and Biopharmaceutics

journal homepage: www.elsevier.com/locate/ejpb

Research paper

In silico modelling of permeation enhancement potency in Caco-2 monolayers based on molecular descriptors and random forest

Søren H. Welling [a,b], Line K.H. Clemmensen [b], Stephen T. Buckley [a], Lars Hovgaard [a], Per B. Brockhoff [b], Hanne H.F. Refsgaard [a,*]

[a] Global Research, Novo Nordisk A/S, Novo Nordisk Park, 2760 Måløv, Denmark
[b] Technical University of Denmark, DTU Compute, 2800 Kgs. Lyngby, Denmark

ARTICLE INFO

Article history:
Received 30 January 2015
Revised 14 May 2015
Accepted in revised form 17 May 2015
Available online 21 May 2015

Keywords:
Permeation enhancers
Caco-2
Random forest
QSAR
Surfactants

ABSTRACT

Structural traits of permeation enhancers are important determinants of their capacity to promote enhanced drug absorption. Therefore, in order to obtain a better understanding of structure–activity relationships for permeation enhancers, a Quantitative Structural Activity Relationship (QSAR) model has been developed.

The random forest-QSAR model was based upon Caco-2 data for 41 surfactant-like permeation enhancers from Whitehead et al. (2008) and molecular descriptors calculated from their structure.

The QSAR model was validated by two test-sets: (i) an eleven compound experimental set with Caco-2 data and (ii) nine compounds with Caco-2 data from literature. Feature contributions, a recent developed diagnostic tool, was applied to elucidate the contribution of individual molecular descriptors to the predicted potency. Feature contributions provided easy interpretable suggestions of important structural properties for potent permeation enhancers such as segregation of hydrophilic and lipophilic domains. Focusing on surfactant-like properties, it is possible to model the potency of the complex pharmaceutical excipients, permeation enhancers. For the first time, a QSAR model has been developed for permeation enhancement. The model is a valuable *in silico* approach for both screening of new permeation enhancers and physicochemical optimisation of surfactant enhancer systems.

1. Introduction

Development of oral delivery systems for proteins and peptides offers the promise of improved patient compliance compared to conventional parenteral administration. However, bioavailability is, in part, limited due to poor absorption of proteins across the intestinal epithelial barrier. To effectively deliver a protein systemically this barrier can be modulated by the presence of permeation enhancers [1].

Quantitative Structural Activity Relationship, QSAR methods have been applied extensively for exploration of structural properties of importance for oral absorption of new chemical entities, e.g., QSAR models have been developed for permeability [2] and solubility [3–5]. To our knowledge, no QSAR model for permeation enhancement has previously been published.

Some permeation enhancers have specific mechanisms of action, e.g., modulating the function of tight junctions in the plasma membrane such as zona-occludens-toxin [6], EDTA [7] or melittin [8]. However, the majority of permeation enhancers are primarily surfactants and will non-specifically disrupt the lipid bilayer packing of phospholipids in the epithelial membrane [1]. Surfactants are molecules having segregated lipophilic and hydrophilic domains. Water soluble surfactants tend to pool in the surfaces of water/air and water/lipid, lowering the surface tension. Lowering of surface tensions of water/air surfaces and the ability to enhance the permeability across lipid bilayers correlated

Abbreviations: C6, sodium hexanoate; C8, sodium octanoate; c8G, octylglucoside; C10, sodium decanoate/caprate; c12PC, dodecylphosphocholine; c12GPC, dodecanoylglycerophosphocholine; c14GP, myristoylglycerophosphate; CART, classification and regression tree; CDC, chenodeoxycholate; DDM, dodecylmaltoside; EDTA, ethylenediaminetetraacetic acid; GCC, glycochenocholate; GH, glycyrrhizinate; LCC, lauroylcarnitinechloride; LOO-CV, leave-one-out cross validation; MOE, molecular operating environment; PCC, palmitoyl carnitine chloride; QSAR, quantitative structural activity relationship; SM, simomenine; RMSE, root mean square error; r_p, Pearsons correlation coefficient; r_s, Spearman rank correlation coefficient; SD, standard deviation; TEER, transepithelial electrical resistance; TDM, tetradecylmaltoside; TDS, sodium tetradecyl sulphate; TDM, tetradodecyl maltoside; TC, taurocholate; T_{pot}, TEER potency; UC, Ursocholate.

* Corresponding author at: Insulin Pharmacology Research, Novo Nordisk A/S, Novo Nordisk Park, 2760 Måløv, Denmark.
E-mail address: hare@novonordisk.com (H.H.F. Refsgaard).

http://dx.doi.org/10.1016/j.ejpb.2015.05.012
0939-6411/

well for a selection of surfactant-like permeation enhancers [9]. General relations between molecular structures and physicochemical properties of surfactants are thoroughly described by Rosen [10]. Several properties of surfactants, including surface pressure, have previously been modelled with a QSAR approach applying both linear regression and non-linear machine learning models as artificial neural networks, support vector machine or random forest [5,11,12]. Combining the above mentioned concepts, it seems plausible that a QSAR-model of surfactant-like permeation enhancement could be constructed.

Our modelling is based on a Caco-2 data set for 41 surfactant permeation enhancers from Whitehead [13,14] tested in cell monolayers across three concentrations. Hereby, trade-offs between potency, pathway and safety amongst a selection of mainly surfactant-like permeation enhancers were investigated. For this article only the potency data was used. *In vitro* Caco-2 monolayers are cultures of functional, differentiated enterocytes and are widely employed to evaluate permeability rates of drug candidates or pre-formulations [15]. The Caco-2 data for permeation enhancers from Whitehead [13,14] together with molecular descriptors calculated from structure of these surfactants were the basis for the QSAR model.

Non-linear machine learning models can have superior predictive capabilities compared to classical statistical explanatory modelling. However, such machine learning models are often complex "black boxes" – difficult to interpret and discuss [16]. This article presents a promising method to elucidate the interplay of features comprising good permeation enhancers within the complex non-linear model of random forest. Therefore, based on the developed model, we here can recommend ranges of the selected molecular descriptors to obtain high permeation enhancement potency.

2. Materials and methods

2.1. Materials

Caco-2 cells (ATTC-HTB-37) were obtained from American Type Culture Collection (Manassas, VA). Cell culture media (Dulbecco's modified essential media (DMEM)) and penicillin/streptomycin were purchased from Lonza (Verviers, Belgium). All other supplements (i.e., foetal bovine serum, HEPES buffer and non-essential amino acids (NEAA)) as well as Hanks' balanced salt solution (HBSS) and trypsin were purchased from Gibco, Life Technologies (Carlsbad, CA). Corning Transwell® filter inserts (1.12 cm^2 surface area, 0.4 μm pore diameter) were purchased from Fisher Scientific (Waltham, MA). Bovine serum albumin (BSA) was purchased from Sigma Aldrich (St. Louis, MO). All other reagents were of the highest analytical grade.

2.2. Cell culture and TEER measurements

Caco-2 cells (passage numbers 41–49) were seeded at a density of 2.5×10^5 cells/flask and grown to 70–90% confluence in DMEM (supplemented with 10% FBS, 100 U/ml penicillin and 100 μg/ml streptomycin and 1% (v/v) NEAA). For transport studies, Caco-2 monolayers were cultured on permeable Transwell® 12 mm diameter inserts at a density of 10^5 cells/cm^2 and used after 14–17 days in culture. Cells were cultured at 37 °C and 5% CO_2 atmosphere and the medium was changed every other day. Monolayers were equilibrated in HBSS-based transport buffer 1 h prior to testing. Transepithelial electrical resistance (TEER) was measured with a chop-stick electrode (Millicell-ERS®, Millipore, Billerica, MA) prior to testing, and monolayers with TEER values <600 Ω cm^2 were discarded. TEER was measured after 1 h exposure to permeation enhancers.

3. Data processing

3.1. Training set

Whitehead et al., tested the ability of 51 permeation enhancers to lower the barrier integrity marker %TEER in Caco-2 cells at 1%, 0.1% and 0.01% (w/v) and published the data set as supplementary materials in two papers [13,14]. Of the 51 permeation enhancers reported, forty-two had computable molecular structures (non-mixtures) and were a wide selection of enhancers which were ascribed to 10 different categories of surfactants: Anionic surfactants, cationic surfactants, zwitterionic surfactants, non-ionic surfactants, bile salts, fatty acids, fatty esters, fatty amines, sodium salts of fatty acids, nitrogen-containing rings and others [13]. EDTA (a calcium chelator) was excluded from the training set because of a non-surfactant-like mechanism together with high potency. The remaining 41 permeation enhancers had surfactant-like structures or low potency e.g., urea could be described as an ineffective surfactant without permeation enhancement effect.

TEER-potency (T_{pot}) was defined to concatenate measurements of TEER%-decrease (EP) at the three different concentrations (0.01%, 0.1% and 1% w/v) into one target variable. T_{pot} was simply defined as the mean TEER%-decrease across the three concentrations as given in Eq. (1). $T_{pot} = 1$ corresponds to a permeation enhancer lowering TEER% completely at 0.01% (w/v) and $T_{pot} = 0$ translates to no effect of a permeation enhancer on TEER% even at 1% (w/v). The TEER%-decrease EP is defined as in Eq. (2) and depends of the TEER% before and after treatment with enhancer plus $TEER\%_+$ the background filter resistance.

$$T_{pot} = \frac{EP[0.01\%] + EP[0.1\%] + EP[1\%]}{3} \quad (1)$$

$$EP = 1 - \frac{TEER\%_{AE} - TEER\%_+}{TEER\%_{noAE} - TEER\%_+} \quad (2)$$

From a statistical point of view the loss of information is minimal, as the TEER%-values of the three concentrations were highly correlated. The loadings of the first principal component of a principal component analysis resembled the definition of T_{pot} and this principal component explained 71% of the variance. From a practical viewpoint T_{pot} could be seen as a linear approximation of pEC50 (−log effective concentration (w/v) of where 50% TEER-decrease is observed), see Eq. (3). pEC50 itself is dimensionless.

Thus, for a given permeation enhancer having a potency of pEC50 = 1 the corresponding value of $T_{pot} = 0.5$.

$$T_{pot} = \frac{pEC50 + 0.5}{3}, \quad \text{for } pEC50 \in [-0.5; 2.5] \quad (3)$$

3.2. Software packages, descriptors and model design

The open source R statistical software (v 3.02) was acquired freely from http://www.r-project.org and Rstudio integrated development environment (v 0.98.501) also acquired freely from http://www.rstudio.com. The R-package 'randomForest' (v.4.6) [17,18] was used in the random forest-QSAR model. CAS identification numbers of compounds in the training set were converted to mol-files through SciFinder [19]. Mol-files bundled in sdf-files were imported to the software application MOE [20] and sequentially pre-processed with the following functions: '*wash*' (simulating an ideal solubilised molecular form), '*partial charges MMFFA96x*' calculating the electron densities necessary for a number of descriptor algorithms, and finally '*energy minimize*' relaxing the molecule in the minimum state. All 2D molecular descriptors provided by MOE were computed. The subgroup of 3D descriptors '*vsurf*' [21] plus the single 3D descriptor '*dipole*' were calculated as they were relatively fast to compute and therefore suitable for

screening purposes. One new descriptor carbon chain length (CCL) was implemented through R. CCL is the length of the longest saturated non-substituted aliphatic carbon chain of a given permeation enhancer. Table 1 explains the simple implementation of CCL. After 266 molecular descriptors were acquired, a variable filtering was performed to increase prediction performance. First, fifteen descriptors were excluded for having the same value for more than 95% of the actual training-set. A descriptor having the same value for all permeation enhancers does not provide any information and is problematic for some algorithms which e.g., divide by the variance, which will be zero. Subsequently, 143 redundant descriptors were filtered off, one at a time, until no remaining descriptor pair-wise correlations exceeded $r_p = 0.9$ (Pearson correlation). This correlation-filtering was a simplified implementation of the CORCHOP routine [22]. Lastly, the remaining descriptors were filtered by their Spearman rank correlation coefficient (r_s) to the target variable, T_{pot}. As Spearman rank correlation utilises the target parameter (T_{pot}), it was computed separately on training data for each fold of the cross-validations, so as to avoid latently overfitting. Nevertheless, the random forest algorithm was a robust model and the root mean square error estimated by leave one out cross-validation ($RMSE_{LOO\text{-}CV}$) exhibited a variation of less than 20% for any reasonable subsets of pruning parameters. The 30 best r_s-correlating (or inverse-correlating) descriptors were included in the model. See Table 2 gives an overview of the descriptors selected for the model. Fig. 1 depicts the data flow from molecular formulas, computation of molecular descriptors, variable filtering, model training and cross-validation.

The default parameters of the random forest model were used as provided in the R-CRAN package 'randomForest', though the number of decision trees grown was set to 10,000 or 50,000 to ensure a conveniently high reproducibility between model-runs. Variable importance was computed for any descriptor and described the deterioration in prediction accuracy of the model, when permuting the particular descriptor. Variable importance was used to rank the importance of the descriptors and did not influence the model predictions. However, variable importance was a valuable tool to identify the molecular descriptors/features most important for predicting surfactant-like permeation enhancement.

To assess the mechanics of the random forest-QSAR, the package rfFC [23,24], which is a diagnostic extension for random forest, was acquired from https://r-forge.r-project.org/projects/rffc/. rfFC provides forest contributions which is the mean contribution of a given variable to the T_{pot} prediction of a given permeation enhancer.

3.3. Experimental test set

A set of 11 compounds and an additional 3 compounds from the training set were tested in Caco-2 monolayers to generate an experimental test set for validation of the developed random forest-QSAR model. Contrary to the experimental setup of the training data, the experimental test conducted for this paper differs in terms of media (HBSS versus DMEM, respectively) and incubation times (60 min versus 15 min, respectively). Likewise, morphology of Caco-2 monolayers is expected to have some inter-lab variation [25]. The most lipophilic permeation enhancers were barely soluble at 1% (w/v) at 37 °C and needed to be maintained at this temperature during the experiment at all times to avoid precipitation. Model predictions were compared to experimentally measured values of T_{pot}. The Squared Pearson correlation coefficient (r_p^2) and the root mean square error of ordinary least square fit ($RMSE_{OLS}$) were used as the validation criteria for the linear relationship between model predictions and experimental values. It is acceptable that the slope and offset deviates from 1 and 0 respectively as the absolute measured T_{pot} is method specific.

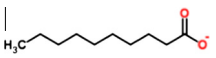
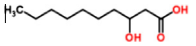
Table 2
Overview of the 30 descriptors applied in the random forest-QSAR model predicting the potency of surfactant-like permeation enhancers in Caco-2 monolayers. Descriptors were computed through MOE (18) except CCL "carbon chain length" implemented for this article.

Group of descriptors:	Amount used	Names of descriptors as available in MOE
Atom counts and bond counts	3	a_nN a_nS b_double
Kier–Hall & Kappa shape:	1	chi1_C
Adjacency and distance matrix:	8	BCUT_SLOGP_0, BCUT_SLOGP_3 BCUT_SMR_3 GCUT_PEOE_0 GCUT_PEOE_3 GCUT_SMR_0 wienerPath
Pharmacophore feature:	1	a_base
Partial charge:	10	Q_PC+ PEOE_RPC+ PEOE_PC+ Q_VSA_PPOS Q_VSA_POL Q_VSA_FPNEG PEOE_VSA_POL PEOE_VSA_FPPOS PEOE_VSA+5 PEOE_VSA+1 PEOE_VSA-1
Surface area, volume and shape:	4	vsurf_IW3 vsurf_ID8 vsurf_CP vsurf_Wp 2
Conformation dependent charge:	1	dipole
Physical properties:	1	log*P*(o/w)
New descriptor in this article:	1	CCL "carbon chain length"

3.4. Literature test set

Based on a literature search, nine permeation enhancers were included as a literature test set (Table 3). For all included

Table 1
Examples of the new descriptor carbon chain length (CCL). CCL estimates the longest sequence of saturated carbon atoms by counting the longest sequence of capital C's in a corresponding SMILES representation of the structure.

Name	Structure	Smiles underscoring	CCL count
Decanoate		O=C(O)CCCCCCCCC	9
3-Hydroxydecanoic acid		CCCCCCCC(CC(=O)O)O	8
Benzoic acid		O=C(O)c1ccccc1	1

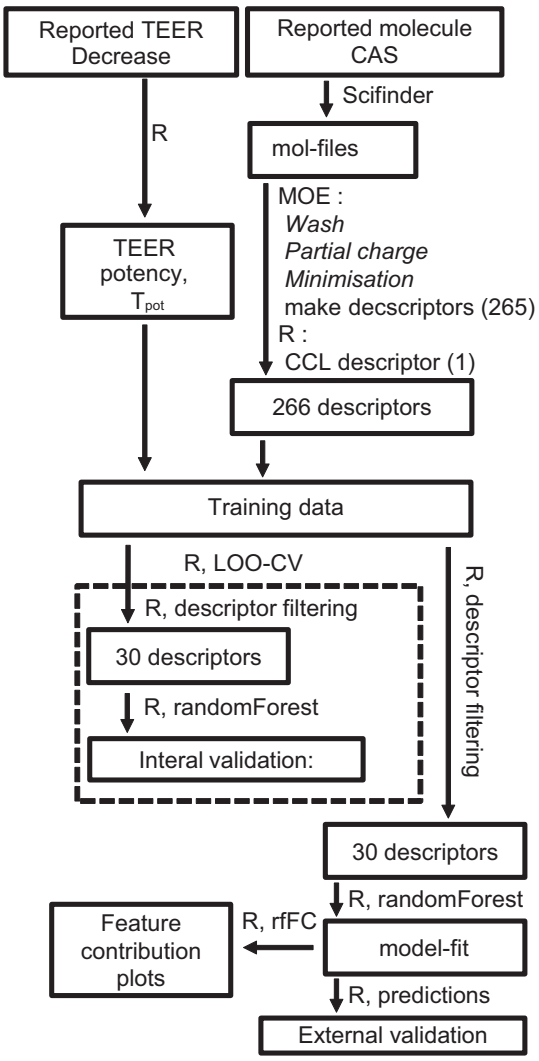


Fig. 1. Scheme of the modelling process. From top, TEER-data and provided chemical identification (CAS) were processed into TEER potency (T_{pot}) and molecular descriptors. Descriptor filtering is embedded in the leave-on-out cross-validation (LOO-CV). The final random forest model was both used for prediction of external test sets and for diagnostic interpretation through the rfFC-package. R (R) and molecular operating environment (MOE) are software applications. CCL, carbon chain length, molecular descriptor, see Table 1.

Table 3
Permeation enhancers from literature tested in Caco-2 monolayers. EC50% is the estimated concentration (w/V)% where the permeation enhancer will lower TEER 50%. pEC50 is the negative logarithm to EC50%. RF predicted potency (w/V)% is the average ability of the permeation enhancer to lower TEER at 1%. 0.1% and 0.01% (w/V)%.

CAS	Compound name	EC50, (w/V)%	pEC50	Refs.
6080-33-7	Simomenine (SM)	2.0	−.30	[26]
81-24-3	Tauro cholate (TC)	.80	−.096	[27,28]
474-25-9	Cheno deoxy cholate (CDC)	.50	.30	[28]
128-13-2	Ursocholate (UC)	.50	.30	[28]
29836-26-8	Glyco octyl (c8G)	.40	.40	[29]
68797-35-3	Glycyrrhizinate (GC)	50	.70	[30]
325465-45-0	Myristoyl glycerol phosphate (c14GP)	.10	1.0	[31]
20559-18-6	Lauryl glycerol phospho choline (c12GPC)	.025	1.6	[31]
29557-51-5	Dodecyl phosphate choline (c12PC)	.021	1.7	[31]

permeation enhancers from the literature, pEC50 was estimated by interpolation to compare across various experimentally applied concentrations. pEC50 is the negative logarithm of EC50 and has an approximate linear relation to T_{pot}, as described in (Eq. (3)). The model performance was validated by the accuracy of the pEC50 prediction for the permeation enhancer in the literature test set. Again the main criteria for comparison were r^2_p and $RMSE_{OLS}$ between interpolated pEC50 values and predicted T_{pot} values.

4. Results

A random forest-QSAR model was developed based on a 41 compound training set from literature [13,14] and permeation enhancement potency (TEER% Caco-2) values were matched with molecular descriptors. The predictability of the model was tested through validation. Three types of validation were applied: Internal leave-one-out cross validation, (LOO-CV), experimental validation and literature validation. Lastly, the mechanics from the autonomous random forest-QSAR model was extracted to provide a complimentary insight into which molecular properties there are important for permeation enhancement potency.

4.1. Model validation

Internal cross-validation was used throughout the process of designing a predictive generalisable model of permeation enhancement. Table 4 summarises the validation outcome. Both the internal and experimental validation showed $RMSE_{OLS}$ = 0.16–0.17. This error was a sixth of the entire 0–1 range of the T_{pot} scale. As T_{pot} summarises three concentration levels 1% to 0.1% to 0.01% (w/v) with a 10-fold span between each step, the accuracy was interpreted as to confirm that the model could predict within which 10-fold concentration a given permeation enhancer was effective. Likewise, for the literature validation the RMSE was 0.39, which corresponds to less than half of one unit on the pEC50 scale. One unit of pEC50 is equal to a 10-fold change in 50% effective concentration.

Fig. 2 shows plots of the three types of validations. In part A and B the predicted T_{pot} values are plotted against the measured values for the training set data from Whitehead et al. [13,14] and for the experimental data set. Fig. 2C depict the correlation between predicted T_{pot} potencies and the actual pEC50 values for the literature test set. The internal LOO validation correlation coefficient was lower, r^2_p = 0.57 (Fig. 2A), than for the external test-sets, r^2_p = 0.65–0.66 (Figs. 2B and 1C).

Eleven permeation enhancers were evaluated in Caco-2 monolayers as an experimental test set (Fig. 2B). Biotin and benzoate are widely used food additives and were intended as negative

Table 4
Summary of the validation of the random forest QSAR model predicting potency (%TEER) of permeation enhancers in Caco-2 monolayer. T_{pot}, a measure of enhancer potency defined as mean decrease of %TEER when applying 1%, 0.1% and 0.01%(w/v) in Caco-2 monolayers. pEC50 is the estimated concentration of which %TEER is decreased 50%. CV-LOO, internal cross validation – LOO. $RMSE_{ols}$, root mean square error of ordinary least square prediction fit. (a) RMSE, root-mean-square-error adjusted to compare across T_{pot} and pEC50 (Eq. (3) in Section 2).

	Training-set	Test-set, experimental	Test-set, literature
Number of enhancers	41	11	9
Data origin	1 article	Experimental	7 articles
Target value	T_{pot}	T_{pot}	pEC50%(T^a_{pot})
Model correlation, r^2_p	57% (LOO-CV)	66%	65%
Model error, $RMSE_{OLS}$	0.17	0.16	0.39(0.16[a])

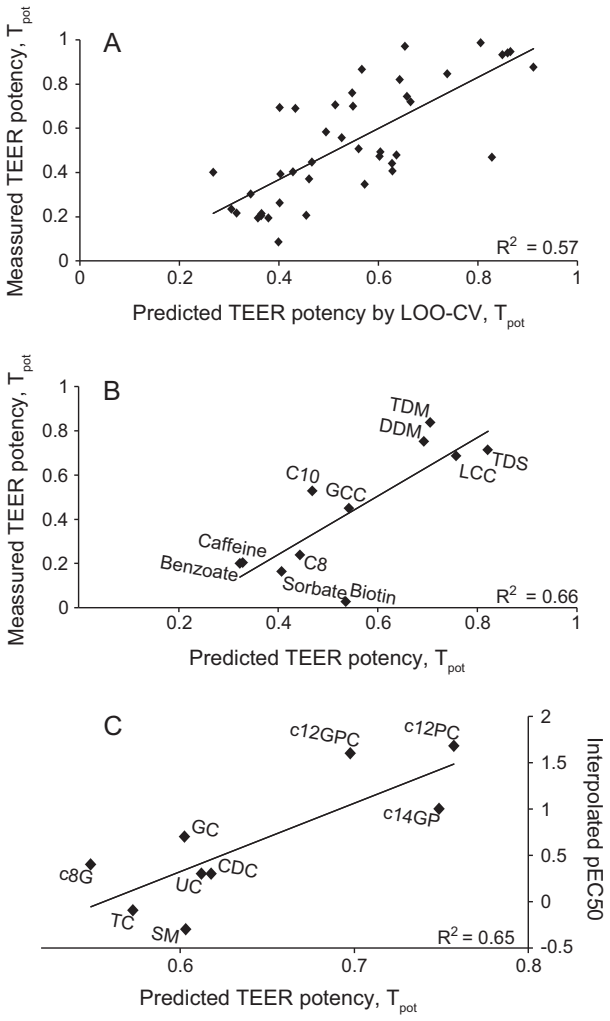


Fig. 2. Validation plots of random forest-QSAR model predicting the potency (%TEER) of permeation enhancers in Caco-2 monolayer. *X*-axis is the predicted target response of permeation enhancers' ability to lower %TEER in the Caco-2 monolayers. *Y*-axis is the true target response. (A) Internal cross-validation plot validating the predictability within the training-set, (B) validation of experimental test set, (C) validation of literature test set. T_{pot}, mean decrease of %TEER in Caco-2 monolayers of a given permeation enhancer applied at concentrations of 1%, 0.1% and 0.01%. pEC50, negated 10 base logarithm of concentration (%w/v) of 50% TEER.

controls. None of these compounds were measured to be potent permeation enhancers. All permeation enhancers in the experimental test set, with the exception of biotin, were well predicted. Biotin was predicted to elicit a moderate potency, but was devoid of any significant effects when tested in Caco-2 cells. Three compounds SDS, C6 and PCC from the training set were retested to verify, that the experimental setup applied here could reproduce findings from Whitehead et al. (data not shown).

Fig. 2C show the predicted T_{pot} potencies and the actual pEC50 values for the literature test set. The nine enhancers were surfactants with a single well defined molecular structure and sufficient data points published to estimate a pEC50 value. The range of interpolated pEC50 values from the literature data ranged from −0.3 to 1.7 corresponding to that the most potent permeation enhancer had ~100 times higher potency than the weakest. Three of the nine compounds Myristoyl glycerol phosphate (c14GP), Lauryl glycerol phospho choline (c12PC) and Dodecyl phosphate choline (c12GPC) were markedly more potent than predicted by the model. The exact predicted rankings of the second most and third most potent compounds of the experimental test-set were not correct, but within the expected uncertainty of the model. The same was seen for the group of low potent permeation enhancers.

4.2. Reviewing descriptors useful for prediction of permeation enhancement

Of the 266 descriptors assessed, 30 descriptors were applied after filtering in the model. Names and grouping of the used descriptors can be seen in Table 2 in the method section. The 16 most important descriptors were included in a Spearman rank correlation matrix depicting their internal rank correlation within the training set (Fig. 3) and their rank correlation with the target parameter T_{pot}. The strongest absolute correlation coefficient, 0.89, was between variables PEOE_VSA-1 and log*P*.oW. No correlation could exceed the correlation filter limit of 0.9, as described in Section 2. All 16 descriptors were found to be rank correlated with the target value T_{pot}. The descriptors absolute rank correlations to T_{pot} ranged from $r_s = 0.34$ to $r_s = 0.63$.

To interpret the precise contribution of each descriptor within the model, a diagnostic method termed 'Feature Contributions' [23,24] was used. A novel diagnostic plot of the feature contributions is presented in Fig. 4 The feature contributions of the 16 most important descriptors describing permeation enhancer-potency (T_{pot}) in the training-set were plotted against their respective descriptor values. This provided an intuitively graphical interpretation of how features within the random forest-QSAR model context affected the T_{pot} prediction. It represents an innovative way to graphically present the computed feature contributions. This expansion of a regular random forest model summarises the total partial descriptor contribution for any permeation enhancer in the training set. The predicted T_{pot} values for a given enhancer are equal to the sum of all partial descriptor contributions, which again is dependent of the actual feature values, as outlined in Fig. 4.

Fig. 4 shows that descriptor [2, BCUT_SLOGP_0], and [3, vsurf_ID8] had sharp thresholds separating the positive (i.e., beneficial) and negative contributions to the T_{pot} value of each permeation enhancer. For [1, dipole] there was also a separation between positive and negative contribution to the T_{pot} value.

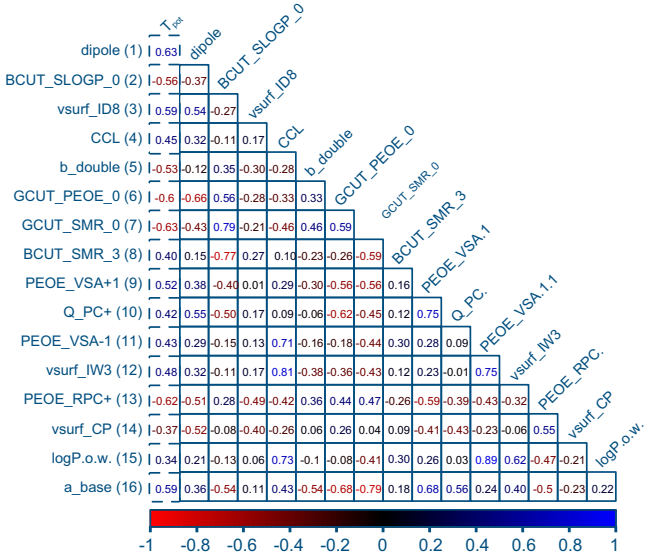


Fig. 3. Spearman-correlation matrix of the 16 most important molecular descriptors listed by decreasing variable importance (most important first) within the random forest-model. The target variable T_{pot}, the ability of permeation enhancers to lower electrical resistance in Caco-2 monolayers across concentrations 0.1–1%(w/v), has also been included.

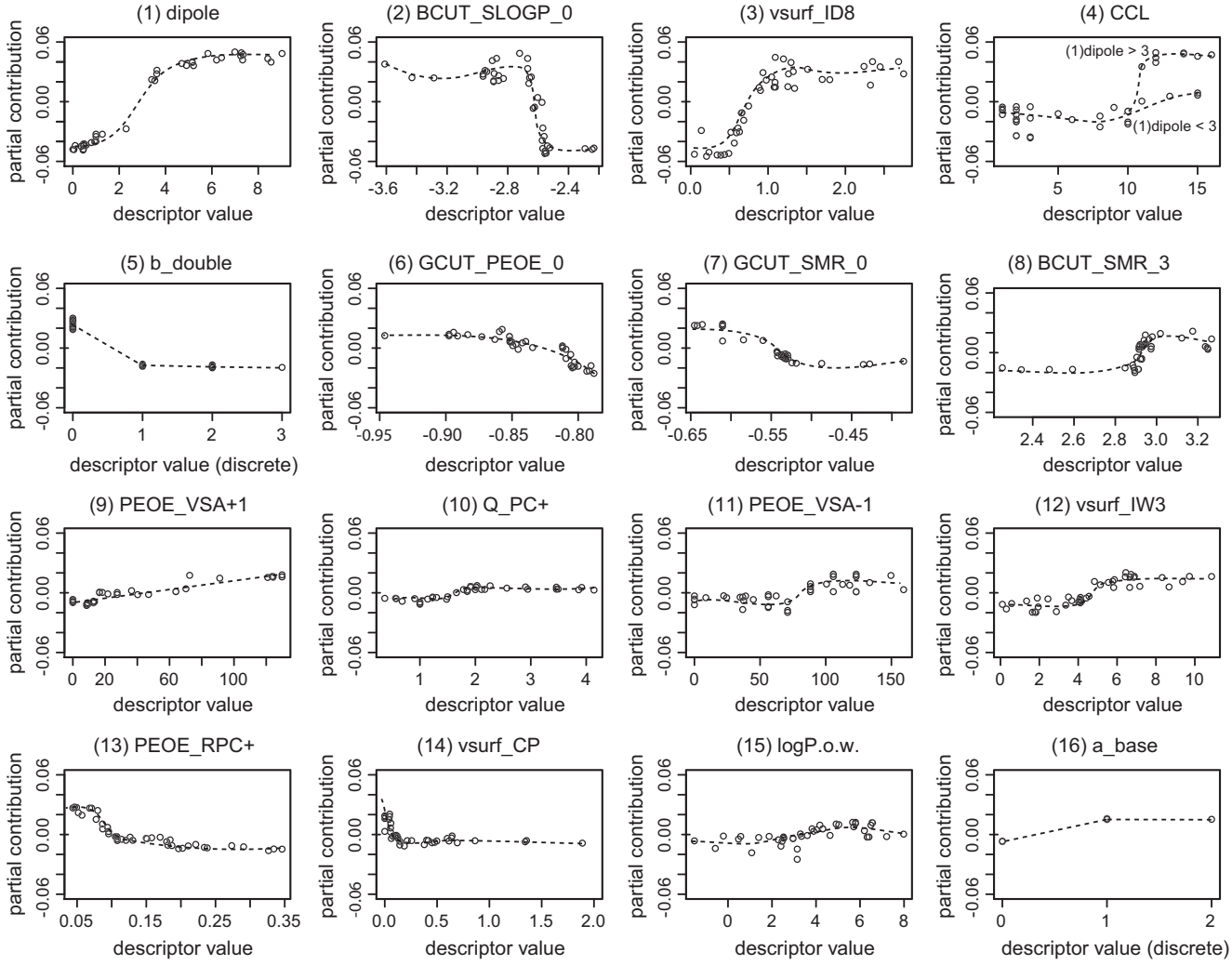


Fig. 4. Random forest feature contributions diagnostics. Scatter plots of partial descriptor contributions versus the individual descriptor values for each observation of the training set for the 16 descriptors with highest 'importance' within the random forest model. Scatter plots enumerated in descending order of 'importance'. Dashed trend lines were added to the plots. Each plot outlines a partial function of a variable as its role in the model.

Good permeation enhancers, compounds with high T_{pot} values, had high dipole > 3, [2, BCUT_SLOGP_0] below −2.7 and [3, vsurf_ID8] > 1.

[1, dipole] reflected the overall dipole moment calculated from the partial charges of the molecule. The descriptor contribution of [1, dipole] within the random forest model was well described as a function of the descriptor value itself. Thus, there was no interaction with other descriptors. On the contrary, the descriptor contributions of descriptor [4, CCL], the maximum aliphatic carbon chain length, varied for many permeation enhancers having the exact same chain length. This pointed to an interaction phenomenon between descriptors. When only emphasising permeation enhancer above the threshold value for dipole > 3, these permeation enhancers were all accredited positively for having a high CCL value. Oppositely, enhancers with dipole < 3 were accredited neutral for any CCL value. The interpretation drawn was that molecules with a high dipole moment are likely to have a hydrophilic domain and if combined with an aliphatic carbon chain of length > 10, the molecules are likely to have surfactant properties. Conversely, compounds with no significant hydrophilic groups such as oils, would not function as enhancers alone despite long carbon chains.

Descriptor [3, vsurf_ID8] reflected the hydrophilic domains separation from the lipophilic domains which was expected to be a central surfactant-like property. More precisely, the [3, vsurf_ID8] reflected the distribution of hydrophobic or hydrated domains and their distance from the mass centre. It was observed that the model evaluated low [14, Vsurf_CP] values as being beneficial. [14, Vsurf_CP] is a micelle critical packing parameter. Cone shaped surfactants would in generally have a low [14, Vsurf_CP] value and thereby a low critical packing number. A low critical packing number favours micellar aggregation, not liposomal. Throughout Fig. 4, the descriptors were generally declining in magnitude of feature contributions, and thus less influential.

5. Discussion

The random forest-QSAR model of permeation enhancement in Caco-2 cells was shown to provide reasonable estimates of permeation enhancer potency. Such a model has to the knowledge of the authors not been developed previously. Within the paradigm, that surfactant-like properties are key features of most enhancers, it was confirmed that a model could be constructed inspired by the *in silico*, *in vitro* and *in vivo* models of surfactant surface tension depression [9,11,12].

In the case of a new modelling area, a large amount of descriptors could possibly become useful. However, the relatively small

size of training examples makes such a selection process challenging. The training set of 41 permeation enhancers and 266 descriptors is an example of sparse (small n – large p) modelling which can lead to overfitted non-generalisable models. Filtering/pruning constants, redundant and non-correlated descriptors improve model performance.

The ensemble method, random forest, is an extension of the classification and regression tree, CART. Decision tree models branch out/split data into increasingly smaller sub groups of samples having the most similar target response. Such CART decision trees are highly adaptable of many types of data, but also easily overfitted to the training data. Thus, the model becomes adaptable but highly noise sensitive. The random forest model is an extra layer to the CART, reducing noise without losing its adeptness. In short, random forest is an ensemble of many of such uncorrelated decision trees (e.g., 500). Though each tree is susceptible to random inference, the average prediction of many decision trees have been shown to be much less prone to overfit thus its conclusions/predictions are more generalisable across data sets [17]. That said, the conclusions from any model approach will always be limited by the diversity of the training set. In this example, scientific literature, the basis of this model training, tend to have a bias towards not reporting any compounds which lack an enhancing effect. In the case of this training set, all compounds elicited at least a low enhancement effect. A feature of decision tree-based models is that, they cannot extrapolate beyond the target range (T_{pot}) of the training set. Likewise, caffeine and benzoate were predicted in absolute terms to be more potent than experimentally measured, as no learning examples could suggest such a weak potency. Nonetheless, this is not of much practical concern in terms of predicting new permeation enhancer candidates. A useful model does not have to distinguish very weak enhancers from non-enhancers.

C10, sodium decanoate, is one of the most described enhancers in literature [7,9,32,33]. Amongst the reported mechanisms of C10 are phosphorylation cascades and intracellular calcium signalling leading to tight junction opening [9,33]. Such mechanisms are far too complex to be captured from a training sample of this size. Conceivably, this may be why C10 was predicted to be a mediocre permeation enhancer, yet elicited a relatively stronger potency (Fig. 2B), caused by components not captured by the model. It is expected that doubling or tripling the size of the training set would improve prediction accuracy significantly. This would require testing another 40–80 permeation enhancers in three concentrations in Caco-2 monolayer.

Fig. 4, the feature contributions versus descriptor values of each training permeation enhancer, provides a novel and very useful way to learn from the random forest-model. For example, a molecule having a dipole > 3, a Vsurf_ID8 > 1 and a BCUT_SLOGP_0 < −2.7 and CCL > 10 would appear to bear promising starting point. Furthermore, the data in Fig. 4 suggested interaction for e.g. CCL > 10, only contributing positively conditioned when dipole > 3. That carbon chain length is only conditionally advantageous matches the general understanding of surfactant-like properties. Such simple rules can help to understand what modifications of an enhancer can be made without incurring a loss of potency. The abundance of partial charge related descriptors (see Table 2) was interpreted as a consequence of, that most surfactants have one or more polar domains neighbouring carbon hydride domains and an induced dipole moment across the border [12].

The feature contributions technique represents a novel approach to data analysis and has the potential to be employed as a powerful explorative tool within many scientific areas. QSAR models based on algorithm models such as random forest are designed to map associations (not necessarily causal) between features and the target parameters to optimise predictions. It should be noted that this is also the case for classical statistical approaches [16]. Nevertheless, as discussed above, the suggestions from the feature contributions are plausible causal from a physicochemical point of view.

Other core aspects relating to oral protein formulation such as solubility, stability and metabolism are not encompassed in the existing approach. Thus, their inclusion is necessary in order to yield a fully predictive model of protein permeation. When designing/screening for new enhancers as excipients in protein-based drug formulations, various other requirements, such as solubility, should be considered.

Thus, by applying the described *in silico* model an *a priori* prediction of the permeation enhancer potency of a surfactant can be determined based upon its structure and hence obviate the need for extensive permeability screening of novel compounds.

6. Conclusions

Random forest-QSAR modelling utilising molecular descriptors calculated from the molecular structure was shown useful for predicting permeation enhancer potency. Although absorption of proteins is a complex biologic phenomena, the surfactant-like properties of permeation enhancers comprise a relatively manageable component.

Sparse data combined with the biological noise (unexplained) component is a challenge to build a robust predictive model. To reduce the estimation error, the prediction challenge was alleviated in two ways:

(1) TEER readings of three concentration levels were joined into a single value target (T_{pot}) to create an approachable modelling question: Is the potency of a new surfactant-like enhancer high, medium or low?
(2) Filtering of correlated descriptors to reduce redundant information and to remove descriptors with no univariate correlation to target parameter was performed to avoid too many descriptors being progressed to the random forest model with few training examples.

From the validations employed i.e., internal cross-validation, experimental validation and literature validation, the model was found to predict potency of permeation enhancers. Furthermore, it was possible to extract common structural features for high potency enhancers. Such knowledge is useful to assess the credibility of the built model and/or inspire our understanding of what makes a surfactant-like permeation enhancer potent.

Hereby, we have outlined how to robustly perform *in silico* screening for permeation enhancers with non-linear random forest, with the possibility to assess and learn from the model. The provided QSAR model forms a good basis for a systematically approach for the development of oral therapeutics formulated with potent permeation enhancers.

Acknowledgments

Christian Vind assisted with molecular descriptors and Sten B. Christensen assisted with literature search.

This work, a part of an industrial Ph.D project for Søren Welling, was granted by The Danish Agency for Science, Technology and Innovation and the company Novo Nordisk A/S.

Søren Welling, Hanne Refsgaard, Stephen Buckley and Lars Hovgaard are employees and/or shareholders of Novo Nordisk A/S.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ejpb.2015.05.012.

References

[1] B.J. Aungst, Absorption enhancers: applications and advances, AAPS J. 14 (2012) 10–18.
[2] H.H.F. Refsgaard, B.F. Jensen, P.B. Brockhoff, S.B. Padkjaer, M. Guldbrandt, M.S. Christensen, In silico prediction of membrane permeability from calculated molecular parameters, J. Med. Chem. (2005) 805–811.
[3] B. Fredsted, P.B. Brockhoff, C. Vind, S.B. Padkjær, H.H.F. Refsgaard, In silico classification of solubility using binary k-nearest neighbor and physicochemical descriptors, Mol. Inform. 26 (2007) 452–459.
[4] D.S. Palmer, N.M. O'Boyle, R.C. Glen, J.B.O. Mitchell, Random forest models to predict aqueous solubility, J. Chem. Inf. Model. 47 (2007) 150–158.
[5] L.D. Hughes, D.S. Palmer, F. Nigsch, J.B.O. Mitchell, Why are some properties more difficult to predict than others? A study of QSPR models of solubility, melting point, and log P, J. Chem. Inf. Model. 48 (2008) 220–232.
[6] A. Fassano, S. Uzzau, Modulation of intestinal tight Junctions by zonula occludens toxin permits enteral administration of insulin and other macromolecules in an animal model, J. Clin. Invest. 99 (1997) 1158–1164.
[7] M. Tomita, M. Hayashi, S. Awazu, Absorption-enhancing mechanism of EDTA, caprate, and decanoylcarnitine, J. Pharm. Sci. 85 (1996) 608–611.
[8] S. Maher, L. Feighery, D.J. Brayden, S. McClean, Melittin as an epithelial permeability enhancer I: investigation of its mechanism of action in Caco-2 monolayers, Pharm. Res. 24 (2007) 1336–1345.
[9] W.J. Xia, H. Onyuksel, Mechanistic studies on surfactant-induced membrane permeability enhancement, Pharm. Res. 17 (2000) 612–618.
[10] M.J. Rosen, Surfactants and Interfacial Phenomena, third ed., Hoboken, New Jersey, 2000.
[11] Z.W. Wang, D.Y. Huanga, G.Z. Li, X.Y. Zhang, L.L. Liao, Effectiveness of surface tension reduction by anionic surfactants-quantitative structure-property relationships, J. Dispers. Sci. Technol. 24 (2003) 653–658.
[12] J. Hu, X. Xhang, Z. Wang, A review on progress in QSPR studies for surfactants, Int. J. Mol. Sci. 11 (2010) 1020–1047.
[13] K. Whitehead, S. Mitragotri, Mechanistic analysis of chemical permeation enhancers for oral drug delivery, Pharm. Res. 25 (2008) 1412–1419.
[14] K. Whitehead, N. Karr, S. Mitragotri, Safe and effective permeation enhancers for oral drug delivery, Pharm. Res. 25 (2008) 1782–1788.
[15] B. Sarmento, F. Andrade, S.B. da Silva, F. Rodrigues, J. das Neves, D. Ferreira, Cell-based in vitro models for predicting drug permeability, Expert Opin. Drug Metab. Toxicol. 8 (2012) 607–621.
[16] G. Shmueli, To explain or to predict?, Stat Sci. 25 (2010) 289–310.
[17] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[18] A. Liaw, M. Wiener, Classification and regression by randomForest, R News 2 (2002) 18–22.
[19] SciFinder Scholar, version 2014, Chemical Abstracts Service. Columbus, OH, 2014; RN (multiple look-ups, +50).
[20] Molecular operating environment (MOE), 2012-2013.8, Chemical Computing Group Inc., 1010 Sherbrooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2013.
[21] G. Cruciania, P. Crivorib, P.A. Carruptb, B. Testab, Molecular fields in quantitative structure–permeation relationships: the VolSurf approach, Theochem 503 (2000) 17–30.
[22] D.J. Linvingstone, E. Rahr, CORCHOP – an interactive routine for the dimension reduction of large QSAR data sets, Quant. Struct. – Act. Relat. 8 (1989) 103–108.
[23] A. Palczewska, J. Palczewskiy, R.M. Robinson, D. Neagux, Interpreting random forest classification models using a feature contribution method, in: T. Bouabana-Tebibel, S.H. Rubin (Eds.), Advances in Intelligent Systems and Computing: Integration of Reusable Systems, Springer, Heidelberg, 2014, pp. 193–218.
[24] V.E. Kuz'min, P.G. Polishchuk, A.G. Artemenko, S.A. Andronati, Interpretation of QSAR models based on random forest methods, Mol. Inform. 30 (2011) 593–603.
[25] R. Hayeshi, C. Hilgendorf, P. Artursson, P. Augustijns, B. Brodin, P. Dehertogh, et al., Comparison of drug transporter gene expression and functionality in Caco-2 cells from 10 different laboratories, Eur. J. Pharm. Sci. 35 (2008) 383–396.
[26] Z. Lu, W. Chen, A. Viljoen, J.H. Hamman, Effect of sinomenine on the in vitro intestinal epithelial transport of selected compounds, Phytother. Res. 24 (2010) 211–218.
[27] U. Werner, T. Kissel, M. Reers, Effects of permeation enhancers on the transport of a peptidomimetic thrombin inhibitor (CRC 220) in a human intestinal cell line (Caco-2), Pharm. Res. 13 (1996) 1219–1227.
[28] S. Michael, M. Thöle, R. Dillmann, A. Fahr, J. Drewe, G. Fricker, Improvement of intestinal peptide absorption by a synthetic bile acid derivative, cholylsarcosine, Eur. J. Pharm. Sci. 10 (2000) 133–140.
[29] P.P. Tirumalasetty, J.G. Eley, Permeability enhancing effects of the alkylglycoside, octylglucoside, on insulin permeation across epithelial membrane in vitro, J. Pharm. Pharm. Sci. 9 (2006) 32–39.
[30] M. Sakai, T. Imai, H. Ohtake, H. Azuma, M. Otagiri, Effects of absorption enhancers on the transport of model compounds in Caco-2 cell monolayers: assessment by confocal laser scanning microscopy, J. Pharm. Sci. 86 (1997) 779–785.
[31] D.Z. Liu, E.L. Lecluyse, D.R. Thakker, Dodecylphosphocholine-mediated enhancement of paracellular permeability and cytotoxicity in Caco-2 cell monolayers, J. Pharm. Sci. 88 (1999) 1161–1168.
[32] S. Maher, T.W. Leonard, J. Jacobsen, D.J. Brayden, Safety and efficacy of sodium caprate in promoting oral drug absorption: from in vitro to the clinic, Adv. Drug Deliv. Rev. 61 (2009) 1427–1449.
[33] T. Lindmark, Y. Kimura, P. Artursson, Absorption enhancement through intracellular regulation of tight junction permeability by medium chain fatty acids in Caco-2, J. Pharmacol. Exp. Ther. 284 (1998) 362–369.