RSC Advances

Computational approaches for the prediction of the selective uptake of magnetofluorescent nanoparticles into human cells

Journal:	*RSC Advances*
Manuscript ID	RA-ART-03-2016-007898.R1
Article Type:	Paper
Date Submitted by the Author:	30-Jun-2016
Complete List of Authors:	PAPA, ESTER; Universita degli Studi dell\'Insubria Dipartimento di Scienze Teoriche e Applicate, ; Universite Paris Diderot, ITODYS Doucet, Jean-Pierre; Universite Paris Diderot, ITODYS Panaye, Annick; Universite Paris Diderot, ITODYS
Subject area & keyword:	Nanomedicine < Nanoscience

SCHOLARONE™
Manuscripts

Please do not adjust margins

RSC Advances

ARTICLE

Computational approaches for the prediction of the selective uptake of magnetofluorescent nanoparticles into human cells

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

E.Papa*,[a]J.P. Doucet[b] and A. Doucet-Panaye[c]

The use of functionalized nanomaterials is of high importance in biomedical applications like the efficient targeting of cancer cells. This paper proposes a comparison of different statistical and mechanistic aspects of new QSAR models generated to predict the selective uptake of a library of surface modified nanoparticles tested in different human cell types. Additionally, a new approach based on the combination of multivariate factorial analysis and QSAR is proposed to generate a 2-dimensional map of the selective uptake of the surface modified nanoparticles into multiple cell types. This map offers an immediate view of the uptake of the nanoparticles, distinguishing among those with high or low uptake in one or more of the studied cells. Finally, QSAR models are generated to predict the coordinates of the studied nanoparticles in the 2D map from their molecular structure. This predictive map is useful to screen new and existing surface modified nanoparticles for diagnostic and biomedical uses.

Introduction

The use of nanomaterials in biomedical applications is of high importance because of the ability of these materials to give specific properties, s10uch as efficient targeting. For instance, magnetofluorescent nanoparticles are among the materials recently studied for their possible application for diagnostic use such as the specific targeting of cancer cells.[1,2]

In a recent study Weissleder and colleagues[1] described the development of a library of 146 magnetofluorescent, surface modified nanoparticles characterized by the same nano-core (iron oxide) and different surface modifications, as the result of the conjugation of the nanoparticles with small organic molecules. The cellular uptake of 109 of these magnetofluoescent nanoparticles was tested on multiple human cell types in order to evaluate the specific cellular affinities, and to gain information useful for the generation of target-specific nanomaterials without a priori knowledge of the potential activity.[1]

Several studies describe *in silico* computational approaches based on Quantitative Structure Activity Relationships (QSAR), applied to the uptake responses measured for this library of 109 NPs.[3-11]

QSAR approaches identify a quantitative relationship between a measured endpoint and the structural representation of the compounds of interest, which is mathematically encoded by structural descriptors.[12-15] The chemical structure (*X*) and the measured response (*Y*) are linked by means of a mathematical function (*f*) so that: $Y=f(X)$.

QSAR approaches are powerful tools, which can be applied for the prediction of missing data for existing or new chemicals. They find application in medicinal chemistry, toxicology and ecotoxicology to predict missing data, to perform pre- or post-synthesis screenings, drug design, hazard assessment, and for regulatory purposes as alternatives to animal testing. [12-15]

QSAR-type approaches have recently found application in the field of nanotoxicology. [3-11,16, 17] In particular, different models have been proposed to predict the uptake in Pancreatic Human Adenocarcinoma epithelial cells (PaCa2)[3-11] and in Human Umbilical Vein endothelial Cells (HUVEC)[4,10] measured for the library of 109 NPs created by Weissleder and colleagues.[1]

All these models share similar predictive ability and different complexity since they are based on different linear, non linear and combinatorial approaches, and involve different combinations of molecular descriptors. However, with the exception of studies by Chau and Kar[6,8], the other studies do not provide an in depth investigation of the anomalies associated with the models, such as the presence of outliers or other problematic chemicals among the surface modifiers. In their recent paper Chau and Yap[6] suggested a further elaboration of the original data published by Weissleder[1], and a classification model specific for the uptake into PaCa2 cells, based on a combinatorial approach. Interestingly this paper reported a detailed analysis of the outliers detected in the model and highlighted possible issues related to these misclassifications, such as the limitations in the domain of the model. Additionally 56 NPs with good/moderate cellular uptake selective for PaCa2 were highlighted. However, no effort was spent to evaluate the selectivity of the 109 surface modifiers taking into account the uptake in cells different from PaCa2.

[a] QSAR Research Unit in Environmental Chemistry and Ecotoxicology, DiSTA, University of Insubria, via J.H. Dunant 3, 21100, Varese, Italy.

Université Paris Diderot, Laboratoire ITODYS, UMR 7086, 15 rue Jean de Baïf, bâtiment Lavoisier, 75013, Paris, France

[b] Université Paris Diderot, Laboratoire ITODYS, UMR 7086, 15 rue Jean de Baïf, bâtiment Lavoisier, 75013, Paris, France

[c] Université Paris Diderot, Laboratoire ITODYS, UMR 7086, 15 rue Jean de Baïf, bâtiment Lavoisier, 75013, Paris, France

Electronic Supplementary Information (ESI) available: Supplementary Information Figures, and Supplementary Information Tables. See DOI: 10.1039/x0xx00000x

Please do not adjust margins

Aim of this paper is to propose new externally validated QSAR models to predict the cellular uptake of 109 surface modified NPs into PaCa2 and HUVEC cells. The comparison of eight modelling methods based on different linear and non-linear functions by using different statistical approaches, the in depth analysis of response outliers and structural aberrations, as well as of the linkage between the chemical class of the surface modifiers and the uptake behaviour, represent some among the multiple innovations proposed in this study. In addition, multivariate analysis was combined with QSAR to generate an interactive 2D map of the uptake tendency of the 109 NPs into different human cell types. This map was applied to predict and screen the selective uptake of 28 new surface modifiers with unknown uptake in human cells. This last approach represents the main innovation proposed in this study, i.e. a QSAR-based screening tool of the potential uptake into multiple cells of new surface modifiers.

Results and discussion

Modelling the uptake into PaCa2 and HUVEC cells

Multiple linear regression models based on ordinary least squares (MLR-OLS)[12,13] were developed on log transformed data to predict the uptake of 109 surface-modified NPs into pancreatic adenocarcinoma (PaCa2) and human umbilical vein cells (HUVEC). The list of surface modifiers and the related uptake data are reported in Table S1. The best models among the best combinations of descriptors selected by genetic algorithm (GA) were chosen by maximizing the internal robustness and external predictive power, evaluated on multiple test sets and several validation metrics. [18,19] The statistical parameters calculated for the best MLR-OLS models developed for PaCa2 and HUVEC cells are reported in Table 1 and in Table S2.

<Table 1>

The best combinations of modelling descriptors selected in MLR models were applied to generate QSARs based on several linear and non linear techniques.[12-32]

Results obtained for the validation of the linear and the non-linear models on five external test set are reported in Table 2 and commented in the following paragraphs.

<Table 2>

Models for PaCa2 cells

The best MLR model selected by the genetic algorithm was based on 8 molecular descriptors. Equations of the externally validated models (i.e. split models), with variables listed in order of importance according to standardized coefficients are reported in Table S3. Plots of experimental vs. calculated values are reported in Figure S1A-S1E.

Table 1 and S2 show that the MLR model is robust and has good comparable internal and external predictivity on the basis of the values of the validation parameters calculated for the five splittings. Moreover, external predictions are better than predictions generated for the training set, if evaluated on the 95% of the NPs in the prediction sets.

Table 2 and Figure S2 show the comparison among performances of linear and non linear models calibrated using the 8 descriptors selected by the Genetic Algorithm.[18]

The analysis of the robustness of the models was conducted by calculation of the residuals and of statistical parameters informative for the dispersion of the error such as the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE).[18,19]

In particular MAE and RMSE values were calculated from predictions generated for the 109NPs only when used as external prediction set in the 5 splittings. Residuals calculated for these predictions are listed in Table S4. Results reported in Table 2 are similar to best results reported in literature for the same dataset, i.e. results obtained by Ensemble learning based nano-QSAR modelling (MAE prediction0.17 (DTB method) - 0.19 (DTF method); RMSE prediction0.22 (DTB method) – 0.25 (DTF Method)).[9]

Moreover, linear and non linear models had similar performance and the quality of predictions was always classified as “good” by the software Xternal Validation Plus[19]. Linear SVM performed better than Radial SVM. K-NN, and GRegNN, which were the most complex approaches among those applied, had lower performances than RBFNN. In general, Radial SVM, K-NN and GRegNN were the least performant methods with the largest RMSE and MAE values. The variation of MAE and RMSE values calculated on the 95% of the prediction set NPs in the five splittings, compared to statistics calculated on the 100% (Table 2, Figure S2), show that the external performances are negatively influenced by a limited number of compounds (i.e. the 5%) with large residuals.

Finally, the average of predictions calculated by the different linear and non-linear models (combinatorial approach), did not generate results which exceed the performance of the best model (i.e. the MLR model). These results clearly highlight that the MLR model can be applied without losing predictive power, instead of more complex non-linear methods.

Three compounds were identified as problematic across the different models i.e. N-methylisatoic anhydride (n° 37), N,N'-Bis(2-aminoethyl)-1,3-propanediamine (n°81) and Diethylenetriamine (n° 76). These chemicals presented always residuals larger than 0.5 log units (Figure S3A – S3E) and the predictions did not improve by applying the combinatorial approach. ID nos°37 and 81 were also detected as outliers with standardized residuals larger than 2.5 standard deviation units. Interestingly, nos° 76 and 37 were highlighted as outliers in the former work by Chau and Yap.[6]

The analysis of the applicability domain (AD) by the leverage–based[12,13,18] and the standardized descriptors-based[33] approaches, led to very similar results (Table S5).

A few surface modifiers were identified outside the AD by the two approaches (Figure S3A – S3E: Palmitic anhydride (n°47), pentafluopropionic anhydride (n°3), 5-chloroisatoic anhydride (n°18), isatoic anhydride (n°36), and bicyclo [2,2,2] oct-7-ene-2,3,5,6-tetracarboxylic anhydride (n°24)). ID nos° 3, 24 and 47 were characterized by rather complex structures with long chains, multiple halogens and ring systems, which explained their influence on the model. Differently, chloroisatoic and isatoic anhydrides (i.e. nos 18 and 36) were detected as high leverage and have experimental uptake values (logPaCa2 = 4.44 and 4.18, respectively) rather different from the response outlier N-methyl isatoic anhydride (n° 37; logPaCa2=3.36), which was always overestimated by the models. It is clear that nos°18 and 36 influenced the model, because of their molecular structure (heteroaromatic cyclic anhydrides with

This journal is © The Royal Society of Chemistry 20xx

two 6 membered fused rings), however the difference in the experimental uptake values across these three anhydrides was the reason for inaccurate prediction of n°37. In addition, n° 29 (lauricanyhydride), n° 82(Pentaethylenehexamine) and n°95 (L-Arg) were highlighted as outside the AD mainly by the Roy's approach[33], while n° 109 (diethylenetriaminepentaacetic dianhydride) only by the leverage approach (Table S5).

Principal Component Analysis (PCA)[34] was performed on the residuals in prediction (Table S4) in order to evaluate similarities across the methods, and to identify the most problematic compounds in the dataset, i.e. NPs predicted with large residuals independently of the method. Results from this PCA are reported in Figure S4-A and are consistent with results calculated by combinatorial approach, i.e. the three outliers (nos 37, 76 and 81) were on the extreme left and on the extreme right of PC1, which explained more than 80% of the total variance.
The loading plot (Fig. S4-B) shows the general similarity of the results calculated by the different approaches (similar weights on PC1) already observed by analysing the performances of the models (Table S2); however PC2 distinctly grouped the best and the worst approaches into opposite clusters.
Moreover, we analysed the distribution of the compounds and of the residuals within the main structural groups in the dataset (i.e. anhydrides, amines, and aminoacids). Figure S4-A shows that residuals were quite evenly distributed across the structural categories; however, the largest residuals mainly belonged to the amines class.

Finally, we analysed the modelling descriptors in order to provide some interpretation of the structural features mainly involved in the uptake into PaCa2.
Standardized coefficients calculated for the eight modelling descriptors selected in the logPaCa2 model gave the following order of importance (the signs of the contribution in the MLR equation is reported in brackets):
nBase (-) >VE3_Dzs (+) > ATSC1v (-) > BIC2 (-) > D070 (-) > maxHxxNH (+) > MATS7s (-).
Correlations among descriptors, response and additional descriptors used for the interpretation of the models are given in Table S6.
These descriptors encode for information mainly related to the presence of basic groups (nBase), and in particular of basic nitrogen, the presence of heteroatoms, and the topological complexity. VE2_Dze, VE3_Dzs are based on the Barysz weighted distance matrix Dz and account simultaneously for the presence of heteroatoms and multiple bonds in the molecule. The autocorrelation descriptors ATSC1v MATS7s are based on the topological distance matrix weighted on van der Waals volumes and on the electrotopological state, respectively. These five descriptors had negative sign in the equations of the models, with the only exception of VE3_Dzs. The other descriptors included in the models were Bonding information content index (BIC2), which takes into account the number and the typology of bonds, the final heat of formation (D070) and one electrotopological state descriptor related to secondary amines (maxHssNH).
As a general observation, we want to highlight that the uptake of NPs into cells is a complex phenomenon governed by multiple mechanisms and influenced by several factors such as NPs size, shape, surface charge and presence of the protein corona. [35-37] Recent literature shows that experimental work is still necessary to clarify the cellular uptake phenomenon and the role of different aspects characterizing the structure of the NPs in different experimental conditions. [36,37]
Therefore, since specific information regarding these factors was not available for the studied NPs, the interpretation of molecular descriptors selected in the models is an *a posteriori* description of the structural features of the surface modifiers, possibly associated with one or more of the aforementioned mechanism/factors. In addition, it is necessary to bear in mind that since QSAR models are usually the result of the combination of two or more molecular descriptors, a straightforward depiction of the mechanistic function of each descriptor is in most of the cases challenging or impossible. [12,13]

In the case of the response logPaCa2 the selection of descriptors in the model matched with the complexity and the heterogeneity of the dataset, which included amines and amino acid based NPs, differently enriched with basic groups, as well as anhydrides with different dimension, shape and presence of heteroatoms. These structural features can influence hydrogen bonding and lipophilicity of the studied NPs, which were highlighted in other studies[3-10, 35-37] among the most relevant properties influencing the uptake into PaCa2. This is consistent with the correlation of nBase (i.e. the most important descriptor in the equation, and with negative sign) with the number of H bond donors[38] (0.81), a descriptor that was not selected in the model but was helpful for the interpretation. This correlation suggests that an increase in the number of H bond donors (i.e. nHBDon in Table S6) would diminish the potential uptake into PaCa2 cells.
Moreover, the descriptor BIC2 has large positive values for simple, small chemicals characterized by low hydrophobicity, and has a negative sign in the equation. Therefore, an increase in the hydrophobicity of the surface modifiers, i.e. a decrease in BIC2 values, may enhance the uptake into PaCa2 cells. This was confirmed by the negative correlation among BIC2 and the descriptor XlogP[38] (-0.61, Table S6).

Models for HUVEC
The best MLR model developed for HUVEC was based on 8 molecular descriptors and had satisfactory fitting and predictive power tested on multiple external prediction sets (Table 1, Table S2, Figure S5A-S5E). The averaged performances calculated for the five splittings and the performances calculated for the full model (i.e. model calibrated on uptake data for all the 109 NPs) were reported in Table 1. These performances were comparable to those of the literature model published by Epa[4] (R^2values: 0.63 and 0.74 for prediction and training sets respectively). However, the literature model had higher complexity being based on 11 descriptors instead of the 8 selected in the new model. Statistics reported in Table S2 confirmed the comparable or better fitting of our MLR model (R^2ranges: 0.56-0.80 and 0.72-0.77 for the prediction and the training sets, respectively) than the literature model. It also demonstrated the robustness and the predictivity of the new HUVEC model evaluated by several parameters such as, cross validated Q^2lmo (range: 0.63 – 0.70), CCCtr (range: 0.84 – 0.87), and multiple parameters calculated for the external validation, i.e different measures of Q^2ext (range: 0.57 – 0.80) and CCCext (range: 0.74 – 0.89). As mentioned above, the model was fairly robust and predictive,

although the random composition of the prediction set influenced the predictivity of the models. In particular in split M3, the model was sensitive to the inclusion of 4-amino-1,8-naphthalic anhydride (n°48) in the prediction set. This chemical induced the largest uptake in HUVEC, and fell outside the AD of the model (Figure S6A-S6E). This example showed that the structural and experimental information associated with NP n°48, which was always well fitted with the exception of M3, was important to stabilize the model and to enlarge the AD.

The results obtained from the application of the selected descriptors to develop QSAR models based on additional linear and non linear techniques, and the related residuals in prediction were reported in Table 2 and Table S7, respectively, and in Figure S7. All the methods had very similar performances with exception of GRegNN and SVM-RAD, which had again the lowest performances and the largest RMSE and MAE values. The combination of the predictions calculated by the different linear and non linear approaches into averaged predictions slightly increased the performance of the models taken singularly. Also in this case all the predictions were judged as of "good quality" by the statistics calculated on the 95% of the external prediction sets[19] (Tables 1, 2 and S2). Moreover, we observed a reduction in the total number of NPs with residuals larger than 0.5 log units, from 14 in MLR and PPR (which were the best approaches) to 11 in the combinatorial model. Among these, Diglycolic anhydride (n° 105), cis-aconitic anhydride (n°107), 1,3-dimethylbutylamine (n°58), 4-amino-1,8-naphthalic anhydride (n° 48) and 4-nitro-1,8-naphthalic anhydride (n°27) had residuals between 0.5 and 1, while 3-hydroxyphthalic anhydride (n°16) and N-methylisatoic anhydride (n°37) had residuals larger than 1 log unit. Possible explanation for some of these errors was that similar structures had rather different uptake value. In the case of 1,3-dimethylbutylamine (n°58) the uptake into HUVEC was larger (logHUVEC: 4.14) than values measured for similar branched amines (i.e. nos 61-66, logHUVEC range: 2.97 -3.91). In the case of 3-hydroxyphthalic anhydride (n°16) the value of uptake (logHUVEC is 3.25) was lower than similar anhydrides (i.e. nos°30-33, logHUVEC range: 4.18 – 4.45); in the case of N-methylisatoic anhydride (n°37) logHUVEC=2.15 was more than 1 log unit lower than isatoic anhydride (n°36, logHUVEC: 3.59). Furthermore, the comparison of the AD calculated using the leverage-based and the standardized descriptors-based approaches led to comparable results (Table S8). In particular, 1,4,5,8-naphtalenetertracarboxylic anhydride (n°13), 3-nitro-1,8-naphthalic anhydride (n°15), 1,2,4-benzenetricarboxylic anhydride (n°30), 4-amino-1,8-naphthalic anhydride (n°48), and diethylenetriaminepenta-acetic dianhydride (n°109), fell outside the AD of the split and the full models. The fact that the four amino-and nitro-1,8 naphthalic anhydrides (i.e. nos°13, 15, 27 and 48) were detected as problematic compounds, because of large residuals, or falling outside of the AD of the model, may indicate that the model was lacking information necessary for the accurate prediction of their uptake into HUVEC. Moreover, nos° 16 and 37 were confirmed as outliers with standardized residuals larger than 2.5 standard deviation units.

Finally, the PCA performed on the residuals (Table S7, Figure S8A,B) was consistent with results calculated for the logPaCa2 response (Figure S4A,B), and with the combinatorial approach calculated on HUVEC predictions. The outliers with the largest residuals in individual models laid on the extreme left and on the extreme right of PC1, which explained the 90% of the total variance (Figure S8A). The loading plot (FigureS8B) showed the general similarity of the results calculated by the different approaches (similar weights on PC1) already observed by analysing the RMSE values; however, the best and the worst approaches were distinctly grouped along PC2. Single compounds isolated on the extreme opposite sides of PC2 (i.e. nos° 13 and 109) were due to large residuals in one or more of the variables with the heaviest weight on PC2. The distribution of the compounds within the main structural groups in the dataset (i.e. symbols in Figure S8A) showed that largest residuals belonged mainly to the anhydrides class.

The eight molecular descriptors selected in the MLR model ranked according to their standardized coefficients, in the following order of importance (signs of the contribution in the MLR equation are reported in brackets): D106 (+) >nF10Heteroring (-) >D346 (+) >VR3_Dt (+) >D094 (-) >MATS3s (-) >ATSC4i (-) >MATS8s (-).

These descriptors encoded, on one hand, for structural information related to local reactivity and electrostatic properties, such as D106 (Minimum partial charge for a N atom), D346 ((1/2)X Beta Polarizability) and D094 (minimum nucleophilic reaction index for O atoms). On the other hand, they reflected molecular topological complexity, such as nF10HeteroRing (Number of 10-membered fused rings containing heteroatoms i.e. N, O, P, S, or halogens). ATSC4i (CenteredBroto-Moreau autocorrelation - lag 4 / weighted by first ionization potential), MATS3s, and MATS8s (Moran autocorrelation - lag 3 and 8 / weighted by I-state), or molecular size, such as VR3_Dt (Logarithmic Randic-like eigenvector-based index from detour matrix).

We observed that the shift from more negative to more positive values of D106 increased the uptake into HUVEC, as well as polarizability (i.e. D346) contributed positively to the uptake. Additionally we observed that D106 was inversely correlated to the best modelling descriptor selected for logPaCa2 i.e. nBase (correlation =-0.70), and with the descriptor number of H bonds donor (correlation= -0.74). Therefore, taking into account the positive or negative signs of D106 and nBase descriptors in the respective equations, these results confirm that the presence of H bonds donors may have a negative effect also on the cellular uptake in HUVEC. Moreover, descriptors related to molecular topological complexity and presence of heteroaromatic rings had negative sign in the model. In particular, the presence of large heteroaromatic rings decreased the uptake into HUVEC (nFHeteroRings was negative in the equation). Finally, VR3_Dt, is an index that increases with the complexity of the molecule in terms of number of bonds. This descriptor has positive sign in the equation and is positively correlated to XlogP (correlation=0.54). This suggests that an increase in the number of bonds, which influences the size, the stability, and the hydrophobicity of the studied molecules, may increase the potential uptake into HUVEC.

Concluding, MLR was here the best modelling option since no better result was generated with other methods to predict the uptake into PaCa2 or HUVEC cells. Linear (PLS) and non linear methods optimized by linear functions (SVM-Linear and PPR)

This journal is © The Royal Society of Chemistry 20xx

had comparable performances to MLR and represented a valid alternative to MLR. Neighbourhood based methods and SVM based on Radial functions had the lowest predictive ability and appeared to be less suitable to model the studied datasets. Few surface modifiers were predicted with residuals larger than 0.5 log units in multiple models and after application of the combinatorial approach. These compounds (i.e. nos°16, 27, 37, 48, 58, 76, 81, 105 and 107) were problematic because induced large variability of the response associated with small variations in the structure of the surface modified NPs, or because the models lack of sufficient structural or experimental information (e.g. some surface modified NPs may be under-represented in comparison to other NPs). Finally, inaccurate predictions may be caused by inaccurate experimental values. A new experimental determination of the cellular uptake of NPs functionalized by nos°16, 27, 37, 48, 58, 76, 81, 105, and 107 may help to confirm the nature of the associated error in prediction (e.g. possibly due to mechanistic or experimental causes).

Map of the selective uptake of magnetofluorescent nanoparticles for multiple cell types

We developed an approach based on multivariate Factorial Analysis[39] in order to provide a map of the selective uptake of the studied NPs for multiple cell types. The final aim was to provide a tool to prioritize NPs on the basis of their multiple cell selectivity as function of the surface modifications, and to evaluate the efficiency of the different chemical classes used to functionalize the NPs.

FA was used since results obtained by PCA generated loadings of each variable, which appeared to be orthogonal, but were placed in between PC1 and PC2 (Figure S9A-B). Therefore, we introduced FA in order to rotate the view and obtain a distribution of the variables more clearly differentiated along the new factors.

We started from the raw data of uptake available for the 109 functionalised NPs measured in five different cell types. The raw uptake data were log transformed, and the FA was performed by applying varimax rotation on the covariance matrix, in order to not overweight the information associated with the different macrophage cells, which we knew to be less responsive to surface modifications than PaCa2 and HUVEC from former studies.[1]

Results from the factorial analysis were reported in Figure 1, Figure S10 and Table S9.

<Figure1>

Figure 1 is a simplified map obtained by plotting Factor 1 and Factor 2 extracted by Factorial analysis. Factor 1 (F1) explains the 60% of the total variance and Factor 2 (F2) the 27%. The 109 surface modified NPs are ranked from right to left along F1 according to decreasing potential uptake mainly into HUVEC, which is the variable with the largest loading value in F1 (Figure S10); NPs are ranked from the top to the bottom along F2 according to decreasing uptake into PaCa2 cells (largest loading value in F2). The uptakes in macrophages are correlated to PaCa2, however they have small values of the loadings, and therefore low influence, in defining the projections of the 109 NP in the F1-F2 space.

Summarizing, NPs placed on the right side of the plot enhance the uptake mainly in HUVEC, while those placed at the top of F2 enhance the uptake mainly in PaCa2 cells. The interpretation of Figure 1 can be simplified by dividing the map in four quadrants, numbered from 1 to 4 moving clockwise. Compounds falling in quadrant 1 (positive score values for F1 and F2) enhance the uptake in all the cell types but GMCSF_Mph (which has negative loadings along F1 and F2). Compounds placed in quadrant 2 (positive and negative F1 and F2 scores, respectively) enhance the uptake in HUVEC but have negative influence on uptake into PaCa2 cells. Compounds placed in quadrant 3 decrease the uptake in all the cell types, while compounds placed in quadrant 4 and in particular those at the top of the map (positive F2 values) enhance the uptake mainly in PaCa2 cells.

According to classes highlighted in Figure 1 it is easy to see that anhydrides are the surface modifiers associated with the largest uptake in all the cell types. In particular glutaric anhydride and its heteroaromatic derivatives (i.e. nos°15, 18, 34, 36, 50) increase the uptake into PaCa2, but inhibit the uptake in HUVEC; large and long chained anhydrides enhance the uptake in HUVEC. The increase in the number of bonds increase the uptake in HUVEC (e.g uptake of succinic anhydride(n°102)<itaconic anhydride (n°104)< cis-aconitic anhydride (n°107)). Nitro substituents increased uptake in HUVEC if they were attached to a single aromatic ring; however, they increased the uptake in PaCa2 if they were attached to a multiple rings system (e.g. 4-nitrophtalic anhydride (n°11), 3-nitrophtalic anhydride (n°32), and 3-nitro-1,8-naphtalic anhidride (n°15)). Linear Amines increased selectivity for PaCa2 and inhibited uptake in HUVEC, while aromatic amines as well as amino acids (quadrant 3 in Figure 1) inhibited the uptake in all the cell types. Branched amines and diamines (quadrant 4, figure 1) had a negative effect on the uptake in HUVEC.

Modelling and prediction of the selective uptake of NPs

Two QSAR MLR-OLS models were generated using F1 and F2 scores as responses, to predict the possible position of new NPs in Figure 1. These models allow for the prediction of the potential selective cellular uptake of NPs from the molecular structure of the surface modifiers. In order to provide external validation F1 and F2 values were split into training and prediction sets. Additionally, in order to show an example of application of the proposed approach, we included 28 NPs listed in the library developed by Weissleder[1] but with unknown cellular uptake values.

The equations for F1 and F2 QSAR models are reported below (descriptors are in order of importance according to standardized residuals). Plots of the experimental vs. predicted values and AD for model F1 and F2 are reported in Figures S11A-B and S12A-B, respectively.

F1=-3.60+2.51PubchemFP614 + 1.76 PubchemFP393 - 0.94 ATSC1p-1.27n6HeteroRing+2.90MATS2s+ 0.82PubchemFP430+1.14GATS2m-0.36minsCH3+0.65VP-7+1.77AATSC7p (1)

N°tr.=88; N°test=21; R^2=0.75; Q^2loo=0.70; Q^2lmo15%=0.69; CCC_{CV}=0.83 R^2_{EXT}=0.78; Q^2_{EXT}Average=0.78; CCC_{EXT}=0.87; RMSEtraining=0.50; RMSEext100%=0.45; RMSEext95%=0.37 MAEext100%=0.36; MAEext95%=0.30

F2=1.55 -0.74MLFER_E–4.77MATS1v+0.02VE3_Dzs-1.60GATS2c+4.57VE1_Dze+0.67PubchemFP637+0.63n6HeteroRing-0.15minHBint3+0.40ATSC8p+0.16GATS6m (2)

N°tr.=88; N°test=21; R^2=0.73; Q^2loo=0.65; Q^2lmo15%=0.65; CCC_{CV}=0.80; R^2_{EXT}=0.79; Q^2_{EXT}Average=0.84; CCC_{EXT}=0.87; RMSEtraining=0.54; RMSEext100%=0.39; RMSEext95%=0.31 MAEext100%=0.31; MAEext95%=0.26

As demonstrated by the values calculated to quantify the internal and external predictivity, considering statistics calculated using 100% or 95% of the external prediction set, the models are robust and predictive, also when they were tested on NPs never included in the model development (i.e. prediction set). The variables that were selected in the two models were similar to those previously selected for individual models in PaCa2, and HUVEC cells. Features like topological complexity, presence of heteroatoms and rings and hydrogen bonding were still present in the new models, and were encoded by several ATS-, MATS-, and GATS-type descriptors (i.e. autocorrelation descriptors), VE3_Dzs and VE1_Dze, n6Heteroring (number of 6 membered heteroaromatic rings), VP7(topological path cluster of order 7) and two electrotopological indices (i.e. minHBint3 and minsCH3). The relevance of these features was furthermore confirmed by some newly selected descriptors such as the excess molar refraction (MLFER_E), which encoded for interactions associated with the polarizability of pi- and n- electrons (E = 0 for saturated alkanes). [40]

Some fingerprints were selected, which encoded for specific substructures in the molecules. FP 393 is a simple atom nearest neighbours counter for the fragment N (~C) (~H). FP 430 is a detailed atom neighbourhoods counter for the pattern C(-C)(-C)(=C). FP 614 is a SMART pattern that describes the presence of the sequence C-C-O-C-C regardless of the count, and was a fundamental descriptor to distinguish between anhydrides and other substituents. FP 637 is also a SMART pattern encoding for the presence of the sequence O-C-C-C-C regardless of the count.

The analysis of the applicability domain (Figure S11B-C; S12B-C) showed that the performances of the models were influenced by a few outliers (i.e. nos° 16 and 42 in F1 and n° 81 in F2), and high leverage compounds (i.e D-Glucosamine (n°59), 4-amino-1,8-naphthalic anhydride (n°48), 5-chloroisatoic anhydride (n°18), N-methylisatoic anhydride (n°37) and 2-sulfobenzoic acid cyclic anhydride (n°19) in F1; pentafluopropionic anhydride (n°3), Palmitic anhydride (n°47), diethylenetriaminepentaacetic dianhydride (n°109) in F2.

Not surprisingly, most of these problematic compounds were identified before as outliers or high leverage in models developed for uptake into PaCa2 and HUVEC cells.

Finally, the two models were applied to predict the F1 and F2 coordinates of 28 new molecules which could be used as surface modifiers, and therefore to screen their selectivity before testing.

The new F1-F2 plot including predictions generated for the new molecules was reported in Figure 2 and S13.

<Figure 2>

The analysis of the applicability domain of the two models (Figures S14 and S15 A,B,C) showed that three compounds out of the 28 tested fell outside the AD of F1 model (i.e. 3,4,9,10-perylenetetracarboxylic dianhydride (n°124), N-Ethyl-N-(2-hydroxyethyl)-4-(4-nitrophenylazo)aniline (n°136) and 3-Mercapto-2-methylpropionyl-L-proline (n°137)). However six compounds fell outside the AD of F2 model (i.e. Hexafluoroglutaric anhydride (n°122); heptafluorobutyric anhydride (n°116); 3,4,9,10-perylenetetracarboxylic dianhydride (n°124); N-2,4-DNP-L-arginine (n°132), Isobutyric anhydride (n° 121) and trymethylacetic anhydride (n°112)). Among these chemicals nos°132, 121 and 112 had large leverage values, and estimated values above and below the experimental range of the training set compounds. This means that these predictions are unreliable and should be discarded.

It was interesting to note that the new molecules were correctly placed in the areas occupied by anhydrides, amine and aminoacids, according to their chemical identity. Figure 2 and S13 showed clearly that aminoacids and amines surface modifiers do not enhance the uptake into the different cell types, while anhydrides are the group that may induce the highest uptake, according to QSAR predictions. ID nos° 112, 121 and 122 among the new surface modifiers are those with the highest selectivity for PaCa2 cells. Unfortunately, we have already explained that these three predictions fall outside the AD of the model, and should be either discarded (i.e. nos° 112 and 121 are largely outside the structural and response domain), or taken carefully (i.e. 122 falls just outside the structural AD of the model (Figure S15)).

Conclusions

In this paper we presented new models useful for the prediction of the uptake of heterogeneous magnetofluorescent NPs with the same core, into different human cell types. The new QSAR models developed for the uptake into PaCa2 and HUVEC cells were consistent with mechanistic findings presented in current literature.[3-11] Several structural features related to electrostatic properties, topological complexity and hydrophobicity were associated with the uptake of the studied NPs into PaCa2 and HUVEC cells. Among the main observed effects, the increasing number of hydrogen bond donors reduces the uptake into PaCa2 and HUVEC cells. In addition, features related to hydrophobicity played an important role and were associated with an increase in the cellular uptake.

Additionally, we showed how the parallel use of different modelling techniques is helpful to identify problematic compounds. Nanoparticles with surface modifiers as nos° 16, 27, 37, 48, 58, 76, 81, 105 and 107 should be newly tested in order to clarify the nature of the error associated with their predictions. An important result would be to confirm large differences in the uptake of very similar structures. This issue, named “activity cliff”, which was originally highlighted by Maggiora[53], can be the cause of errors in prediction across similar structures and should be taken into account in read across procedures. Results reported in this study show that read across would be unsuitable for NPs such as nos°16, 37 and 58, and similar structures. Additionally we have demonstrated that multivariate analysis is a powerful tool to simplify the interpretation of the behaviour of NPs described by multiple variables (i.e. uptake into different cell types) into a 2D map. An important result was to provide by a simple 2D scatterplot a clear representation of the uptake behaviour of

This journal is © The Royal Society of Chemistry 20xx

the different NPs, and to show that anhydrides in general tended to enhance the uptake into all the here studied cells, while amino acids inhibited the cellular uptake. Finally, we provided a predictive map of the potential uptake of NPs according to the known uptake into different cell types and we demonstrated that this tool could be easily applied to generate predictions for new surface modifiers. The prediction of the three new surface modifiers with the highest selectivity for PaCa2 cells (i.e.trimethylacetic anhydride (n°112), isobutyric anhydride (n°121), and hexafluoroglutaric anhydride (n° 122)) outside the domain of F2 model, impose to consider these predictions as less reliable. However, this observation highlighted two important points: i) the applicability domain should always be identified and quantified, to avoid unreliable extrapolations; ii) the domain of the F2 QSAR model can be improved if new data will become available. This draws attention to the need for new data, which are necessary to build robust and predictive models with the largest possible applicability domain.

Concluding, we think that the proposed approaches can serve as examples of how models can be developed and combined to extract as much information as possible from the analysis of predictions, residuals and domains, and that our results will be useful for the future development of new nanoparticles with different cell-selectivity for use in more efficient biomedical applications.

Experimental

Experimental data set

A library of supermagnetic fluorescent nanoparticles sharing an iron oxide core, a dextran coating and surface modified with 146 different small molecules, was generated by Weissleder and colleagues.[1] The full list of 146 surface modifiers was reported by Fourches et al.[3]

The cellular uptake was tested in different human cell types (i.e. primary resting human Macrophages (RestMph), granulocyte macrophage colony stimulating factor-stimulated human macrophages (GMCSF_Mph), U937 human macrophage-like cell line (U937), human pancreatic adenocarcinoma epithelial cells (PaCa2), and human umbilical vein endothelial cells (HUVEC)). Experimental data measured for 109 surface modifications[1] are available online at https://csb.mgh.harvard.edu/information/links, in the section “NP screening data” from the list of topics (i.e. Data from "Weissleder R, Kelly K, Sun EY, Shtatland T, Josephson L. Nat Biotechnol. 2005 Nov;23(11):1418-23.")).

Uptake was quantified by well fluorescein isothiocyanate (FITC) concentrations. Data expressed as picomoles/Liter (pM) were modelled by QSAR and log-transformed (log10) prior to modelling. The 109 surface modifiers tested on the different cell types, were listed in Table S1. Twenty-eight additional surface modifiers were also listed in TableS1 for a total of 137 chemicals. These 28 chemicals were extracted after exclusion of salts from the 37 surface modifiers with unknown uptake reported in the list published by Fourches[3], in addition to the 109 with measured response.[1]

Calculation of the molecular descriptors

Molecular descriptors were calculated for the 137 molecules used to modify the surface of the iron oxide NPs. 3D structures were designed and energetically optimized, using both the Semi-empirical method AM1 and the Allinger molecular mechanical method (MM+), in the HYPERCHEM program.[41] The software PaDEL Descriptors (v2.18)[38] and CODESSA[42]were used to compute mono and bi-dimensional molecular descriptors starting from the optimized structures. Constant, near constant and highly correlated descriptors (R>95%) were excluded, by using QSARINS[18], to reduce redundant and non-useful information. At the end of this procedure, a final set of 612 descriptors was used as input for the modelling.

Explorative analysis

Principal component analysis (PCA)[34] was performed on residuals in prediction calculated by the different models proposed in this study. The 109 NPs were labelled according to the different chemical classes they belonged to i.e. amines, anhydrides and amino acids. PCA was performed on autoscaled data in the software SCAN.[43]

QSAR Modelling and validation

The QSAR approach applied in this study was based on the regression of the structural properties of 109 chemicals, used as surface modifiers, against the cellular uptake measured for 109 NPs generated after conjugation of the surface modifiers with the same supermagnetic nano-core. The basic assumption was that the core shared by all the nanoparticles was a constant element, and as such, it was excluded from the structure-activity analysis to focus on the structural differences responsible for the measured cellular uptake. The approach used here offers the possibility to calculate a large variety of different molecular descriptors and was successfully applied in former studies.[3-10] Moreover, as was explained by Weissleder[1], only PaCa2 and HUVEC were characterized by sufficient variability in the response depending on surface modification suitable for the development of QSAR models. Therefore, we developed specific regression QSAR models for these two cell types by using different linear and non-linear methods.

Due to the large amount of structural descriptors available as input for QSAR generation, the best combinations of descriptors were identified by using variables subset selection methods available for the development of Multiple Linear Regression based on Ordinary Least Squares (MLR-OLS) in the software QSARINS.[18]

The selected descriptors were used as input for the development of other linear and non linear models.

Multiple Linear regression Models for PaCa2 and HUVEC responses

Multiple Linear Regression technique [12-15] attempts to find a linear relationship between a dependent variable (y) and more than one independent variables (x_j). This relationship can be reported as:

$$\hat{\boldsymbol{y}} = \boldsymbol{b_0} + \boldsymbol{b_1 x_1} + \boldsymbol{b_2 x_2} + \cdots + \boldsymbol{b_j x_j}$$

Where $\hat{\boldsymbol{y}}$ is the calculated response (dependent variable), b_j are the coefficients of the models, and x_j are the predictors (independent variables).

The b elements of the vector of the coefficients are estimated, using the ordinary least squares method (OLS), from the X matrix of the independent variables according to the following formula:

$$b=(X^T X)^{-1} X^T y$$

The best modelling variables to generate MLR-OLS models were selected from the initial pool of over 600 molecular descriptors calculated for chemicals in the training set. The selection was performed in two steps starting with an exhaustive search, i.e. by exploration of the statistical quality of MLR-OLS models generated by all the possible combinations of up to two of the available experimental descriptors, followed by Genetic Algorithm[18]. The output of the GA was a population of models including up to eight descriptors. Models were intentionally kept as simple as possible, as recommended by the parsimony principle (Ockham's Razor) and the inclusion of a new variable in the models was stopped when the increase in the models complexity did not increase the models performance. The best models were chosen by using Q^2 leave-one-out (Q^2loo) as optimization value. Furthermore, the correlation between the modelling descriptors and the modelled response was checked by the QUIK rule[44], to exclude models with co-linearity and exclude chance correlation. Additionally, Y-scrambling was applied to verify that the models were not based on a chance correlation of descriptors with the response. Low R^2 values of the models, which were calculated on scrambled responses (i.e. R^2SY), confirmed the absence of chance correlation in the original model (results reported in Table S2). Moreover, the robustness of the models was evaluated by applying the leave many out (15-30%) procedure Q^2lmo (2000 iterations). Standardized residuals were calculated to identify outliers for the response (chemicals with standardized residuals greater than 2.5 standard deviation units).

The external predictivity of models was evaluated on multiple random prediction sets (5 for each training set) manually generated by unbiased random splitting (without taking into consideration response or descriptors distributions), leaving out about 20% of the original data sets.

Different populations of models selected by the genetic algorithm were generated independently for each training set, and were tested on the respective prediction set. The external predictivity of the models was quantified by analysis of different external parameters calculated by QSARINS[18] for each model in the independent populations i.e. three differently calculated External Q^2 and the Concordance Correlation Coefficient.[45,46]

In addition, the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), were used to evaluate the prediction accuracy[18,19]. In particular, the validation criteria proposed by Roy et al.[19] based on MAE and RMSE statistics calculated for the 100% and the 95% of the prediction sets, were used to further confirm the predictivity of the models. These parameters were calculated by the software Xternal Validation Plus (http://teqip.jdvu.ac.in/QSAR_Tools/).[19]

The best models were chosen as best options taking into consideration internal and external predictivity, number of outliers and applicability domain. Once identified the best MLR models for the two responses of uptake, only the external predictions (generated by these models for the five independent prediction sets set used to perform the external validation) were combined in order to have only externally predicted values for all the 109 nanoparticles. These predictions were used to compare the external predictivity of linear and non linear models.

PLS and Non-Linear Regression Models for PaCa2 and HUVEC responses

Different linear and non linear approaches were explored in addition to MLR-OLS, i.e. Partial Least Squares (PLS) regression[20], Projection Pursuit Regression (PPR)[21], support vector machines (SVM)[22-25], K-NearesNeighbours (K-NN)[26], Radial Basis function neural networks (RBFNN)[27-29] and general regression neural networks (GRegNN).[30-32]

Since so different methods may overemphasize some structural characteristics and ignore or underestimate others, is therefore not a priori obvious that they will lead to similar results.[47] The performances and the sensitivity of the different methods were compared in the same modelling conditions. The multiple predictions provided the basis for the application of combinatorial approach to analyse residuals and reduce the prediction errors.[16,17,48]

As mentioned above all the modelling techniques tested in addition to MLR, were non linear machine learning approaches, with the only exception of PLS. Non linear methods do not provide any explicit, directly usable, formula for property evaluation, and therefore they appear as more directed toward activity prediction or data mining than mechanistic interpretations. However, despite these drawbacks, they have the advantage of easy settings, rapid training, and guarantee to find the global minimum on the error surface. Owing to these beneficial aspects, it is not surprising that a large amount of QSPR/QSARs now rely on non-linear approaches, with successful applications also in the field of nanotoxicology. [3-11, 16, 17]

Calculations for the additional methods were generated by using the Caret package [49] of the Cran-R software[50], and Matlab routines for RBFNN[51] and GRegNN[52]. The quality of the models was evaluated by quantification of internal fitting (R^2) and predictivity (Q^2loo).

The external predictivity of the models was evaluated on the same external prediction sets used for MLR models described before, and quantified by R^2. RMSE and MAE criteria[18,19] were used to measure and compare prediction accuracy in the training and in the prediction sets. All these parameters, obtained in comparable conditions, were used for comparing the quality of the predictions using linear or non-linear models.

A brief description of the basic principles of approaches different from MLR is given as follows. More information can be found in specific literature.[20-32]

Tuneable parameters were adjusted through optimization of cross-validated RMSE or Q^2loo, while the other parameters were fixed, as far as possible, at their default value. Specific settings used for the setup of the models generated for the full data sets are given in Table S10.

We used a common methodology relying on a grid-type search to optimize the parameters for the methods implemented in the package caret[49] i.e. PLS, PPR, KNN, SVM. The tuneable parameters for these methods concern the number of components (PLS), the number of successive projections (PPR) or the number of neighbours (KNN). In particular, for a series of possible values (typically from 5 (PLS) to 10 (SVM)), proposed by the program or user-defined, we performed a 5-fold cross-validation repeated 3 times. The best mean RMSE value for these runs determined the choice of the best parameters. The reproducibility of the results was ensured by

This journal is © The Royal Society of Chemistry 20xx

the use of a seed-value (here seed=2), which was used to select subsamples in cross validation.

Partial Least Squares (PLS)[20] generates by linear combination of the original variables a limited set of orthogonal components (latent vectors). Unlike the aforementioned Principal Component approach, which is an explorative technique, the set of latent vectors in PLS, is determined as representing at best the variability in both the descriptor (*X*) and also in the property (*Y*) space.

Projection Pursuit Regression (PPR) is a nonparametric method developed by Friedman and Stueltzle[21] and may be considered as a (empirically determined) sum of nonlinear local smooth (univariate) functions iteratively determined. Given a trial direction vector *a*, the descriptor matrix *X* is projected as:

$$Z=a^T X$$

The model operates in the space of the *Z* projections, which are linear combinations of the initial variables. PPR approximates the regression function (relying the property y to associated predictors X) by a finite sum of smooth ridge functions of the new predictor variables Z. The software, after setting the smoothing function, automatically defines the number of projections, by optimizing cross-validation results (i.e. RMSEcv). In this study, the first projection was sufficient to generate a satisfactory QSARs.

Support Vector Machine (SVM)[22-25] approach privileges robustness of the model over the search for an optimal recall of the data in order to get more generalization ability. A kernel function is used to project data in a higher dimensional space, where it may be expected that a linear representation would work better than in the original descriptor space. The choice of the kernel function (and the related hyper-parameters) is of crucial importance for the optimization of the SVM's performance. In this work, we used Linear and Gaussian kernels that are the most commonly employed in QSAR studies.
The application of SVM with a Linear kernel requires to adjust the regularization parameter C, which balances between the complexity of the model and its precision. In addition, the diameter of the "epsilon insensitive tube" (where the errors are neglected during the model development) was left at default value, i.e. 0.1.
In Radial SVM the Gaussian kernel additionally requires the definition of the "inverse radius" gamma (γ) of the Gaussian, which is computed as $\exp(-\gamma(x_j-x_i)^2)$, x_i and x_j being independent feature vectors. The software caret defines γ value by the kernlab program starting from the input variables (here the molecular descriptors).
A supplementary user-driven optimisation of γ slightly improved the results and was not systematically carried out.
The best settings for Linear and Gaussian kernels, were defined through optimization of RMSEcv values.

Neighbourhood based method such as k-Nearest-Neighbours (k-NN), Radial Basis Function Neural Networks (RBFNN) and General Regression Neural Network (GRegNN) directly rely (at different levels) to neighborhood relationships between samples. [26-32]
In k-NN[26] the property (class membership or activity value) is not calculated by fitting a model, but evaluating a weighted average value over the k_{th} most similar compounds. Here, the best k value was k=1.

In RBFNN and GRegNN [16,17,27-32] the investigated property for a compound is evaluated as a weighted average of its values on selected neighbouring compounds.
For RBFNN the Orr's algorithm[28,51] automatically determines the number and location of the hidden units (that are chosen among the data points). The radius (s) of the Gaussian function defines the activation of hidden units. The Gaussian function is then calculated as $\exp(-(x-ci)/s)^2$ where x and c_i represent the predictors of the investigated pattern and hidden center c_i, (note that a unique radius is chosen for all hidden unit).
A series of user-defined values of "s" were tested to optimize the RMSE calculated in leave-one-out for the training sets.

The same approach was used for GRegNN. Here, the whole dataset is involved in evaluating neighbourhood relationship. However, a unique tuneable parameter i.e. the radius (r), is needed to adjust the Gaussian weighting function (i.e. $\exp[-(x-xj)^T(x-xi)/2r^2]$) intervening in Parzen's estimator[32], which balances the influence of "neighbouring" compounds.
Note that the selected neighbours (i.e. the whole data set (GRegNN) or only some data points (RBFNN) are fixed for a given data set whereas in k-Nearest Neighbour method (k-NN) they vary for each submitted pattern.

Factorial analysis and 2D map of the selective uptake of 137 surface modified Nanoparticles.
Factorial Analysis (FA)[39] was applied to generate a 2D map the selective uptake of the 109 surface modified NPs with known uptake into different human cell types. FA was performed on log transformed values starting from covariance matrix, in order not to overweight uptake into less sensitive cells (i.e. macrophages).[1] The principal components method to extract the factors and the varimax rotation were performed in the software SCAN.[43]
MLR-OLS QSAR models were subsequently generated using as response the coordinates (i.e. score values) of the 109 NPs in the space of the first two rotated factors. Values calculated for the F1 and F2 scores are reported in Table S9. Models development and validation was performed as described in the MLR section. In order to perform the external validation compounds were split according to the scheme "four trainings and 1 prediction", after being ordered according to the respective response value (i.e. Factor 1 (F1) Scores and Factor 2 (F2) scores).
The best split models, were newly calibrated using the empirical information available for the 109 NPs and applied as "full models" to predict the F1 and F2 scores for additional 28 NPs with unknown behaviour in the studied cells.

Applicability Domain
The structural Applicability Domain (AD) of MLR models was quantified by the leverage approach[12,13,18] in order to verify the presence of influential objects (i.e. NPs) in the training set, and to verify the reliability of predictions for objects not included in the training set (i.e. reliable predictions should fall within the AD of a model).
The leverage matrix *H*, which includes n training set samples and p modelling descriptors, is calculated from the X matrix as follows:

$$H=(X(X^TX)^{-1}X^T)$$

Diagonal elements ($h_{i/i}$) of the *H* matrix quantify the influence of each object on the regression results, i.e. the leverage of each object ($h_{i/i}$) in the space of the model. The value h*, which is calculated as 3(p+1)/n (p= number of variables in the model, n=number of compounds in the training set) is the cut-off value for the domain. Compounds which "influence" the mathematical structure of the model have leverage values greater than h* and fall outside structural AD of the model. Predictions calculated for high leverage chemicals in the prediction set should be considered as less reliable (i.e. extrapolated values).[13]

The Applicability domain of MLR models was further inspected by graphic approach. The Williams graph[18,43] is the plot of hat diagonal values vs. standardized residuals and gives an immediate view of NPs falling within the structural AD of the models (i.e. $h_{i/i}$<h*), and of response outliers which are characterized by standardized residuals larger than 2.5 standard deviation units.

Furthermore, the method proposed by Roy et al.[33] which is independent of the MLR statistics, was used to verify the AD for models different than MLR. This approach relies on the range covered by standardized descriptor values. Details regarding this method, which can be applied by the software Applicability Domain (using standardization approach), available online at http://teqip.jdvu.ac.in/QSAR_Tools/, are reported in the related literature.[33]

Acknowledgements

We would like to acknowledge Prof. Paola Gramatica for providing useful comments on this study. We thank Prof.Kunal Roy for providing freeware license for the software Xternal Validation Plus, and Applicability Domain (using standardization approach)http://teqip.jdvu.ac.in/QSAR_Tools/.

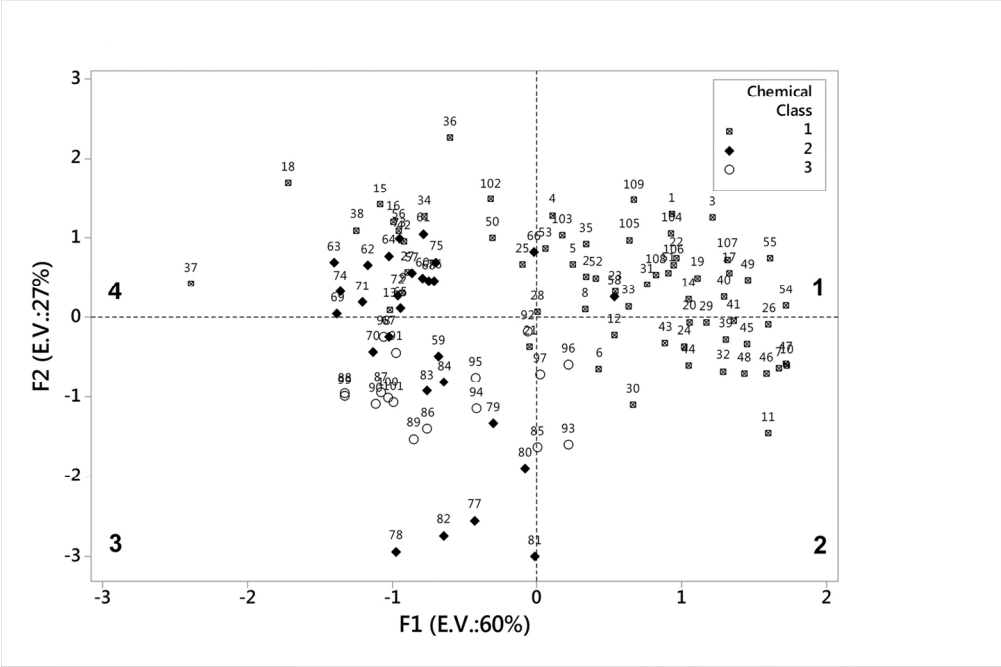
Notes and references

1. R. Weissleder, K. Kelly, E. Y. Sun, T. Shtatland, L. Josephson, *Nat. Biotechnol.*,2005, **23**, 1418-1423.
2. E. Y. Sun, L. Josephson, K. a. Kelly, R. Weissleder, *Bioconjug. Chem.*, 2006, **17**, 109–113.
3. D. Fourches, D. Pu, C. Tassa, R. Weissleder, S. Y. Shaw, R. J. Mumper, A. Tropsha, *ACS Nano.*, 2010, **4**, 5703–5712.
4. V. C. Epa, F. R. Burden, C. Tassa, R. Weissleder, S. Shaw, D. A. Winkler, *Nano Lett.*, 2012, **12**, 5808–5812.
5. M. Ghorbanzadeh, M. H. Fatemi, M. Karimpour, *Ind. Eng. Chem. Res.*, 2012, **51**, 10712–10718.
6. Y. T. Chau, C. W. Yap, *RSC Adv.*, 2012, **2**, 8489-8496.
7. A. Toropov, A. P. Toropova, T. Puzyn, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, *Chemosphere*, 2013, **92**, 31-37.
8. S. Kar, A. Gajewicz, T. Puzyn, K. Roy, *Toxicol. In Vitro*, 2014, **28**, 600-606.
9. K. P. Singh, S. Gupta, *RSC Adv.*, 2014, **4**, 13215 – 13230.
10. D. A. Winkler, F. R. Burden, B. Yan, R. Weissleder, C. Tassa, S. Shaw, V. C. Epa, *SAR QSAR Environ. Res.*, 2014, **25**, 161-172.
11. C. Oksel, C.Y. Ma, J.J. Liu, T. Wilkins, X.Z. Wang, *Particuology*, 2015, **21**, 1–19.
12. Organisation for Economic Co-operation and Development, ENV/JM/MONO(2007)2, Guidance document on the validation of (Quantitative) Structure Activity relationships [(QSAR)] models, http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2007)2&doclanguage=en (Accessed February 2, 2016).
13. P. Gramatica, *QSAR Comb. Sci.*, 2007, **26**, 694–701.
14. .J.P. Doucet, and A. Panaye, *Three-dimensional QSAR. Applications in Pharmacology and Toxicology*,CRC Press, BocaRaton, FL, 2010.
15. C.Y. Liew and C.W.Yap, in *Statistical Modelling of Molecular Descriptors in QSAR/QSPR, Volume 2*, ed. M. Dehmer, K. Varmuza, and D.Bonchev, Wiley-VCH, Verlag GmbH, 2012, 1, 1-31.
16. E. Papa, J.P. Doucet, A. Doucet-Panaye, *SAR QSAR Envir. Res.*, 2015, **26**, 647-665.
17. E. Papa, J.P. Doucet, A. Sangion, and A. Doucet-Panaye, *SAR QSAR Envir. Res.*, 2016, early view doi: 10.1080/1062936X.2016.1197310
18. P. Gramatica, N. Chirico, E. Papa, S. Cassani, S. Kovarich, *J. Comput. Chem.*, 2013, **34**, 2121–2132.
19. K. Roy, R.N. Das , P. Ambure, RB Aher, *Chemom. Intell. Lab. Sys.*, 2016, **152**, 18–33.
20. P. Geladi, B.R. Kowalski, *Anal. Chim. Acta.*, 1986, **185**, 1-17.
21. J.H. Friedman, W. Stuetzle, *J. Am. StatsAssoc.*, 1981, **76**, 817-823.
22. C. Cortes, V. Vapnik, *Machine Learning*, 1995, **20**, 273-297.
23. N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*, Cambridge University Press, 2000.
24. O. Ivanciuc, in *Reviews in Computational Chemistry*. Ed. K.B. Lipkowitz, T.R. Cruciani, Wiley-VCH, Weinheim, 2007, 23, 291-400.
25. J.P. Doucet, F. Barbault, H. Xia, A. Panaye, B.T. Fan, *Curr. Comput.-Aided Drug Des.*, 2007, **3**, 263–289.
26. W. Zheng, A. Tropsha, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 186-194B.
27. B. Walczak, D.L. Massart*, Chemom. Intell. Lab.*, 2000, **50**, 179–198.
28. M.J.L. Orr, *Introduction to Radial Basis Function Networks*, Centre for Cognitive Science, Edinburgh University, Edinburgh, U.K., 1996.
29. X.J. Yao, A. Panaye, J.P. Doucet, R.S. Zhang, H.F. Chen, M.C. Liu, Z.D. Hu, B.T. Fan, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1257–1266.
30. P.D. Mosier, P.C. Jurs, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 1460-1470.
31. D.F. Specht, IEEE Trans. *Neural Netw.*, 1991, **2**, 568–576.
32. E. Parzen, Ann. Math. Stat., 1962**, 3**, 1065–1076.
33. K. Roy, K. S. Kar, and P. Ambure. *Chemom. Intell. Lab. Sys.*,2015, **145**, 22-29.
34. S. Wold, K. Esbensen, P. Geladi, *Chemom. Intell. Lab. Syst*., 1987, **2**, 37-52.
35. C. Wilhelm, C. Billotey, J. Roger, J.N. Pons, J.C. Bacri, F. Gazeau F., Biomaterials, 2003, **24**, 1001-1011.
36. A. Albanese, P.S. Tang; W.C.W. Chan, *Annu. Rev. Biomed. Eng.*, 2012, **14**, 1-16
37. J. Voigt, J. Christensen, V P. Shastri, PNAS, 2014, **111**, 2942–2947
38. C.W. Yap, *J. Comput. Chem.*, 2011, **32**, 1466-1474.
39. A. Gie Yong, S. Pearce, *Tutor. Quant. Methods Psychol.*, 2013, **9**, 79–94.

 This journal is © The Royal Society of Chemistry 20xx

40. M.H. Abraham, A. Ibrahim, A.M. Zissimos, Y.H. Zhao, J. Comer, D.P. Reynolds, *Drug Discovery Today*, 2002, **7**, 1056-1063.
41. HYPERCHEM v. 7.0, Hypercube, inc. http://www.hyper.com/
42. A.R. Katritzky, M. Karelson, R. Petrukhin, COmprehensiveDEscriptors for Structural and Statistical Analysis, www.codessa-pro.com/
43. SCAN Software for Chemometric Analysis, rel. 1.1, Minitab Inc.
44. R. Todeschini, A. Maiocchi, V. Consonni, *Chemom. Intell. Lab.* 1999, **46**, 13–29.
45. N. Chirico, P. *Gramatica, J. Chem. Inf. Model.*, 2011, **51**, 2320–2335.
46. N. Chirico, P. Gramatica, *J. Chem. Inf. Model.*, 2012, **52**, 2044-2058.
47. J.P. Doucet, A. Doucet-Panaye, *SAR QSAR Environ Res.*, 2014, **25**, 589-6
48. E. Papa, L. van der Wal, J. Arnot, P. Gramatica, *Sci. Total Environ.*, 2014, **470-471**, 1040-1046.
49. M. Kuhn, *J. Stat. Soft.*, 2008, **28**, 1-26.
50. R.D.C. Team. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria, 2014, http://www.R-project.org (accessed February 25, 2016).
51. M.J.L. Orr, *MATLAB Routines for Subset Selection and Ridge Regression in Linear Neural Networks*, Centre for Cognitive Science, Edinburgh University, Edinburgh, U.K., 1996.
52. A. Panaye, B.T. Fan, J.P. Doucet, X.J. Yao, R.S. Zhang, M.C. Liu, Z.D. Hu, *SAR QSAR Environ. Res.*, 2006, **17**, 75–91.
53. G.M. Maggiora, *J. Chem. Inf. Model.*, 2006, **46**, 1535.

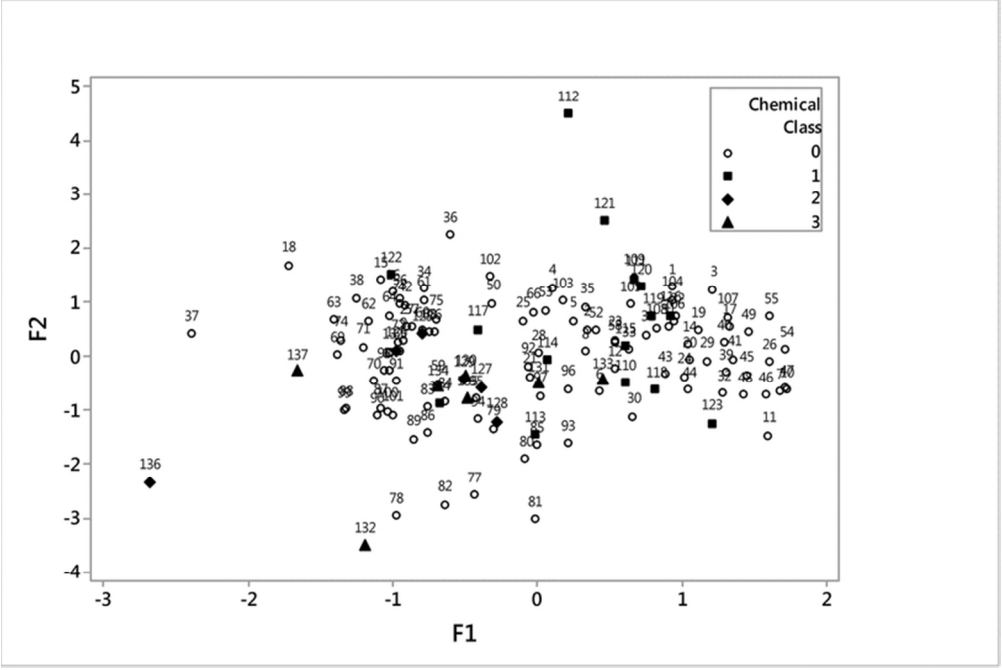
This journal is © The Royal Society of Chemistry 20xx



Rotated score plot of Factor 1 (F1) and Factor 2 (F2) generated by Factorial Analysis on log-transformed data of uptake in different human cell types measured for 109 nanoparticles. Nanoparticles are labelled according to chemical classes of the surface modifiers (i.e. 1=anhydrides, 2=amines, 3=aminoacids).

Figure 1

135x90mm (300 x 300 DPI)



Plot of empirical (i.e. calculated by Factorial Analysis), and predicted (i.e. by QSAR) F1 and F2 values for 109 nanoparticles (NPs) with known cellular uptake, and 28 NPs functionalized by new surface modifications. The 109 NPs are labelled as empty circles; the 28 new NPs are labelled according to chemical classes (i.e. 1=anhydrides, 2=amines, 3=aminoacids).

Figure 2

68x45mm (300 x 300 DPI)

Table 1. Performances of split (averages on 5 external prediction sets) and full MLR-OLS models.

PaCa2	R^2	Q^2	Qlmo	CCCtr	R^2YS	RMSE Tr	MAE Training	R^2ext	Average Q^2ext	CCC ext	RMSE ext (100% data)	MAE ext (100% data)	RMSE ext (95% data)	MAE ext (95% data)	MAE+3*SD ext (95% data)
Av. Split	0.74	0.67	0.64	0.85	0.09	0.21	0.17	0.74	0.72	0.84	0.22	0.17	0.18	0.14	0.46
Full Model	0.74	0.69	0.66	0.85	0.07	0.21	0.1656	-	-	-	-	-	-	-	-
HUVEC	R^2	Q^2	Qlmo	CCCtr	R^2YS	RMSE Tr	MAE Training	R^2ext	Average Q^2ext	CCC ext	RMSE ext (100% data)	MAE ext (100% data)	RMSE ext (95% data)	MAE ext (95% data)	MAE+3*SD ext (95% data)
Av. Split	0.75	0.69	0.67	0.86	0.09	0.30	0.23	0.69	0.69	0.82	0.33	0.27	0.27	0.22	0.67
Full Model	0.75	0.7	0.68	0.86	0.07	0.3	0.2372	-	-	-	-	-	-	-	-

Table 2. Comparison of the predictive performances of linear and non-linear approaches evaluated for the 109 NPs in five external validation sets on the basis of Mean Absolute Errors (MAE) and Root Mean Squared Errors (RMSE) values calculated with the software Xternal Validation Plus (available online at http://dtclab.webs.com/software-tools) . Combined=calculated on averaged predictions.

PaCa2	MLR	SVM-LIN	SVM-RAD	PPR	PLS	KNN	GRegNN	RBFNN	Combined	Prediction Quality
RMSEext(100% data)	0.22	0.23	0.28	0.22	0.22	0.26	0.26	0.22	0.22	GOOD
RMSEext(95% data)	0.19	0.20	0.23	0.19	0.19	0.22	0.20	0.19	0.18	GOOD
MAEext(100% data)	0.17	0.18	0.21	0.18	0.17	0.20	0.19	0.17	0.17	GOOD
MAEext(95% data)	0.15	0.16	0.18	0.16	0.15	0.18	0.17	0.15	0.15	GOOD
MAEext+3*SD (95% data)	0.49	0.49	0.61	0.47	0.49	0.56	0.51	0.48	0.46	GOOD
HUVEC	MLR	SVM-LIN	SVM-RAD	PPR	PLS	KNN	GRegNN	RBFNN	Combined	Prediction Quality
RMSEext(100% data)	0.34	0.34	0.42	0.34	0.34	0.35	0.39	0.35	0.33	GOOD
RMSEext(95% data)	0.28	0.29	0.36	0.28	0.29	0.29	0.32	0.29	0.27	GOOD
MAEext(100% data)	0.27	0.27	0.34	0.26	0.27	0.28	0.30	0.27	0.26	GOOD
MAEext(95% data)	0.23	0.24	0.30	0.23	0.24	0.24	0.26	0.24	0.23	GOOD
MAEext+3*SD (95% data)	0.71	0.72	0.87	0.72	0.71	0.75	0.81	0.72	0.68	GOOD

Table of contents

Modelling and screening the selective uptake of magnetofluorescent nanoparticles into human cells by combining QSAR and multivariate analysis.