Nanoscale

PAPER

View Article Online
View Journal | View Issue

Cite this: *Nanoscale*, 2015, **7**, 9664

Prediction of nanoparticles-cell association based on corona proteins and physicochemical properties

Rong Liu,[a,b] Wen Jiang,[a] Carl D. Walkey,[c] Warren C. W. Chan[c,d,e] and Yoram Cohen*[a,b,f]

Cellular association of nanoparticles (NPs) in biological fluids is affected by proteins adsorbed onto the NP surface, forming a "protein corona", thereby impacting cellular bioactivity. Here we investigate, based on an extensive gold NPs protein corona dataset, the relationships between NP-cell association and protein corona fingerprints (PCFs) as well as NP physicochemical properties. Accordingly, quantitative structure–activity relationships (QSARs) were developed based on both linear and non-linear support vector regression (SVR) models making use of a sequential forward floating selection of descriptors. The SVR model with only 6 serum proteins and zeta potential had higher accuracy (R^2 = 0.895) relative to the linear model (R^2 = 0.850) with 11 PCFs. Considering the initial pool of 148 descriptors, the *APOB*, *A1AT*, *ANT3*, and *PLMN* serum proteins along with NP zeta potential were identified as most significant to correlating NP-cell association. The present study suggests that QSARs exploration of NP-cell association data, considering the role of both NP protein corona and physicochemical properties, can support the planning and interpretation of toxicity studies and guide the design of NPs for biomedical applications.

Received 9th March 2015,
Accepted 21st April 2015

DOI: 10.1039/c5nr01537e

www.rsc.org/nanoscale

1. Introduction

Engineered nanomaterials (ENMs) are now routinely used in a myriad of products and various applications given their novel/useful chemical, electrical, magnetic, optical, thermal, and mechanical properties that arise from their small size and structural features.[1] It is estimated that ENMs are constituents of over 1600 consumer products.[2] Aside from their use in various industrial and consumer products, there are also rapid advancements in the application of nanotechnology in medical-treatment and diagnosis as exemplified by the use of nano-scaled systems for safe and effective delivery of therapeutic agents into targeted sites.[3] Although many of the unique properties of ENMs are beneficial, there is concern regarding potential risks that ENMs may pose to human health and the environment.[4,5] Various studies have reported that certain ENMs can lead to adverse biological impacts.[6,7] For example, CuO and ZnO nanoparticles (NPs) have been observed to cause pulmonary inflammation,[8] Ag and Pt NPs may interfere with zebrafish embryo hatching,[9] CdSe quantum dots were found to affect cell viability,[10] and carbon nanotubes have been linked to pulmonary fibrosis.[11]

Experimental[5,8,12] and modeling studies[13,14] have shown that bioactivity of ENMs is strongly correlated with their physicochemical properties. Based on the assumption that ENMs of similar physicochemical properties will have similar bioactivity, a number of (quantitative) structure–activity-relationships ((Q)SARs) have been successfully developed for various ENMs, such as metal oxide NPs,[14–17] surface modified iron oxide NPs,[18–22] and carbon nanotubes.[23] Physicochemical properties used as (Q)SAR descriptors include NP primary[15–18] and aggregate[15] sizes, zeta potential in the media of exposure,[15,17,18] concentration (*e.g.*, mass concentration[15] and volume fraction[16]), relaxivities,[18,19] energy/enthalpy information (*e.g.*, atomization and band gap energies[16,17] and formation enthalpies[14,17]), and structures of NP surface-modifying molecules.[18,20–22,24,25] In a physiological environment, NPs suspended in a biological fluid (*e.g.*, blood, plasma, or interstitial fluid) will adsorb proteins that form a "protein corona" on the NP outer surface.[26–30] It has been suggested that the protein corona constitutes the first stage of nano-bio interaction/interface that governs subsequent NP fate and transport (*e.g.*, aggregation, dissolution, and mobility) which in turn impacts NP bioactivity (*e.g.*, cellular association/uptake

[a] Center for Environmental Implications of Nanotechnology, University of California, Los Angeles, CA90095, USA. E-mail: yoram@ucla.edu
[b] Institute of the Environment and Sustainability, University of California, Los Angeles, CA 90095, USA
[c] Institute of Biomaterials and Biomedical Engineering, University of Toronto, Canada M5S 3G9
[d] Department of Chemistry, Department of Chemical Engineering, Department of Materials Science and Engineering, University of Toronto, Canada M5S 3G9
[e] Donnelly Center for Cellular and Biomolecular Research, University of Toronto, Canada, M5S 3G9
[f] Chemical and Biomolecular Engineering Department, University of California, Los Angeles, CA 90034, USA

and toxicity).[29] In this situation, the NPs assume "biological identity" after exposure to biological environment that is different than their "synthetic identity".[26,27,31] The biological identity is "seen" by cells and impacts the observed biological behavior.[26–29] Therefore, the protein corona, as described by the types of serum proteins and their abundance on the NP surface,[30] are thought to encode information that may be useful in predicting NP bioactivity.[27,31]

It has been argued that cellular association with NPs, attributed to the protein corona, is a critical upstream event influencing various downstream biological responses, such as pharmacological and toxicological effects.[29,30] For example, in a recent study Au NPs were incubated with human serum, purified, and the proteins were identified by mass spectrometry.[27] In parallel, cell association, using A549 human lung epithelial carcinoma cells, was quantified for a compositionally diverse library of 105 NPs (Fig. 1) of 15, 30, or 60 nm Au core with neutral, anionic, or cationic ligands.[27] Liquid chromatography tandem mass spectrometry was employed to detect the adsorbed serum proteins on the NP surface, providing a comprehensive quantitative characterization of the NP protein corona.[27] For the entire library of Au NPs, a total of 785 distinct serum proteins were detected with 129 identified as suitable for abundance quantification. The relative abundance[27] of the 129 serum proteins adsorbed on the NP surface was used as a "fingerprint" to characterize the protein coronas. QSARs were then developed for 84 NPs based on the 129 protein fingerprints, with the remaining 21 NPs of neutral surface ligands excluded due to their negligible adsorption of serum proteins.[27] QSAR for $\log_2$-transformed cell association (2.60×10^{-3} to 2.51 mL μg(Mg)$^{-1}$) of Au NPs was developed *via* partial least squares regression (PLSR)[32] using 6 optimal principal components (PCs) calculated from 64 fingerprints identified *via* sequential forward selection (SFS).[33] The developed QSAR demonstrated prediction accuracy in terms of R^2 = 0.81 and 0.61 (coefficient of determination between predicted and observed cell association) in leave-one-out (LOO) validation (*i.e.*, R^2_{LOO} = 0.81) and 4-fold cross-validation (*i.e.*, R^2_{4CV} = 0.61), respectively.

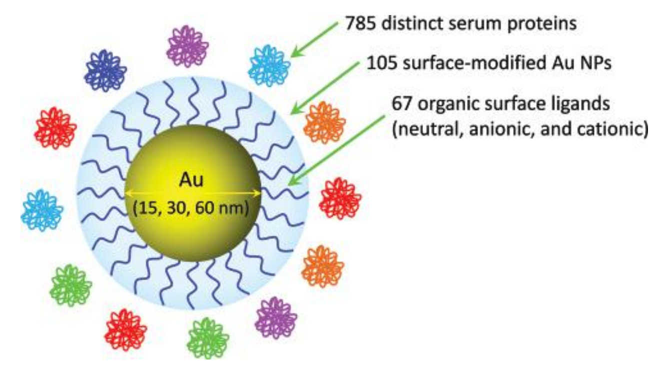
785 distinct serum proteins
105 surface-modified Au NPs
67 organic surface ligands (neutral, anionic, and cationic)
Au
(15, 30, 60 nm)

Fig. 1 Structure of Au NP-protein complex. The NP library contains 105 NP formulations of a 5, 30, or 60 nm Au core with 67 neutral, anionic, or cationic ligands. For the entire NP library, a total of 785 distinct serum proteins were detected on their surfaces with 129 identified as suitable for relative abundance as "fingerprints" for protein corona characterization.

The above study also reported that QSARs for Au NP-cell association, developed based on a total of 39 physicochemical properties of the Au NPs (as synthesized and w/serum) of various surface ligands,[27] were not of an acceptable prediction accuracy. The 39 physicochemical properties included TEM and DLS size characterization, zeta potential, absorbance spectrophotometry, and the amount of adsorbed serum protein on the NP surface obtained from the bicinchoninic acid (BCA) assay. The best performing QSAR developed in the above work[27] was using 7 PCs calculated from 52 descriptors that included both protein corona fingerprints and NP physicochemical properties; however, only marginal performance improvement was attained (R^2_{LOO} = 0.86 and R^2_{4CV} = 0.63) relative to the QSAR utilizing only the protein corona fingerprints. Given the above, it was concluded that the protein corona encodes relevant biological information regarding cell association with Au NPs.[27]

The above work provided an important basis for advancing the understanding of nano-bio interactions given the publication of comprehensive fingerprint profiles of protein corona on the surface of Au NPs, along with QSAR analyses of their correlation to cell association.[27] Although definitive correlations of protein corona fingerprints with cell association were demonstrated, the QSAR descriptors were selected based on SFS process that provided limited exploration of the descriptor space.[33,34] Also, LOO validation of the derived QSARs, while valuable, usually leads to over-optimistic estimate of prediction accuracy for small datasets, while cross-validation (CV) can lead to estimate of excessive variance.[35,36] It is also important to note that the QSARs reported in the above work[27] were developed using PLSR that are in fact "full models", in which each PC represents a linear combination of all the original descriptors (*e.g.*, the 64 selected fingerprints).[37,38] The above provided practical compact QSARs the use of a small number of PCs, but at the cost of forfeiting unambiguous and direct links between cell association and the specific protein corona fingerprints.[38,39] The analysis presented in the above work provided a sufficient and compelling case regarding the importance of the protein corona fingerprints and the lesser dominance of physicochemical properties.[27]

With the above dataset it should be possible to identify the specific and quantitative significance of various corona proteins and physicochemical properties to NP-cell association. Accordingly, in the current work we demonstrate that the specific corona serum proteins and/or NP physicochemical properties, which are most relevant to correlating NP-cell association, can be identified *via* rigorous QSAR analysis. QSAR descriptors were identified by sequential forward floating selection (SFFS)[34] with prediction accuracy of developed linear and non-linear models validated by a bootstrapping based approach that is suitable for relatively small datasets.

2. Results

2.1. QSAR development

The correlation of cell association of Au NPs (modified with different ionic/cationic surface ligands) with corona proteins and physicochemical properties was investigated *via* QSAR analysis of a recently published dataset.[27] QSARs were developed following a workflow (Fig. 2; Methods section) consistent with the OECD guidelines.[40] In the dataset[27] of cellular association of 84 Au NPs, 9 of the 129 identified protein corona fingerprints (PCFs) were found in the coronas of <5% of the 84 Au NPs. Thus, these 9 PCFs did not provide sufficient information for developing QSARs of acceptable generalization capability and were hence omitted. The Au NPs were characterized by a total of 19 physicochemical properties (NPPs; Table 1) which included TEM core size, four hydrodynamic size measures (based on *z*-average, volume mean, number mean and intensity mean), primary NP surface area and volume, zeta potential, and localized surface plasmin resonance (LSPR) index and peak position. NPs in serum were also characterized by the same hydrodynamic size measures, LSPR index, and zeta potential as above, in addition to the adsorbed proteins (total mass and adsorbed surface density) and the NP dose (Table 1). Accordingly, three different initial descriptor sets were utilized, which comprised of 120 PCFs, 19 NPPs (Table 1), and a composite of the two sets. Both linear regression and non-linear epsilon support vector regression (ε-SVR) models[41–43] were used for QSAR development. The most suitable descriptors for a given QSAR were selected from the above three sets of descriptors *via* sequential forward floating selection (SFFS).[34] QSAR prediction accuracy was assessed *via* a bootstrapping (*i.e.*, sampling with replacement) based validation approach that has proven particularly suitable for a limited number of training samples.[17,35,36] Robustness validation of the developed QSARs was carried out based on Y-randomization[17–19] to ensure that the QSARs were not "chance" correlations. Applicability domain analysis using William's plot[44,45] was subsequently conducted to map the descriptor space for which reliable QSAR predictions can be attained.

2.2. Impact of number of selected descriptors on QSAR prediction accuracy

Correlation of cell association with protein corona fingerprints (PCFs) and NP physicochemical properties (NPPs) was assessed *via* linear and non-linear ε-SVR QSARs. The developed QSARs demonstrated prediction accuracy in terms of R^2_{E632}(= 0.368R^2_{resub} + 0.632R^2_{boot}, where, R^2_{resub} is the model prediction accuracy assessed using the training set and R^2_{boot} denotes the prediction accuracy in bootstrapping validation;[35,36] Methods section) that increased with increasing number of selected descriptors (Fig. 3). The prediction accuracy of the ε-SVR QSARs was higher relative to linear QSARs (Fig. 3), at the cost of increased model complexity. For the linear QSARs, the inclusion of NPPs with PCFs did not result in significant difference in prediction accuracy compared to those developed based on PCFs alone (Fig. 3); this implies that information contained in the NPPs in a linear correlation to cell association of Au NPs could be substituted by certain PCFs. On the other hand, of the ε-SVR QSARs developed based on descriptors selected solely from the NPPs, the highest prediction accuracy of R^2_{E632} = 0.695 was attained with 3 NPPs, including zeta potential (as synthesized), density of protein on NP surface, and number mean hydrodynamic diameter (w/serum). The above result should, however, not be interpreted to suggest that NPPs have no relevance to cell association, since the ε-SVR QSAR demonstrated significantly improved prediction accuracy (R^2_{E632} = 0.895, Fig. 3) upon the inclusion of zeta potential (as synthesized) with the selected PCFs.

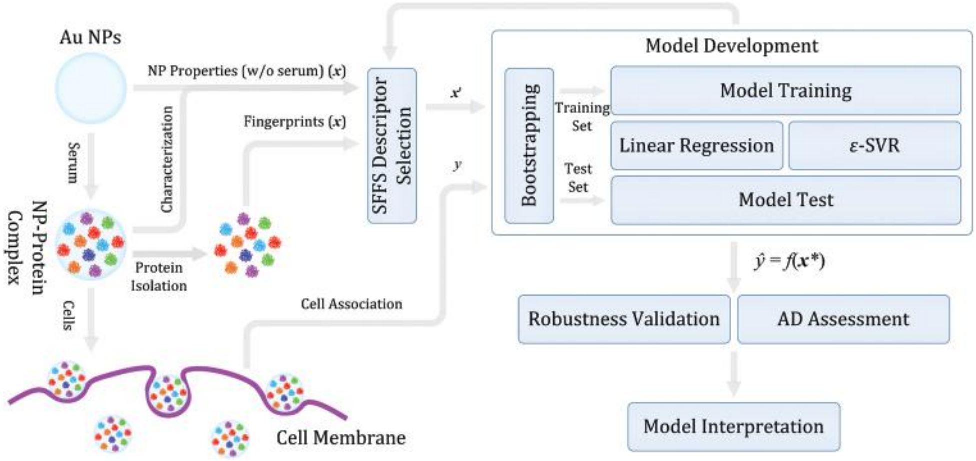

Fig. 2 Workflow for QSAR development. Both linear and nonlinear (ε-SVR) regressions were used in the QSAR development for cellular association of Au NPs. The regressions were coupled with sequential forward floating selection (SFFS) to identify suitable QSAR descriptors from the three descriptor sets, including 120 PCFs, the 19 NPPs, and a composite of the two sets. The QSAR prediction accuracy was assessed *via* a bootstrapping based validation approach known as 0.632 estimator. Once the desired QSAR was identified, Y-randomization was then conducted to assess its robustness followed by William's plot for applicability domain analysis.

Table 1 Protein corona fingerprints and NP physicochemical properties considered in the QSAR analysis

NP physicochemical property (NPP)

As synthesized	w/serum
Four different NP hydrodynamic diameter (nm) (measures based: *Z*-average, volume mean, number mean, and intensity mean)	Four different NP hydrodynamic diameter (nm) (measures based: *Z*-average, volume mean, number mean, and intensity mean)
Zeta potential (mV)	Zeta potential (mV)
Localized surface plasmon resonance (LSPR) index (AU)	Localized surface plasmon resonance (LSPR) index (AU)
LSPR peak position (nm)	Total adsorbed protein[a] (BCA assay) (µg)
TEM core size (nm)	Total NP surface area[a] (cm^2)
Surface area of a NP (nm^2)	Density of protein on NP surface[a] (µg cm^{-2})
Single NP Volume (nm^3)	

Protein corona fingerprint (PCF)

Abbrev.	Full name	Abbrev.	Full name
A1AT	Alpha-1-antitrypsin	*AMBP*	Protein AMBP
ANT3	Antithrombin-III	*APOB*	Apolipoprotein B-100
APOE	Apolipoprotein E	*APOF*	Apolipoprotein F
CO3	Complement C3	*FA10*	Coagulation factor X
FA11	Coagulation factor XI	*FA12*	Coagulation factor XII
HRG	Histidine-rich glycoprotein	*IC1*	Plasma protease C1 inhibitor
IGHG4	Ig gamma-4 chain C region	*IGLL5*	Immunoglobulin lambda-like polypeptide 5
ITIH3	Inter-alpha-trypsin inhibitor heavy chain H3	*ITIH4*	Inter-alpha-trypsin inhibitor heavy chain H4
KLKB1	Plasma kallikrein	*KNG1*	Kininogen-1
PLMN	Plasminogen	*PON1*	Serum paraoxonase/arylesterase 1
PROS	Vitamin K-dependent protein S	*TETN*	Tetranectin
THRB	Prothrombin	*TTHY*	Transthyretin

[a] Based on the total NP content in the serum.

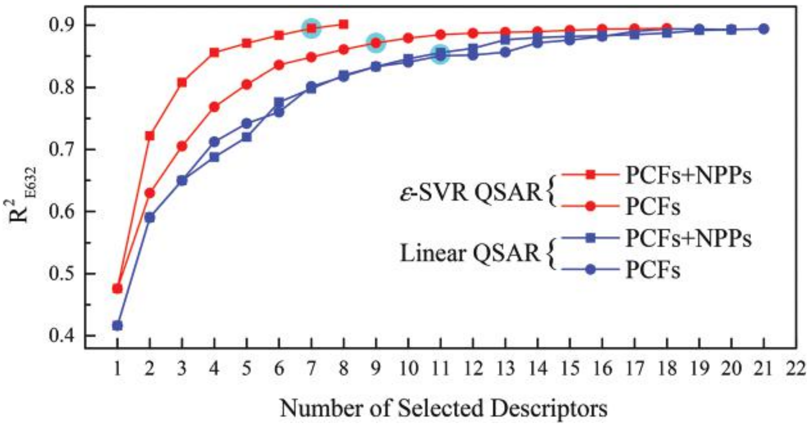

Fig. 3 Prediction accuracy (R^2_{E632}) of the linear and ε-SVR QSARs as increasing number of selected descriptor. The circles identify the "turning-points" where the increase in R^2_{E632} upon adding a new descriptor is ≲1%. In the above figure, PCFs + NPPs identifies the composite descriptor set of the protein corona fingerprints (PCFs) and NP physicochemical properties (NPPs).

It is noted that prediction accuracy increased with added descriptors (Fig. 3). In order to keep the number of QSAR descriptors at a reasonable level, while maintaining acceptable accuracy, the number of selected descriptors was set at the point where the increase in R^2_{E632} upon adding a new descriptor was ≲1%. The above "turning-point"[46] was in the range of 7–11 descriptors (Table 2) for the linear and ε-SVR QSARs. The above number of QSAR descriptors is within the suggested 1/10 to 2/10 ratio of number of descriptors relative to the number of total samples (*i.e.*, NPs).[47] It is noted that 18 of the 24 distinct PCFs identified in the present analysis (Table 2) were also identified previously as suitable QSAR descriptors;[27] moreover, 9 of these 18 PCFs were also identified in previous work with the same dataset[27] as being highly relevant to correlating NP-cell association.

2.3. Non-linear QSAR

Among the QSARs corresponding to the four identified descriptor set (Table 2), the ε-SVR QSAR developed using the

Table 2 Most suitable descriptors selected for linear and ε-SVR QSARs[a]

Linear QSAR	PCF	APOB	ANT3	KLKB1	TTHY	A1AT	IGHG4	ITIH3	PLMN	APOF	FA11	KNG1
	VIP[b]	1.19	1.15	0.98	1.51	1.61	1.00	1.51	—	—	0.85	0.87
	+NPP	APOB	ANT3	KLKB1	TTHY	AMBP	ITIH4	PON1	HRG	VOL_{Au}	IC1	FA10
	VIP	1.19	1.15	0.98	1.51	1.67	1.51	—	0.94	—	—	0.70
ε-SVR	PCF	APOB	A1AT	IGLL5	ANT3	KLKB1	THRB	CO3	PROS	TETN		
	VIP	1.19	1.61	0.80	1.15	0.98	—	1.37	0.67	0.78		
	+NPP	APOB	A1AT	IGLL5	ZP_{Syn}	HRG	FA12	APOE				
	VIP	1.19	1.61	0.80	—	0.94	—	0.87				

[a] A total of 24 distinct PCFs and 2 NPPs (zeta potential (as synthesized, ZP_{Syn}) and volume of a NP (VOL_{Au})) were identified. [b] VIP: variable importance (influence) on projection[32] reported in the previous work[27] for the PCFs. Lack of the VIP value (identifies by "—") indicate that the PCF was not identified in the partial least squares regression QSAR developed in the previous work.[27]

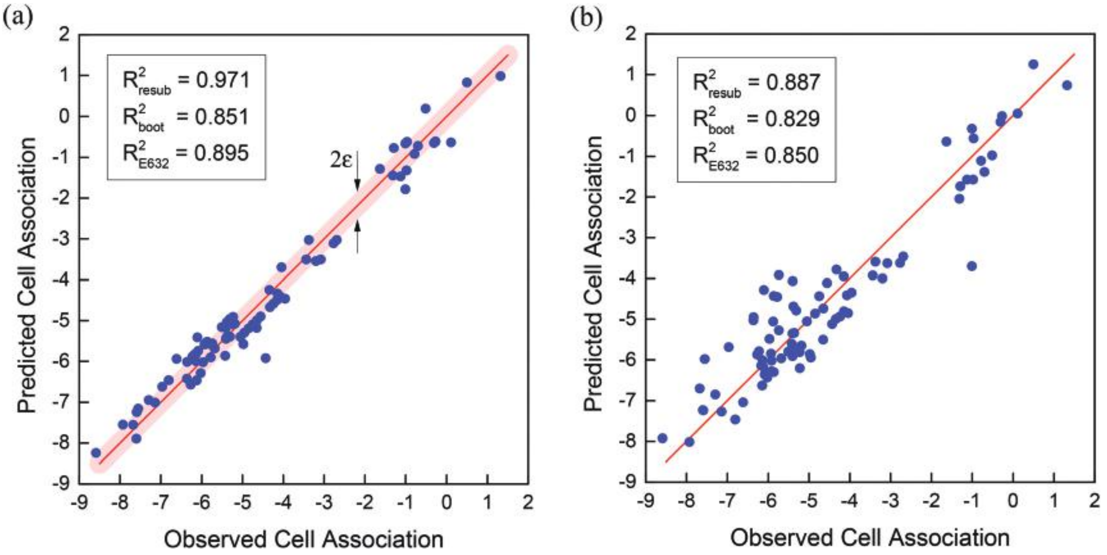

Fig. 4 Observed cell association (2.60 × 10^{-3} to 2.51 mL μg(Mg)$^{-1}$, $\log_2$-transformed) of Au NP *versus* those predicted by (a) the ε-SVR and (b) linear QSARs. In (a), the points on or outside the ε-tube (ε = 0.344) are support vectors of the ε-SVR.

7 most suitable descriptors had the highest prediction accuracy of R^2_{E632} = 0.895 (R^2_{Boot} = 0.851 ± 0.049 and R^2_{resub} = 0.971; Fig. 4a). The above best performing ε-SVR QSAR, which kernel width (γ), regularization factor (C), and tube size (ε) were determined (based on the recommended parameter selection approaches;[42,48,49] Methods section) as 0.125, 11.310, and 0.344, respectively, can be expressed as

$$y = 4.479 + \sum_{i=1}^{51} \alpha_i \exp(-0.125 \parallel \mathbf{z} - \mathbf{z}_i \parallel^2) \qquad (1)$$

in which y represents the $\log_2$-transformed NP cell association (mL μg(Mg)$^{-1}$), while z_i denote NPs (represented by their standardized descriptors) identified as support vectors with their weights given by a_i (Fig. 5). In total, 54 Au NPs (64%) were identified as support vectors (Fig. 5) for the ε-SVR QSAR, which signifies its reasonable sparsity. It is noted that it is acceptable for the number of support vectors in a reasonable ε-SVR model to amount to ~50% of the total samples.[48] The reasonable sparsity together with the relatively small descriptor number demonstrates acceptable complexity of the ε-SVR QSAR. The ε-SVR QSAR (eqn (1)) also exhibited good robustness in a 100-round Y-randomization with R^2_{E632} = −0.208 ± 0.109 and is associated with a well spanned applicability domain (Methods section) covering all but two (G15.DTNB and G60.SPP) of the Au NPs (Fig. 6a). It is also noted that significantly improved prediction accuracy of R^2_{4CV} = 0.862 ± 0.026 as determined in a 100-round 4-fold cross-validation (Methods section) was obtained with the above ε-SVR QSAR (eqn (1)) relative to previously developed partial least squares regression QSAR (R^2_{4CV} = 0.63 ± 0.16).[27]

2.4. Linear QSAR

The developed linear QSARs did not reveal marked difference in prediction accuracy of those constructed from the descriptor sets of PCFs *versus* PCFs plus NPPs. Therefore, based on the linear QSARs, one cannot convincingly argue that NPPs are descriptors of relevance to linear correlation of NP-cell association. Thus, only the PCF descriptors can be justified for the

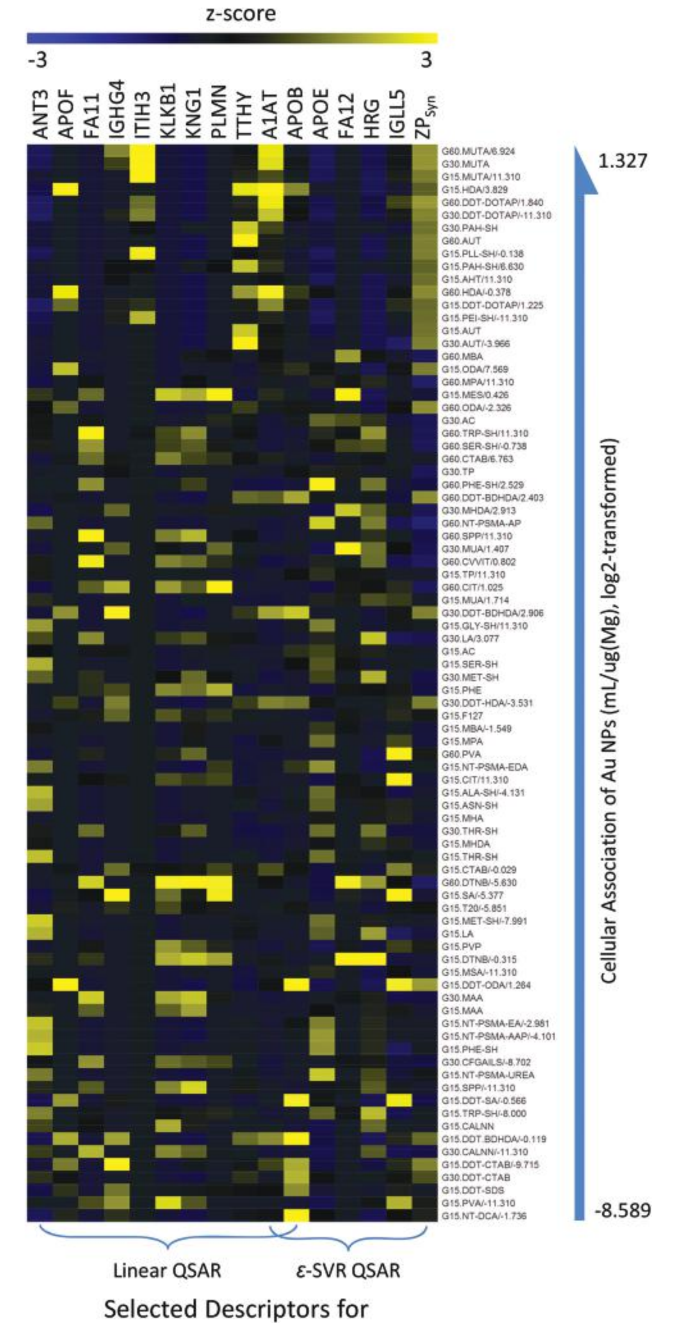

Fig. 5 Selected descriptors (standardized per the *z*-score) for the linear and the ε-SVR QSARs where by the Au NPs are ordered on the basis of their cellular association. A brief explanation about these descriptors is provided in Table 1. The values given following "/" are the weight factors of the NPs identified as support vectors.

linear QSAR. Accordingly, the best performing linear QSAR with PCF descriptors was with 11 descriptors as given by

$$\begin{aligned} y = & -19.167x_{APOB} - 7.188x_{ANT3} + 251.252x_{PLMN} + 51.815x_{ITIH3} \\ & + 31.432x_{A1AT} - 598.507x_{IGHG4} - 106.242x_{KLKB1} + 66.079x_{TTHY} \\ & + 21.937x_{FA11} + 68.110x_{APOF} - 12.531x_{KNG1} - 3.530 \end{aligned} \quad (2)$$

in which y is the log$_2$-transformed NP cell association (mL μg (Mg)$^{-1}$) and x_i is the fingerprint (*i.e.*, relative abundance) of protein i (Table 1). The above linear QSAR provided good correlation of the observed cell association data for the Au NPs (Fig. 4b), quantified by prediction accuracy of R^2_{E632} = 0.850 (R^2_{Boot} = 0.829 ± 0.044 and R^2_{resub} = 0.887). In addition, the above linear QSAR (eqn (2)) also had higher prediction accuracy of R^2_{4CV} = 0.843 ± 0.015 than the partial least squares regression QSAR (R^2_{4CV} = 0.63 ± 0.16) developed previously.[27] Moreover, the present linear QSAR (eqn (2)) exhibited good robustness in a 100-round Y-randomization with average R^2_{E632} = −0.208 ± 0.109; this negative value signifies that the developed QSAR is not a "chance" correlation. The linear QSAR was shown to have a well spanned application domain (Fig. 6b), which covers all but three (G15.DDT-ODA, G15.MES, and G15.AHT) of the 84 Au NPs.

3. Discussion

The developed QSARs provide a means of assessing the relative significance of the identified protein corona and physicochemical descriptors. This was achieved by determining the R^2_{E632} decrease for a given descriptor in 100-round descriptor randomization (Fig. 7; Methods section). We note that one should also consider a descriptor that is selected by multiple models to be of higher significance than those selected by a single model. Among the 7 descriptors used in the best performing ε-SVR QSAR (eqn (1)), *APOB* (Apolipoprotein B-100) was the most commonly selected descriptor (Table 2). This descriptor also demonstrated the greatest impact on the correlation of Au NP cell association. As a major protein found in both low-density lipoprotein (LDL) and very low-density lipoprotein (VLDL), *APOB* has been reported to be responsible for cellular uptake of LDL particles from plasma.[50] Studies have also shown that *APOB* acts as a ligand for LDL receptors in various cells throughout the body as well as a bridge to deliver cholesterol into tissue.[50,51] *APOB* has been identified in protein corona of other NPs (*e.g.*, silica[52] and polystyrene[53] NPs). On the NPs, the protein function is altered due to changes in the structure and this affects the intracellular trafficking, fate, and transport of NPs in cells and animals.[54] The second significant descriptor in the ε-SVR QSAR (eqn (1)) is *A1AT* (Alpha-1-antitrypsin). It is noted that there is evidence that this protein is associated with a broad group of proteases and protects the lungs from cellular inflammatory enzymes.[55] The anti-apoptotic function of *A1AT* has been reported both *in vitro* and *in vivo* for lung microvascular endothelial cells and epithelial cells.[56] In other words, *A1AT* is an anti-inflammation protein serving as an immune system regulator of NPs that impact lymphocyte proliferation and cytotoxicity and mediates monocyte and neutrophil functions.[55] In addition, the association of *A1AT* with anti-inflammation could reduce the attraction of macrophages to the site of NP deposition.[57,58] The third ranking descriptor in the ε-SVR QSAR (eqn (1)) is the zeta potential (as synthesized). This descriptor has been reported

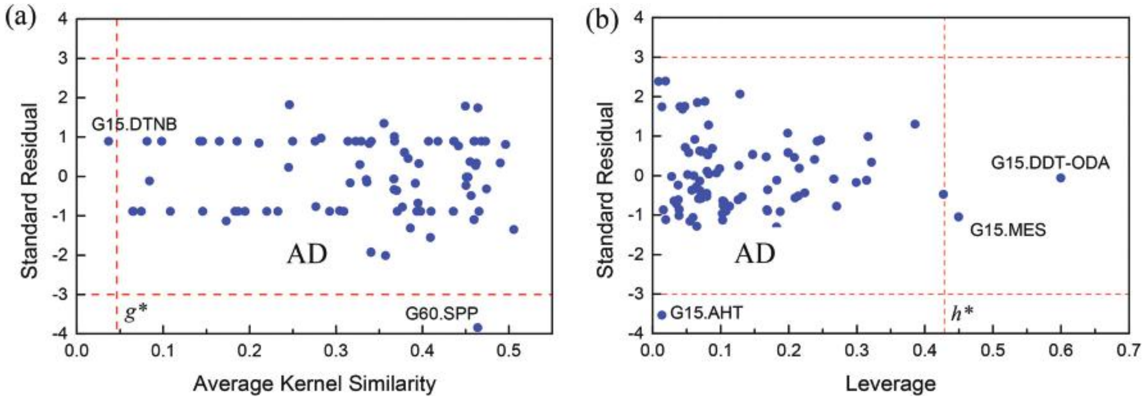

Fig. 6 William's plots for the applicability domains (AD) of (a) the ε-SVR and (b) linear QSARs. The leverage/average kernel similarity quantifies the similarity in NPs while the standardized residual reflects the prediction quality. The critical leverage was identified as h^* = 0.43 and g^* = 0.05 for the linear and ε-SVR QSARs, respectively. In the above figure, G15.DTNB, G15.MES, G15.AHT, G15.DDT-ODA, and G60.SPP identify the NPs of 15 nm Au Core with 5,5′-dithiobis(2-nitrobenzoic acid), 2-mercaptoethanesulfonate, 6-amino-1-hexanethiol, 1-dodecanethiol @ octadecylamine surface modifiers, respectively and NP of 60 nm Au Core with Bis(p-sulfonatophenyl) phenylphosphine surface modifier.[27]

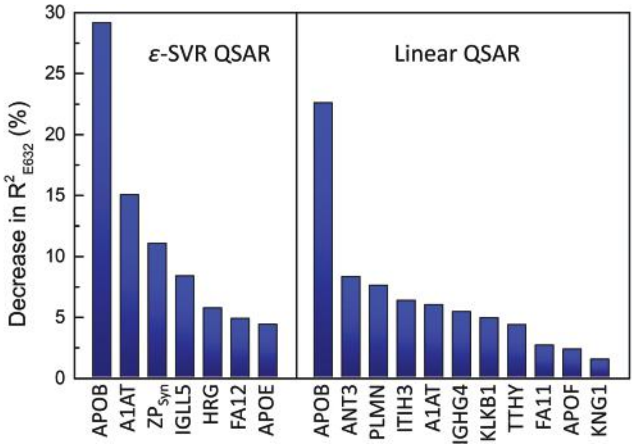

Fig. 7 Descriptor importance assessed by the decrease in prediction accuracy (R^2_{E632}) of the ε-SVR and linear QSARs developed with a given descriptor randomly permutated. A descriptor whose random permutation leads to a large R^2_{E632} decrease is considered to be of increased correlation to cellular association of Au NPs.

as a significant parameter in nanotoxicology and has been widely used as a QSAR descriptor.[16–19,24] NP surface charge, while "screened" by the protein corona, will impact the adsorption of proteins and thus their relative abundance in the protein corona, which will in turn affect NP-cell association.[53] In general, zeta potential plays a fundamental role in the fate and transport of NPs[59] (*e.g.*, stability and aggregation of NPs in aquatic environments[60]) and also affects their behavior at the bio-nano interface.[60,61]

Analysis of the relative significance of the 11 PCFs used in the linear QSAR (eqn (2)), consistent with the ε-SVR QSAR (eqn (1)), also identified *APOB* as having the greatest correlating importance for NP-cell association. The above consistency increases the level of confidence regarding the significance of *APOB* as an important protein that affects cell association of Au NPs. The second significant descriptor in the linear QSAR (eqn (2)) is *ANT3* (Antithrombin III). As a serine protease inhibitor,[62] *ANT3* can exert anti-inflammatory properties *via* inhibition of NF-κB activation and subsequent production of growth factors and cytokine.[63] Therefore, *ANT3* adsorbed onto NP surfaces could alter inflammatory processes during cellular uptake of NPs.[64] The third ranking descriptor in the linear QSAR (eqn (2)) is *PLMN* (Plasminogen), which is known as a zymogen released by plasmin from the liver. The activated or open form of *PLMN* is reported to be responsible for facilitating protein–protein interactions between plasmin and fibrin in blood.[65] Existence of *PLMN* on cell surfaces is important for positive regulation of cell surface plasmin proteolytic activity that facilitates both physiological and pathological processes.[66] Accordingly, the activation of *PLMN* to plasmin could be significantly enhanced when *PLMN* on the surface of NPs binds to cells, indicating the critical role of *PLMN* in macrophage recruitment during the inflammatory response.[67] It is noted that *A1AT*, which was identified as the second most significant descriptor in the ε-SVR QSAR (eqn (1)), was the fifth highest ranking descriptor in the linear QSAR (eqn (2)). The higher ranking of *A1AT* in the ε-SVR QSAR (eqn (1)) could indicate a significant non-linear correlation between *A1AT* and Au NP cell association, which is not captured by the linear QSAR. On the other hand, the simple linear QSAR (eqn (2)) is beneficial in that its descriptors can be categorized as being either positively (*i.e.*, "promoter") or negatively (*i.e.*, "inhibitor") correlated with NP-cell association. For example, the linear QSAR (eqn (2)) indicates that *APOB* is an "inhibitor" while *A1AT* is a "promoter" of Au NP cell association.

4. Conclusions

In summary, the correlation of NP-cell association with protein corona fingerprints (PCFs) and NP physicochemical properties

(NPPs) was explored using both linear and non-linear quantitative structure–activity relationships (QSARs). The analysis of a cell association dataset of a combinatorial library of 84 gold nanoparticles (NPs) of 15, 30, or 60 nm cores with cationic or anionic surface ligands, included evaluation of a set of 129 PCFs and 19 NPPs as QSAR descriptors. Serum proteins, such as *APOB*, *A1AT*, *ANT3*, and *PLMN*, were identified, along with NP zeta potential, as being significant PCFs for correlating NP cell association. The best performing linear QSAR with the most suitable 11 PCFs had a high prediction accuracy of R^2_{E632} = 0.850, which was improved to R^2_{E632} = 0.895 by a non-linear ε-SVR QSAR using 6 only PCFs and NP zeta potential. Both the QSARs demonstrated good robustness and well spanned applicability domain. Good performance of the developed QSARs demonstrated that exploration of the descriptor space can provide important information about NP-cell association that can potentially guide toxicity studies by identifying the set of proteins of potential relevance. Identification of the relevant proteins *via* data mining through QSAR can also provide information that could be useful in developing NP with targeted NP protein adsorption for bio-medical application and to render NPs non-toxic.

5. Methods

QSAR development (Fig. 2) included the use of both linear and non-linear ε-SVR regression models.[41–43] The suitable descriptor sets for a given QSAR were selected *via* sequential forward floating selection (SFFS).[34] QSAR prediction accuracy was assessed *via* a bootstrapping (*i.e.*, sampling with replacement) based validation approach that has proven particularly suitable for a limited number of training samples.[17,35,36] Robustness validation of the developed QSARs was carried out based on Y-randomization[17–19] to ensure that the QSARs were not "chance" correlations. Applicability domain analysis using William's plot[44,45] was subsequently conducted to map the descriptor space for which reliable QSAR predictions can be attained.

5.1. QSAR models

Linear QSARs, for the $\log_2$ transformed cell association (y), can be expressed as

$$y(\boldsymbol{x}) = b + (\boldsymbol{w}, \boldsymbol{x})$$

where ($\boldsymbol{w}$, $\boldsymbol{x}$) denotes the inner product of the NP descriptor vector $\boldsymbol{x}$ (*e.g.*, PCFs or NPPs) and weight vector $\boldsymbol{w}$. The weight vector $\boldsymbol{w}$ and the intercept term b are parameters determined from the data by the least square algorithm.[41]

Non-linear QSARs, developed using ε-SVR,[41–43] can be expressed as

$$y(\mathbf{x}) = b + (\mathbf{w}, \varphi_i(\mathbf{x})) = b + \sum_{i=1}^{l} \alpha_i k(\mathbf{x}, \mathbf{x}_i) \tag{4}$$

where, $k(\boldsymbol{x}, \boldsymbol{x}_i) = (\phi(\boldsymbol{x}), \phi(\boldsymbol{x}_i))$ represents a kernel function, set for the present QSAR development as the commonly used Gaussian kernel[41,49] $k(\boldsymbol{x}, \boldsymbol{x}_i) = \exp(-\gamma||\boldsymbol{x} - \boldsymbol{x}_i||^2)$ (γ is known as kernel width). In eqn (4), $\boldsymbol{x}_i$'s are support vectors determined together with b and $\boldsymbol{w}$ $\left(= \sum_{i=1}^{l} \alpha_i \phi(\mathbf{x}_i)\right)$ from the data by solving the following optimization problem[41–43] with two slack variables ξ_i and $\xi_i{}^\star$.

$$\begin{array}{ll} \min\limits_{\mathbf{w},b,\xi,\xi^\star} & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum\limits_{i=1}^{n}\xi_i + C\sum\limits_{i=1}^{n}\xi_i{}^\star \\ \text{subject to} & \mathbf{w}^T\varphi(x_i) + b - y_i \le \varepsilon + \xi_i \\ & y_i - \mathbf{w}^T\varphi(x_i) - b \le \varepsilon + \xi_i{}^\star \\ & \xi_i, \xi_i{}^\star \ge 0, i = 1, \ldots, n \end{array} \tag{5}$$

In the above formulation, n denotes the total number of training samples, while C and ε are known as regularization factor and tube size, respectively. The ε-SVR model performance depends on proper setting of these two parameters (C and ε) as well as the kernel width γ. In the present QSAR development, according to a practical model parameter selection approach,[48] C and ε were determined as:

$$C = \max(|\bar{y} + 3\sigma_y|, |\bar{y} - 3\sigma_y|)$$

$$\varepsilon = \frac{3\sigma}{\sqrt{n}} \tag{7}$$

where, $\bar{y}$ and σ_y denote the average and standard deviation of the response variable y (*i.e.*, $\log_2$ transformed cell association), respectively. The standard deviation of data noise σ in eqn (7) was estimated for each of the descriptor sets by:[48]

$$\sigma^2 \approx \frac{n^{1/5}k}{n^{1/5}k - 1} \cdot \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{8}$$

where, $\hat{y}$ is the response estimated by k-nearest-neighbor method with the recommendation[48] of k = 3. The optimal kernel width, γ, was determined using a "grid-search"[42,49] from $\gamma = 2^{-9}, 2^{-7}, \ldots, 2^3$. It is noted that, for the ε-SVR development, all the descriptors (PCFs and NPPs) were standardized as $z_i = (d_i - \bar{d})/\sigma_\text{d}$ (*i.e.*, z-score[43]), in which $\bar{d}$ and σ_d denote the mean and standard deviation of descriptor d.

5.2. QSAR performance validation

The performance of the developed QSARs was quantified by the widely used coefficient of determination R^2 as recommended by the OECD guidelines for QSAR development and validation.[40] It is noted that R^2, defined as 1-MSE/σ_y, in which MSE = $\Sigma_i(y_i - y(\boldsymbol{x}_i))/n$ is the mean squared error, can be negative for a QSAR of MSE > σ_y, which would indicate lack of QSAR predictive ability.[68] The 0.632 estimator[35,36] was adopted for QSAR validation since it is particularly suitable for performance validation of models based on small datasets. Accordingly, model prediction accuracy was assessed as $R^2_{\text{E632}} = 0.368R^2_{\text{resub}} + 0.632R^2_{\text{boot}}$, where, R^2_{resub} is the model prediction accuracy assessed using training set (*i.e.*, re-substitution validation) and R^2_{boot} denotes the prediction accuracy in bootstrapping validation.[35,36] For the latter, the out-of-bootstrap samples (*i.e.*, un-sampled ones) used for model testing were ~36.8% ($\approx(1 - 1/n)^n \approx 1/\text{e}$) of the complete dataset.[69] The bootstrapping process was repeated for 200 rounds, consistent with the recommended range[36] of 25–200. In addition, the prediction accuracy of the developed QSARs was also assessed *via* 4-fold cross-validation (CV).[41,43,47] In 4-fold CV, the NPs were randomly partitioned into 4 mutually exclusive subsets, with three subsets used for training and one for validation. This 4-fold CV was repeated 100 times and the average prediction accuracy was used for model assessment.

5.3. Descriptor selection

QSAR development and performance were evaluated in relation to the number of included model descriptors. Consistent with the goal of arriving at QSARs with a reasonably small number of descriptors, in the current analysis the maximum number of descriptors was set at ∼3/10 of the total number of NPs. It is interesting to note that a recommended rule of thumb[47] is that the number of QSAR descriptors should be in the range of 1/10–2/10 of the total NPs. Accordingly, the analysis also included the range of recommended number of descriptors in relation to the number of samples. In order to identify a reasonably small QSAR descriptor subset of relevance to cell association, descriptor selection[33,38,39] was accomplished by sequential forward floating selection (SFFS),[34] which represents an improvement of the traditionally used sequential forward selection (SFS).[33] At each selection step, SFFS first conducts a forward selection to identify the descriptor that leads to the greatest increase in model performance, then backward elimination to evaluate whether previously selected descriptors should be removed due to the addition of the newly selected one.[34] The above selection process proceeds until the prescribed number of descriptors selected. In order to avoid early termination of the descriptor selection process (*i.e.*, addition of a new descriptor that does not improve model performance prior to reaching the target number of descriptors), speculative steps of up to ∼1/10 of the total number of the Au NP was used. The speculative steps proceed the descriptor selection process by including the additional identified descriptor, even in the absence of increased model performance. Based on the on model performance, with respect to the selected descriptors, the suitable descriptor number was then determined by locating the "turning point" being defined when the addition of a new descriptor led to insignificant improvement (*e.g.*, ≲1% increase in R^2) in model performance.[46]

5.4. QSAR robustness and descriptor importance assessment

In order to assess if the developed QSARs are indeed robust models (*i.e.*, not "chance" correlations), they were further validated *via* a 100-round Y-randomization.[17–19] In Y-randomization, models were built using the same set of selected descriptors but for randomized cell association values. The built models are considered as "chance" correlations when compared with the developed QSAR, since Y-randomization distorts the correlation (if a correlation exists) between the selected descriptors and cell association. This accepted approach confirms the robustness of a QSAR if it significantly outperforms the "chance" correlations.[17–19]

The randomization approach was also used to estimate the importance of the selected QSAR descriptors. In this approach, a decrease in the performance of a model developed with randomization of a given descriptor was used to assess the descriptor's importance.[70] Accordingly, a significant decrease in model performance is indicative of increased importance of the descriptor that was randomized.[70]

5.5. Applicability domain analysis

The applicability domain for the linear QSAR was analyzed by William's graph[44,45] in order to identify the descriptor space region in which reasonable QSAR predictions can be made. William's graph depicts a QSAR's applicability domain with a two-dimensional scatter plot. The first dimension is the leverage $h_i = \boldsymbol{x}_i(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_i^{\mathrm{T}}$ (where $\boldsymbol{X}^{\mathrm{T}} = [\boldsymbol{x}_1^{\mathrm{T}}, \boldsymbol{x}_2^{\mathrm{T}}, \ldots, \boldsymbol{x}_n^{\mathrm{T}}]$ identifies the complete NP dataset) that represents the distance of a given NP ($\boldsymbol{x}_i$) to the center of those NPs used for QSAR development in the descriptor space.[44] NPs of smaller h value are more similar to the dataset used for QSAR development and thus are within the NP descriptor domain, in which cell association can be predicted more reliably. In practice, the critical leverage value is commonly set as $h^* = 3(m+1)/n$ covering ∼99% of normally distributed training samples,[44,45] where m is the number of QSAR descriptors and n is the number of training samples (n = 84 in the present work). The second dimension of William's graph is the standardized prediction residual $(e_i - \bar{e})/\sigma_{\mathrm{e}}$, (where $e_i = y_i - y(\boldsymbol{x}_i)$ with $\bar{e}$ and σ_{e} denoting its average and standard deviation respectively) with $[-3, 3]$ as a generally accepted range.[44]

It is emphasized that in the traditional Willam's graph, the leverage only quantifies linear similarity defined by $(\boldsymbol{x}_i, \boldsymbol{x}_j)$. However, in the non-linear ε-SVR model, similarity is defined by a non-linear kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Therefore, leverage may not be suitable for assessing the applicability domain of ε-SVR QSARs. The similarity measure for applicability domain assessment has to be consistent with the one used for QSAR development. Therefore, determination of the application domain for the ε-SVR QSARs was accomplished with replacement of the leverage in William's graph by the average kernel similarity[71] defined below.

$$g_i = (\varphi(\mathbf{x}_i), \frac{1}{n}\sum_{j=1}^{n}\varphi(\mathbf{x}_j)) = \frac{1}{n}\sum_{j=1}^{n}k(\mathbf{x}_i, \mathbf{x}_j) \qquad (9)$$

Similar to the leverage, the average kernel similarity (eqn (9)) quantifies the similarity of a NP $\boldsymbol{x}_i$ to the center of those used for QSAR development as measured by the kernel function. Here, analogous to h^*, the critical kernel similarity g^* was defined as the 99% percentile of the kernel similarity for all training samples.

Acknowledgements

This material is based upon work supported by the National Science Foundation and the Environmental Protection Agency under Cooperative Agreement Number DBI-0830117, the Natural Sciences and Engineering Research Council of Canada, Canadian Institute of Health Research, and Canadian Research Chairs Program. We acknowledge Drs. A. Emili, J. Olsen and H. Guo for their contribution to building the protein corona dataset. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Environmental Protection Agency. This work has not been subjected to EPA review and no official endorsement should be inferred.

References

1 Z. Guo and L. Tan, *Fundamentals and Applications of Nanomaterials*, Artech House, 2009.

2 The Wilson Center, Inventory Finds Increase in Consumer Products Containing Nanoscale Materials, 2013, at http://www.nanotechproject.org/cpi/.

3 D. Peer, *et al.*, Nanocarriers as an emerging platform for cancer therapy, *Nat. Nanotechnol.*, 2007, **2**, 751–760.

4 A. Nel, T. Xia, L. Mädler and N. Li, Toxic potential of materials at the nanolevel, *Science*, 2006, **311**, 622–627.

5 S. Sharifi, *et al.*, Toxicity of nanomaterials, *Chem. Soc. Rev.*, 2012, **41**, 2323–2343.

6 W. Jiang, B. Y. S. Kim, J. T. Rutka and W. C. W. Chan, Nanoparticle-mediated cellular response is size-dependent, *Nat. Nanotechnol.*, 2008, **3**, 145–150.

7 A. Kahru and H.-C. Dubourguier, From ecotoxicology to nanoecotoxicology, *Toxicology*, 2010, **269**, 105–119.

8 H. Zhang, *et al.*, Use of metal oxide nanoparticle band gap to develop a predictive paradigm for oxidative stress and acute pulmonary inflammation, *ACS Nano*, 2012, **6**, 4349–4368.

9 P. V. Asharani, Y. Lianwu, Z. Gong and S. Valiyaveettil, Comparison of the toxicity of silver, gold and platinum nanoparticles in developing zebrafish embryos, *Nanotoxicology*, 2011, **5**, 43–54.

10 R. Hardman, A toxicologic review of quantum dots: toxicity depends on physicochemical and environmental factors, *Environ. Health Perspect.*, 2006, **114**, 165–172.

11 X. Wang, *et al.*, Dispersal state of multiwalled carbon nanotubes elicits profibrogenic cellular responses that correlate with fibrogenesis biomarkers and fibrosis in the murine lung, *ACS Nano*, 2011, **5**, 9772–9787.

12 S. Y. Shaw, *et al.*, Perturbational profiling of nanomaterial biologic activity, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 7387–7392.

13 Y. Cohen, R. Rallo, R. Liu and H. H. Liu, In silico analysis of nanomaterials hazard and risk, *Acc. Chem. Res.*, 2012, **46**, 802–812.

14 T. Puzyn, *et al.*, Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles, *Nat. Nanotechnol.*, 2011, **6**, 175–178.

15 C. Sayes and I. Ivanov, Comparative study of predictive computational models for nanoparticle-induced cytotoxicity, *Risk Anal.*, 2010, **30**, 1723–1734.

16 R. Liu, *et al.*, Classification NanoSAR development for cytotoxicity of metal oxide nanoparticles, *Small*, 2011, **7**, 1118–1126.

17 R. Liu, *et al.*, Development of structure–activity relationship for metal oxide nanoparticles, *Nanoscale*, 2013, **5**, 5644–5653.

18 D. Fourches, *et al.*, Quantitative nanostructure–activity relationship modeling, *ACS Nano*, 2010, **4**, 5703–5712.

19 R. Liu, *et al.*, Nano-SAR development for bioactivity of nanoparticles with considerations of decision boundaries, *Small*, 2013, **9**, 1842–1852.

20 Y. T. Chau and C. W. Yap, Quantitative nanostructure–activity relationship modelling of nanoparticles, *RSC Adv.*, 2012, **2**, 8489–8496.

21 M. Ghorbanzadeh, M. H. Fatemi and M. Karimpour, Modeling the cellular uptake of magnetofluorescent nanoparticles in pancreatic cancer cells: A quantitative structure activity relationship study, *Ind. Eng. Chem. Res.*, 2012, **51**, 10712–10718.

22 A. A. Toropov, *et al.*, QSAR as a random event: Modeling of nanoparticles uptake in PaCa2 cancer cells, *Chemosphere*, 2013, **92**, 31–37.

23 C.-Y. Shao, *et al.*, Dependence of QSAR models on the selection of trial descriptor sets: a demonstration using nanotoxicity endpoints of decorated nanotubes, *J. Chem. Inf. Model.*, 2013, **53**, 142–158.

24 V. C. Epa, *et al.*, Modeling biological activities of nanoparticles, *Nano Lett.*, 2012, **12**, 5808–5812.

25 S. Kar, A. Gajewicz, T. Puzyn and K. Roy, Nano-quantitative structure–activity relationship modeling using easily computable and interpretable descriptors for uptake of magnetofluorescent engineered nanoparticles in pancreatic cancer cells, *Toxicol. In Vitro*, 2014, **28**, 600–606.

26 C. D. Walkey, J. B. Olsen, H. Guo, A. Emili and W. C. W. Chan, Nanoparticle size and surface chemistry determine serum protein adsorption and macrophage uptake, *J. Am. Chem. Soc.*, 2012, **134**, 2139–2147.

27 C. D. Walkey, *et al.*, Protein corona fingerprinting predicts the cell association of gold nanoparticles, *ACS Nano*, 2014, **8**, 2439–2455.

28 C. D. Walkey and W. C. W. Chan, Understanding and controlling the interaction of nanomaterials with proteins in a physiological environment, *Chem. Soc. Rev.*, 2012, **41**, 2780–2799.

29 X.-R. Xia, N. A. Monteiro-Riviere and J. E. Riviere, An index for characterization of nanomaterials in biological systems, *Nat. Nanotechnol.*, 2010, **5**, 671–675.

30 S. Wan, *et al.*, The "Sweet" Side of the Protein Corona: Effects of Glycosylation on Nanoparticle–Cell Interactions, *ACS Nano*, 2015, **9**, 2157–2166.

31 A. Albanese, *et al.*, Secreted biomolecules alter the biological identity and cellular interactions of nanoparticles, *ACS Nano*, 2014, **8**, 5515–5526.

32 S. Wold, M. Sjöström and L. Eriksson, PLS-regression: A basic tool of chemometrics, *Chemom. Intell. Lab. Syst.*, 2001, **58**, 109–130.

33 H. Liu and H. Motoda, *Computational Methods of Feature Selection*, Chapman and Hall/CRC, 2008.

34 P. Pudil, J. Novovičová and J. Kittler, Floating search methods in feature selection, *Pattern Recognit. Lett.*, 1994, **15**, 1119–1125.

35 U. M. Braga-Neto and E. R. Dougherty, Is cross-validation valid for small-sample microarray classification?, *Bioinformatics*, 2004, **20**, 374–380.

36 B. Efron, Estimating the error rate of a prediction rule: Improvement on cross-validation, *J. Am. Stat. Assoc.*, 1983, **78**, 316–331.

37 R. Liu, R. Rallo and Y. Cohen, Unsupervised feature selection using incremental least squares, *Int. J. Inf. Technol. Decis. Mak.*, 2011, **10**, 967–987.

38 R. Liu and Y. Shi, Spatial distance join based feature selection, *Eng. Appl. Artif. Intell.*, 2013, **26**, 2597–2607.

39 H. Liu and L. Yu, Toward integrating feature selection algorithms for classification and clustering, *IEEE Trans. Knowl. Data Eng.*, 2005, **17**, 491–502.

40 Guidance Document on the Validation of (Quantitative) Structure–Activity Relationship [(Q)SAR] Models, 2007.

41 C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

42 C.-C. Chang and C.-J. Lin, LIBSVM: A library for support vector machines, *ACM Trans. Intell. Syst. Technol.*, 2011, **2**, 27.

43 J. Han, M. Kamber and J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2011.

44 J. Jaworska, N. Nikolova-Jeliazkova and T. Aldenberg, QSAR applicabilty domain estimation by projection of the training set descriptor space: a review, *Altern. Lab. Anim.*, 2005, **33**, 445–459.

45 T. Puzyn, D. Leszczynska and J. Leszczynski, Toward the development of 'Nano-QSARs': Advances and challenges, *Small*, 2009, **5**, 2494–2509.

46 J. Xu, L. Wang, H. Zhang, X. Shen and G. Liang, Quantitative structure-property relationships studies on free-radical polymerization chain-transfer constants for styrene, *J. Appl. Polym. Sci.*, 2012, **123**, 356–364.

47 T. Puzyn, J. Leszczynski and M. T. Cronin, *Recent Advances in QSAR Studies: Methods and Applications*, Springer, 2010.

48 V. Cherkassky and Y. Ma, Practical selection of SVM parameters and noise estimation for SVM regression, *Neural Networks*, 2004, **17**, 113–126.

49 C.-W. Hsu, C.-C. Chang and C.-J. Lin, *et* al., A practical guide to support vector classification, 2003, at http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

50 Y. Fujioka, T. Taniguchi, Y. Ishikawa and M. Yokoyama, Significance of acidic sugar chains of apolipoprotein B-100 in cellular metabolism of low-density lipoproteins, *J. Lab. Clin. Med.*, 2000, **136**, 355–362.

51 J. H. Contois, G. R. Warnick and A. D. Sniderman, Reliability of low-density lipoprotein cholesterol, non-high-density lipoprotein cholesterol, and apolipoprotein B measurement, *J. Clin. Lipidol.*, 2011, **5**, 264–272.

52 S. Tenzer, *et al.*, Nanoparticle size is a critical physicochemical determinant of the human blood plasma corona: A comprehensive quantitative proteomic analysis, *ACS Nano*, 2011, **5**, 7155–7167.

53 M. Lundqvist, *et al.*, Nanoparticle size and surface properties determine the protein corona with possible implications for biological impacts, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 14265–14270.

54 R. Cukalevski, *et al.*, Structural changes in apolipoproteins bound to nanoparticles, *Langmuir*, 2011, **27**, 14360–14369.

55 N. Pirooznia, S. Hasannia, A. Lotfi and M. Ghanei, Encapsulation of Alpha-1 antitrypsin in PLGA nanoparticles: In Vitro characterization as an effective aerosol formulation in pulmonary diseases, *J. Nanobiotechnol.*, 2012, **10**, 20–35.

56 S. Sohrab, *et al.*, Mechanism of alpha-1 antitrypsin endocytosis by lung endothelium, *FASEB J.*, 2009, **23**, 3149–3158.

57 S. Dekali, *et al.*, Cell cooperation and role of the P2X7 receptor in pulmonary inflammation induced by nanoparticles, *Nanotoxicology*, 2013, **7**, 1302–1314.

58 G. Qi, L. Li, F. Yu and H. Wang, Vancomycin-modified mesoporous silica nanoparticles for selective recognition and killing of pathogenic gram-positive bacteria over macrophage-like cells, *ACS Appl. Mater. Interfaces*, 2013, **5**, 10874–10881.

59 H. H. Liu, S. Surawanvijit, R. Rallo, G. Orkoulas and Y. Cohen, Analysis of nanoparticle agglomeration in aqueous suspensions via constant-number Monte Carlo simulation, *Environ. Sci. Technol.*, 2011, **45**, 9284–9292.

60 K. Garner and A. Keller, Emerging patterns for engineered nanomaterials in the environment: a review of fate and toxicity studies, *J. Nanopart. Res.*, 2014, **16**, 2503.

61 Z.-G. Yue, *et al.*, Surface charge affects cellular uptake and intracellular trafficking of Chitosan-Based nanoparticles, *Biomacromolecules*, 2011, **12**, 2440–2446.

62 D. K. Strickland and M. Z. Kounnas, Mechanisms of cellular uptake of Thrombin-Antithrombin II complexes role of the low-density lipoprotein receptor-related protein as a serpin-enzyme complex receptor, *Trends Cardiovasc. Med.*, 1997, **7**, 9–16.

63 D. Berry, D. M. Lynn, R. Sasisekharan and R. Langer, Poly(β-amino ester)s promote cellular uptake of heparin and cancer cell death, *Chem. Biol.*, 2004, **11**, 487–498.

64 C. Oelschlager, *et al.*, Antithrombin III inhibits nuclear factor kappaB activation in human monocytes and vascular endothelial cells, *Blood*, 2002, **99**, 4015–4020.

65 L. A. Miles, *et al.*, The plasminogen receptor, Plg-Rkt, and macrophage function, *J. Biomed. Biotechnol.*, 2012, **2012**, 250464.

66 J. E. Testa and J. P. Quigley, Protease receptors on cell surfaces: new mechanistic formulas applied to an old problem1, *J. Natl. Cancer Inst.*, 1988, **80**, 712–714.

67 S. J. Busuttil, *et al.*, A central role for plasminogen in the inflammatory response to biomaterials, *J. Thromb. Haemost.*, 2004, **2**, 1798–1805.

68 R. D. Cramer III, J. D. Bunce, D. E. Patterson and I. E. Frank, Crossvalidation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies, *Quant. Struct. Relatsh.*, 1988, **25**, 18–25.

69 J. B. O. Mitchell, Machine learning methods in chemoinformatics, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2014, 468–481.

70 V. Svetnik, *et al.*, Random forest: A classification and regression tool for compound classification and QSAR modeling, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1947–1958.

71 N. Fechner, A. Jahn, G. Hinselmann and A. Zell, Estimation of the applicability domain of kernel-based machine learning models for virtual screening, *J. Cheminform.*, 2010, **2**, 2.