The Application of QSAR Approaches to Nanoparticles

Jacques Ehret,[1] Martina Vijver[2] and Willie Peijnenburg[2,3]

[1]Helmholtz-Zentrum München — German Research Centre for Environmental Health (GmbH), Institute of Molecular Exposomics, Neuherberg, Munich, Germany; [2]Leiden University, Institute of Environmental Science, Leiden, The Netherlands; [3]RIVM, Centre for Safety of Substances and Products, Bilthoven, The Netherlands

Summary — Nanoparticles (NPs) are increasingly used throughout the world for many purposes. The resulting exposure increases the relevance of efforts to assess their effects. The activities of NPs are related to many structural features, including their shape, composition and size. Applying Quantitative Structure–Activity Relationship (QSAR) methods to nanoscale systems becomes challenging, due to the lack of data and insight into the fate and effects of NPs. In this study, the possible use of QSAR methods on NPs is investigated. To this intent, several ways of representing and describing NPs were tested by using different data mining methods. The main conclusion is that QSAR methods are relevant for the study of the activity of NPs, but this should be confirmed by using larger and more diverse sets of data. Moreover, representing the constitution of NPs (in terms of core, coating and surface modification) significantly increases the prediction accuracy of the models. In our case, the most significant features to be represented were found to be the core and surface modification.

Key words: *nanoparticles, QSAR*.

Address for correspondence: *Jacques Ehret, Helmholtz-Zentrum München — German Research Centre for Environmental Health (GmbH), Institute of Molecular Exposomics, Ingolstädter Landstrasse 1, Neuherberg D-85764, Munich, Germany.*
E-mail: *jacques.ehret@helmholtz-muenchen.de*

Introduction

Nanoparticles

Nanoparticles (NPs) are defined as "a sub-classification of ultrafine particles with lengths in two or three dimensions greater than 0.001 micrometer (1 nanometer) and smaller than about 0.1 micrometer (100 nanometers)" (1). Thus, nanomaterials have structural features in between those of atoms and bulk materials. To modify, or optimise, surface properties (i.e. stability in solution, reactivity, selectivity), it is usual to coat NPs with atoms, molecules, or particles. NPs are becoming considerably more-widely used for a variety of purposes, such as in cosmetics, sunscreens and food packaging. As a consequence, there is an increase in the exposure of living beings to NPs.

Shevchenko *et al*. (2) stated that "an ensemble of nanoparticles is a strongly nonequilibrium nonlinear multivariant system. There are no grounds to believe that, in the course of the evolution, this ensemble should tend to homogenisation rather than to a new hierarchic order according to the self-organisation principle. This suggests that the structural inhomogeneity is a fundamental property of the nanostate". In other words, a system of NPs is constantly evolving. Therefore, it is uncertain whether an observed effect is due to the particle itself or to its evolution (i.e. the formation of agglomerates or aggregates). Moreover, NPs have a high structural diversity to begin with — nine categories have been described, including different shapes, and diversity among coatings and surface modifications (3). NPs are functionally diverse, because many features affect their activity, namely, size distribution, agglomeration state, shape, porosity, surface area, chemical composition, structure-dependent electronic configuration, surface chemistry, surface charge, and crystal structure (4). These features, which are highly diverse and constantly evolving, serve to influence the activity of NPs and thus make accurate measurements both costly and challenging. Quantitative Structure–Activity Relationship (QSAR) approaches can be used to lower costs and double-check measurements.

Quantitative Structure–Activity Relationships

Quantitative Structure–Activity Relationship (QSAR) approaches apply informatics possibilities to the Structure–Activity Relationship (SAR)

assumption, which states that there is a correlation between the structure of a molecule and its activity.

A database of molecules, which includes a targeted activity or property, is used. Structure and activity cannot be linked directly, so one first has to represent the molecule. The different types of molecular representation (for example, Lewis, Cram) are partial reflections of the chemical reality. One representation does not provide information about the whole chemical reality of a molecule, and biased representations lead to a false SAR (e.g. the activity would be linked to structural features that were not related). Therefore, different representations of the structural information have to be tested.

Then, a link between representations and computations has to be made, and it was in this context that chemometricians developed so-called 'descriptors'. According to Todeschini and Consonni (5): "A molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardised experiment." Descriptors are numbers, and the representation of a molecule is described by a vector of descriptors. All those vectors are merged into a matrix of descriptors, which is the description of the initial molecular data set.

Subsequently, data mining algorithms are used on the descriptor matrix, in order to extract latent knowledge to build prediction models. Some examples of machine learning algorithms are Artificial Neural Networks (ANN; 6, 7), Support Vector Machines (SVM; 8), Partial Least Squares (PLS; 9, 10) and Multi Linear Regression (MLR; 11).

To analyse a model's accuracy, statistical parameters are calculated. Usually the coefficient of determination (called r^2 or q^2) is calculated for this purpose (12), as well as the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE; see Equations 1–3). Determination coefficients are a measure of the correlation between the prediction that is made and the perfect prediction (where prediction = measurement). The MAE and RMSE provide information about the average error of predictions, with the difference that RMSE is more sensitive to outliers.

— Coefficient of determination (q^2)

$$q^2 = 1 - \frac{\sum_{i=1}^{n} (y_i exp - y_i pred)^2}{\sum_{i=1}^{n} (y_i exp - \bar{y}_i pred)^2} \qquad \text{[Equation 1]}$$

— Mean Absolute Error (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i pred - y_i exp| \qquad \text{[Equation 2]}$$

— Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i pred - y_i exp)^2}{n}} \qquad \text{[Equation 3]}$$

Experimental data contain noise (13). Learning from noise is called over-fitting (14) and decreases the model's accuracy. To reduce such an effect, a validation set and test sets are usually used. One of the most common validations is the k-fold cross validation. This validation method requires the division of the data set into k folds: k–1 folds are used to train a model that is then used to predict the values of the last fold (called the test set). New models are calculated until all the data were predicted once. The overall accuracy is calculated on the predicted values of all test sets.

Quantitative Nanostructure–Activity Relationship (QNAR) Modelling: Applying QSAR Modelling to Nanoparticles

When applying QSAR modelling to nanoparticles, the representation and description of the particles have to be carefully considered. Data mining methods should work with matrices built on NP information (since such methods can be used on any kind of data), but features in NPs that influence their activity (the ones to be represented and described) are highly diverse and fluid compared to those of usual molecules.

An ontology for NPs has been developed. An ontology is a formal, explicit representation of knowledge belonging to a subject area (15). Based on this ontology, an accurate representation of NPs requires information on the core, coating, shell and surface of each NP. In addition, as NPs are in a state of meta-stable equilibrium, information on the medium–particle interface and the evolution of the particle should also be provided. A more thorough description of a NP's surface modifications is conducted, because it is, by definition, at the interface with the medium. To describe the surface–medium interface, surface area, charge and zeta potential have to be considered. Zeta potential is a measure of the stability of a particle in a given medium.

Moreover, specific features of NPs should also be considered, such as particle size, shape and mass. Size, for instance, can significantly modify the properties of NPs. To provide information about the size of particles, the Average Particle Size (APS) value is often used. However, one usually has a mixture of NPs that are either different (in order to combine two specific properties) or just at different aggregation states. In such cases, the APS is irrelevant, because if two different sizes of particles are present in a single medium, then

information on the average size is not representative of the true situation.

A data set taking into account the biological activity of NPs (assessed by multiple physiological cell-based assays, in different cell types, and at various doses) was published by Shaw *et al.* in 2008 (16). Composed of 50 NPs, chemical information about the core, coating and surface modification is provided for each NP, as well as four experimental values: zeta potential, relaxivities R1 and R2, and APS. Relaxivities describe how fast spins of NPs return to the equilibrium distribution after a nuclear magnetisation. Fourches *et al.* (17) showed, with Shaw's data set, that QNAR modelling is a valid approach. However, these authors did not use the intrinsic structure of NPs, but only the experimental measurements (zeta potential, R1, R2, APS) after separating their data set in three clusters of molecules (determined by using a hierarchical clustering procedure).

Materials and Methods

The aim of our study was to test whether representing and describing features of the structure of a NP (such as the core, coating and surface modifications) improves the accuracy of QSAR models, as compared to the use of only experimental data. Therefore, we used Shaw's data set of NPs, represented it according to NP ontology (15), calculated descriptors from such representation, and mined the resulting data.

Building the models

The data set records information on the core, coating and surface modification of NPs, together with experimental data. It comprises 50 NPs with two different cores, five different coatings, 17 surface modifications, and four different experimental values. The endpoint used is the biological activity of a nanomaterial, assessed by multiple physiological cell-based assays in multiple cell types, and at multiple doses. The biological activity data that were used are from Fourches *et al.* (17). All structural information was manually converted to a format that could be read by the software used to calculate the descriptors, i.e. the line notation SMILES (Simplified Molecular-Input Line-Entry System; 18). The websites of chemical providers (19) and ChemSpider (20) were used to calculate the SMILES notation. The structures of some surface modifications were not available (not known or not published by the manufacturer). Therefore, we calculated the models by using only particles for which full details were available, and excluded the unknowns from the data set. The particles excluded due to lack of information were: NP6, NP22, NP23, NP33, NP37 (the number corresponds to the IDs in the Fourches data set; 17).

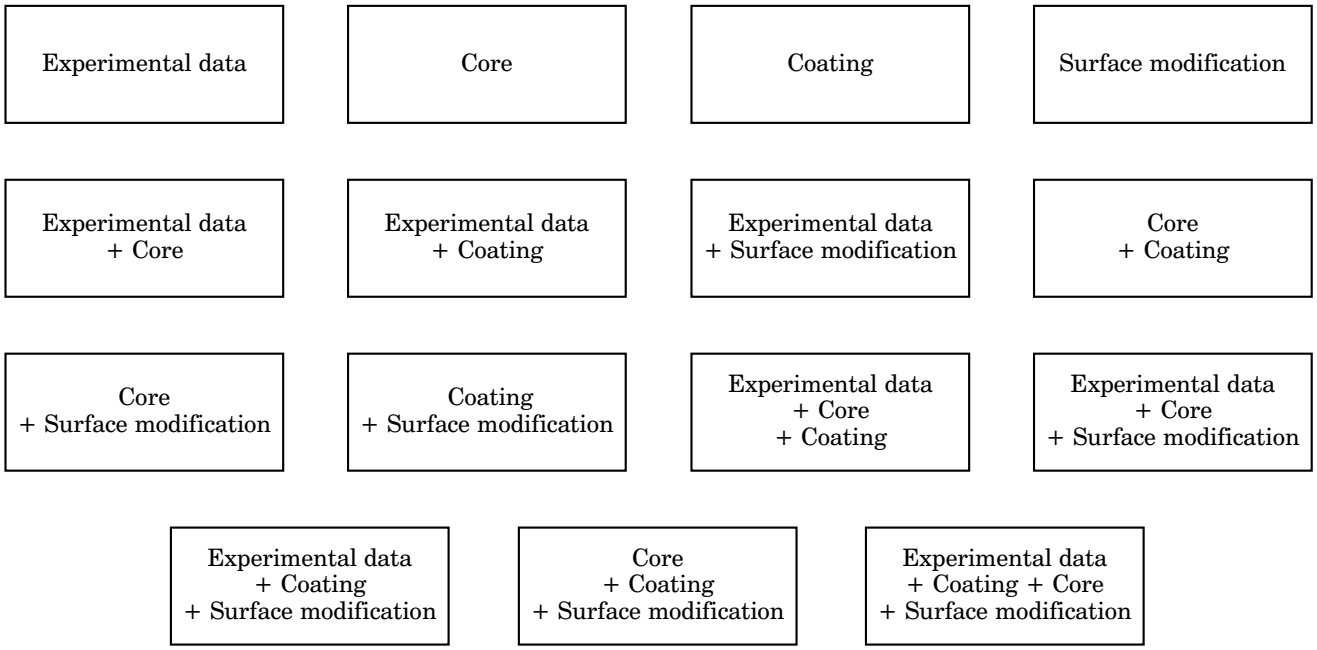
Several representations were tested: with and without experimental data, core, coating, and surface modification. In addition, for each combination, all the different ways of describing data were tested. The different representations are illustrated in Figure 1 and the descriptions are shown in Figure 2. As an example, for the representation comprising information on Experimental Data + Surface Modification, we used five descriptor matrices. All contained the four experimental values; one contained PaDEL descriptors, one Estate descriptors, one topological descriptors, one electronic descriptors, and one topological and electronic descriptors.

To calculate descriptors we used DRAGON (21) and PaDEL software (22). Three-dimensional (3-D) descriptors were not calculated, due to the lack of reliable 3-D information inherent to the NPs. With DRAGON, we calculated Estate (23, 24), topological and electronic descriptors. In the case of topological descriptors, the following DRAGON categories were included: constitution, topological and walking path. These represent 164 descriptors. With regard to electronic descriptors, Estate and charges descriptors were integrated, which resulted in 185 different descriptors. Subsequently, both descriptors (electronic and topological) were merged to obtain a 349-descriptor matrix, in order to check whether both types of information were needed to accurately represent the molecule. With PaDEL, we calculated 61 descriptors from the Chemistry Development Kit (CDK; 25, 26). We also used energy values, and random numbers to check whether data mining algorithms needed measured information on the composition of the NP, or needed only differentiated categories. The energy values included are dipole momentum, electron energy, nuclease repulsion, and binding energy. Figure 2 shows the types of descriptors that were calculated and used for each part of the NPs.

To mine all the descriptor sets, the open source software WEKA, developed by the University of Waikato (27), was used. It was considered accurate, as it was peer-reviewed, and convenient, due to the possibility of launching calculations by using Bash scripts. The data were filtered by removing nominal values and useless values (which do not vary at all), and normalising the descriptors between 0 and 1. All this filtering was unsupervised. Several data mining methods, such as ANN, MLR, PLS and SVM, were evaluated. We then focused on the methods that gave the best results, which were PLS, SVM with Radial Based Function (RBF) and SVM with PolyKernel.

For each data mining method, specific parameters were modified. For PLS, the number of principal components was changed from 1 to 15. For both

Figure 1: An overview of the 15 different representations that were tested in our approach



For each representation all the corresponding descriptions were calculated. The descriptions are shown in Figure 2.

Figure 2: Description of the cores, coatings and surface modifications that were calculated for each part of the NPs

Core	**Coating**	**Surface modification**
4 Cases:	**6 Cases:**	**5 Cases:**
— PaDEL	— PaDEL	— PaDEL
— Information about energies	— Random number	— DRAGON Estate
— Random number	— DRAGON Estate	— DRAGON topologic
— DRAGON Estate	— DRAGON topologic	— DRAGON electronic
	— DRAGON electronic	— DRAGON both
	— DRAGON both	

Each of these descriptions were used on every different representation (Figure 1). To describe experimental features, all experimental data were used.

SVMs, the tolerance parameter was modified from 0.010 to 0.700 in increments of 0.012. Moreover, each model was cross-validated by using five folds. To lower the impact of chance on the selection of each fold, different seeds for random numbers were used, from 0 to 20, in increments of 1, for each data mining method.

Assessment of model accuracy

We used the open source software R (28) to analyse the accuracy of our prediction models. To assess statistical significance between two representations (which means that the difference between them is unlikely to have happened by chance), we performed

a paired Student's *t*-test to compare two representations with the same parameters (PC, average seed and tolerance). For instance, comparing the average of the 21 values calculated by using different seeds with the method parameter '3 principal components' with data set 1, paired with method parameter '3 principal components' with data set 2. We could use this test because the values followed a normal distribution. The output of this test is whether representation A is significantly better than representation B, considering all the different parameters.

Then, to compare the accuracies of the prediction models with different descriptions, we used a binomial test. For example, to answer the question 'Does describing the core by using random numbers improve the model or not?', we compared all descriptions that used random numbers to describe the core, to all descriptions that did not use them. For both tests, the p value had to be lower than 0.01 to be considered significant.

It was observed that high values for parameters (14–15 principal components for PLS, and over 0.550 tolerance for SVM) led to inconsistent outcomes. For PLS, a large amount of principal components left the models prone to over-fitting, which decreased their predictive accuracy. In addition a convergence was observed for both SVMs (no further modification of the prediction's accuracy after a defined tolerance value). To avoid biased results in significance tests, we excluded values above a cut-off of 13 principal components in PLS, and over 0.550 tolerance for SVM.

Results

Experimental data

We first calculated the accuracy of the predictions with only experimental data. Such models performed really badly. However, when some information about the constitution of the NPs was added (e.g. representation of the core, coat or surface modifications), the accuracy of the predictions increased.

We then compared models that used and did not use experimental data (see Table 1). When experimental data were used, there was a significant decrease in the accuracy of the predictions for PLS, SVM RBF and SVM Polykernel (SVMK2) methods for q^2, RMAE and RRMSE (which are the relative MAE and RMSE). This was surprising, because information regarding the surface–medium interface is relevant for a NP's activity.

Core

We calculated that there is a significant improvement in the accuracy of the predictions when the NP's core is represented (p value < 0.01 with paired Student's t-test, comparing the use of core description).

The performance of the different descriptors depended on the data mining method used — random descriptors that used PLS, and DRAGON descriptors that used SVM Polykernel, gave the best results. However, it was observed that neither describing the core by using many (DRAGON or PaDEL) or few descriptors (energy values), nor by using random numbers (1 for the first core, 2 for the second core), resulted in significant differences. It means that, statistically speaking, describing the core with random numbers provided results as good as those obtained with other descriptors. To illustrate this issue, Table 2 shows the results of the comparison between core random description and core description with energies. It shows an improvement in the accuracy of the predictions. However, this improvement is not significant, when random values are used.

Table 1: A summary of the results of the paired Student's *t*-test comparing a representation with and without experimental data

With ED compared to without ED		Prediction accuracy	Significance of the difference
PLS	q^2	Significant	Decreases
	RMAE	Significant	Decreases
	RRMSE	Significant	Decreases
SVM	q^2	Significant	Decreases
	RMAE	Significant	Decreases
	RRMSE	Significant	Decreases
SVMK2	q^2	Significant	Decreases
	RMAE	Significant	Decreases
	RRMSE	Significant	Decreases

The comparisons show that using experimental data (ED) lowers the model's prediction accuracy.

Coating

We calculated that representing the coating did not significantly improve the prediction models (p value < 0.01 with paired Student's t-test between the presence or absence of coating description). However, better results were obtained with molecular descriptors than with only random numbers; according to binomial tests, this improvement was not judged to be significant (p value = 0.23). Furthermore, there was no significant difference between the use of all descriptors or only topological ones. Although a difference was observed when only electronic descriptors were used, the significance of this improvement was not consistent throughout all the machine learning algorithms. Adding electronic description of the coating of NPs improved the accuracy of the models, but the most important features to describe are topological.

Surface modification

According to our results, surface modification was the most important feature that should be represented and described. The most significant improvements in the accuracy of the prediction models (according to the paired Student's t-test) were achieved by using representations of surface modifications.

We then compared topological and electronic descriptions, separately and in combination. Even if the representation of surface modification was found to be necessary, the choice of description was not consistent and depended on the machine learning method that was used. No significant improvement was found between the different descriptors investigated, as long as there was a description of the surface modification.

Discussion

An issue with the data set

The data set we used comprises 50 NPs, which is rather small for QSAR modelling. Describing the core, coating and surface modification separately, with the usual descriptors, for this data set, would result in an unbalanced matrix (easily 50 rows by more than 300 columns). Such situations lead to over-fitting. Moreover, we could not split such a data set fairly into three sets (training, test and validation), because too little information would be within the training set. That is why all the calculated models were cross-validated by using a 5-cross validation and not externally tested afterwards. We believe this is a good compromise between not over-fitting data and having too little knowledge to train a model.

Comments on the results

Experimental data in the form of zeta potential, relaxivities and particle size contain information that matters for the activity of NPs. Therefore, we expected to obtain better results with experimental values. However, the contrary was observed. A possible explanation is that experimental values are not precise enough. For instance, APS (one of the four properties investigated) can be insufficiently descriptive, if there is more than one size of particle present. APS is the average of all particle sizes in the solution, so it would not describe all particles correctly. An APS of 50nm could describe a suspension of 50nm-sized particles or a mixture ratio 1:1 of particles 30nm and 70nm in size. For NP26–30, the size (according to Shaw's paper [16]) is between 20–60nm, which confirms the descriptivity of the experimental data. Fourches *et al.* used the average size, i.e. 40nm (17). Moreover, for NPs with IDs from 26 to 44, there is imprecision with regard to the relaxivities — in Shaw's set, the value is literally < 0.5, while Fourches used 0.5 as the value for his models. This is why we observed better results when not relying on experimental data.

It is noteworthy that our data set groups only two types of core, which is very limited. We showed that it is necessary to describe cores, but a random description (values 1 and 2) works just as well as an accurate description. This indicates that models need only to discriminate between the two different cores. With this set, we cannot state whether molecular descriptors, simple energies, or other descriptions should be used. Since cores are often metallic, with a crystalline organisation, we thought that it could be relevant to use group theory-based descriptors (which provide information about molecular symmetry). However, lack of data prevented us from doing this.

We also found that representing the coating did not improve the models. This could be due to the fact that the coating is not necessary for the targeted activity (although this would be counter-intuitive), or due to a poor description of the coating. To solve this issue, a more-direct description of the coating should be carried out. Indeed, our coatings are polymers, and descriptors were calculated according to their monomeric structure. Finally, surface modification is the most important feature to describe. This can be understood by the fact that it is situated at the interface between medium and particle, and thus will directly interact with the target.

Conclusions

With this work, our aim was to show that QSAR is indeed a valid approach for predicting the activity

Table 2: A summary of the results of the binomial test comparing descriptions of the core of NPs using energy values and random values

Energy values compared to random values		Prediction accuracy	Significance of the difference
PLS	q^2	Significant	Increases
	RMAE	Significant	Increases
	RRMSE	Not significant	Increases
SVM	q^2	Not significant	Increases
	RMAE	Not significant	Increases
	RRMSE	Not significant	Increases
SVMK2	q^2	Not significant	Increases
	RMAE	Not significant	Increases
	RRMSE	Not significant	Increases

of NPs, and that representing them according to our nanoparticle ontology provides better results than the representations used in other earlier articles.

We have shown that the use of experimentally-measured properties is relevant for QSARs related to NPs, but the accuracy of those data is of real importance. For instance, uncertainty about NP size tremendously lowers the accuracy of the prediction. A system has to be correctly represented by such values — for example, APS is not relevant enough, but size distribution could be. Furthermore, describing all the constituents (core, coating, surface modification) of NPs significantly increases the accuracy of the models' predictions. Even though we cannot state which kind of description should be performed (we did not identify the descriptors to be used on the different elements of a NP), we proved that molecular descriptors can accurately describe surface modifications. Besides, we confirmed that our nanoparticle ontology approach is valid, and further research should be carried out to further develop it.

We reveal that the emphasis with NPs should not be placed on the concept and development of new descriptors but instead on their representation. We now believe that cores, coatings and surface modifications should be described with descriptors reflecting the specificities of each part — i.e. making use of orbital and symmetry-based descriptors for metals, topological descriptors for organic groups, and so on. This description should be combined with accurate measurements of interactions of the particle within the medium, and specific parameters, such as size and porosity. To carry out a proper QSAR on a NP, there should be enough particles to create an external validation set, which should have roughly the same amount of compounds as the training set. The data set used in our study was small for QSAR approaches, contained some unknown structural features, and some data were not accurate enough. Therefore, further work should be done on larger data sets, including training and external test sets, and more-accurate experimental data with knowledge of each part of the NPs used.

Acknowledgments

This study was supported by the FP7 MC ITN project 'Environmental ChemOinformatics' (Grant Agreement No. 238701). In addition, the work was sponsored by the NATO SFP project 'Ecotoxicity of metal and metal oxide nanoparticles: Experimental study and modeling' (Research Grant EAP SFPP 984401). The first author thanks Dr Igor V. Tetko and Prof. Dr Karl-Werner Schramm for their support.

References

1. Anon. (2012). *ASTM Standard E 2456-06(2012)*, 4pp. West Conshohocken, PA, USA: ASTM International.
2. Shevchenko, V.Y., Madison, A.E. & Mackay, A.L. (2003). The structural diversity of the nanoworld. *Glass Physics & Chemistry* **296**, 577–582.
3. Maynard, A.D. & Aitken, R.J. (2007). Assessing exposure to airborne nanomaterials: Current abilities & future requirements. *Nanotoxicology* **11**, 26–41.
4. Oberdörster, G., Maynard, A., Donaldson, K., Castranova, V., Fitzpatrick, J., Ausman, K., Carter, J., Karn, B., Kreyling, W., Lai, D., Olin, S., Monteiro-Riviere, N., Warheit, D. & Yang, H. (2005). Principles for characterizing the potential human health effects from exposure to nanomaterials: Elements of a screening strategy. *Particle & Fibre Toxicology* **28**.
5. Todeschini, R. & Consoni, V. (2003). *Handbook of Molecular Descriptors, Volume 3*, pp. 12. Weinheim, Germany: Wiley-VCH.
6. Livingstone, D.J. (ed.) (2008). *Artificial Neural Networks: Methods & Applications*, pp. 185–202. New York, NY, USA: Humana Press.
7. Ajay, A. (1993). A unified framework for using neural networks to build QSARs. *Journal of Medicinal*

Chemistry **36**, 3565–3571.
8. Cortes, C. & Vapnik, V. (1995). Support-vector network. *Machine Learning* **203**, 273–297.
9. Wold, H. (1982). Soft modeling: The basic design and some extensions. In *Systems Under Indirect Observation*, Vol. 2 (ed. K.G. Jöreskog & H. Wold) pp. 1–54. Amsterdam, The Netherlands: Elsevier Science Ltd.
10. Wold, S. (1995). PLS for multivariate linear modeling. In *Chemometrics Methods in Molecular Design*, pp. 195–218. Weinheim, Germany: Wiley-VCH.
11. Clementi, S. & Wold, S. (1995). How to choose the proper statistical method. In *Chemometrics Methods in Molecular Design*. pp. 319–336. Weinheim, Germany: Wiley-VCH.
12. Golbraikh, A. & Tropsha, A. (2002). Beware of q2! *Journal of Molecular Graphics & Modelling* **20**, 269–276.
13. Nettleton, D.F., Orriols-Puig, A. & Fornells, A. (2010). A study of the effect of different types of noise on the precision of supervised learning technique. *Artificial Intelligence Review* **33**, 275–306.
14. Hawkins, D.M. (2004). The problem of overfitting. *Journal of Chemical Information & Computer Sciences* **44**, 1–12.
15. Thomas, D.G., Pappu, R.V. & Baker, N.A. (2011). Nanoparticle ontology for cancer nanotechnology research. *Journal of Biomedical Informatics* **441**, 59–74.
16. Shaw, S.Y., Westly, E.C., Pittet, M.J., Subramanian, A., Schreiber, S.L. & Weissleder, R. (2008). Perturbational profiling of nanomaterial biologic activity. *Applied Biological Science* **10,521**, 7387–7392.
17. Fourches, D., Pu, D., Tassa, C., Weissleder, R., Shaw, S.Y., Mumper, R. & Tropsha, A. (2010). Quantitative nanostructure–activity relationship modeling. *ACS Nano* **410**, 5703–5712.
18. Anon. (undated). *SMILES — A Simplified Chemical Language*. Laguna Niguel, CA, USA: Daylight Chemical Information Systems, Inc. Available at: http://daylight.com/dayhtml/doc/theory/theory.smiles.html (Accessed 10.03.14).
19. Johnson, S. & Spence, M.T.Z. (ed.) (2010). *The Molecular Probes® Handbook*. Paisley, UK: Life Technologies. Available at: http://de-de.invitrogen.com/site/de/de/home/references/molecular-probes-the-handbook.html/ (Accessed 05.03.14).
20. RSC (2013). *ChemSpider*. Cambridge, UK: Royal Society of Chemistry. Available at: http://www.chemspider.com/ (Accessed 20.12.13).
21. Mauri, A., Consonni, V., Pavan, M. & Todeschini, R. (2006). DRAGON software: An easy approach to molecular descriptor calculations. *MATCH Communications in Mathematical & in Computer Chemistry* **562**, 237–248.
22. Yap, C.W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, **327**, 1466–1474.
23. Hall, L.H. & Kier, L.B. (1995). Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information. *Journal of Chemical Information & Computer Sciences* **356**, 1039–1045.
24. Hall, L.H., Kier, L.B. & Brown, B.B. (1995). Molecular similarity based on novel atom-type electrotopological state indices. *Journal of Chemical Information & Computer Sciences* **356**, 1074–1080.
25. Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E. & Willighagen, E.L. (2003). The chemistry development kit (CDK): An open-source java library for chemo- and bioinformatics. *Journal of Chemical Information & Computer Sciences*, **432**, 493–500.
26. Steinbeck, C., Hoppe, C., Kuhn, S., Floris, M., Guha, R. & Willighagen, E.L. (2006). Recent developments of the chemistry development kit (CDK): An open-source java library for chemo- and bioinformatics. *Journal of Chemical Information & Computer Sciences* **1217**, 2111–2120.
27. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I.H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations* **11**, 10–18.
28. Ihaka, R. & Gentleman, R. (1996). R: A language for data analysis & graphics. *Journal of Computational & Graphical Statistics* **53**, 299–314.