Perspectives from the NanoSafety Modelling Cluster on the validation criteria for (Q)SAR models used in nanotechnology

Tomasz Puzyn[a,1]*, Nina Jeliazkova[b], Haralambos Sarimveis[c], Richard L. Marchese Robinson[d,1,2], Vladimir Lobaskin[e], Robert Rallo[f], Andrea-N. Richarz[d], Agnieszka Gajewicz[a], Manthos G. Papadopulos[g], Janna Hastings[h,3], Mark T. D. Cronin[d], Emilio Benfenati[i], Alberto Fernandez[j]

[a] *Laboratory of Environmental Chemometrics, Faculty of Chemistry, University of Gdansk, Gdansk, Poland*
[b] *IdeaConsult Ltd., Sofia, Bulgaria*
[c] *Unit of Process Control and Informatics, School of Chemical Engineering, National Technical University of Athens, Athens, Greece*
[d] *School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Byrom Street, Liverpool (UK)*
[e] *School of Physics, University College Dublin, Belfield, Dublin 4, Ireland*
[f] *Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Tarragona, Catalunya, Spain*
[g] *Ethniko Idryma Erevnon, National Hellenic Research Foundation, Athens, Greece*
[h] *European Bioinformatics Institute, Hinxton, Cambridgeshire, UK*
[i] *Istituto di Ricerche Farmacologiche "Mario Negri", Milan, Italy*
[j] *Departament d'Enginyeria Quimica, Universitat Rovira i Virgili, Tarragona, Catalunya, Spain*

1. *These authors contributed equally*
2. *Present address: School of Chemical and Process Engineering, University of Leeds, Leeds (UK)*
3. *Present address: Babraham Institute, Babraham, Cambridge UK*

***Corresponding author**: Tomasz Puzyn; Laboratory of Environmental Chemometrics, Department of Environmental Chemistry and Radiochemistry, Faculty of Chemistry, University of Gdansk, Wita Stwosza 63, 80-308 Gdansk, Poland;
phone: +48 58 523 5248; fax: +48 58 523 5012; e-mail: t.puzyn@qsar.eu.org

Nanotechnology and the production of nanomaterials have been expanding rapidly in recent years. Since many types of engineered nanoparticles are suspected to be toxic to living organisms and to have a negative impact on the environment, the process of designing new nanoparticles and their applications must be accompanied by a thorough exposure risk analysis. (Quantitative) Structure-Activity Relationship ([Q]SAR) modelling creates promising options among the available methods for the risk assessment. These *in silico* models can be used to predict a variety of properties, including the toxicity of newly designed nanoparticles. However, (Q)SAR models must be appropriately validated to ensure the clarity, consistency and reliability of predictions. This paper is a joint initiative from recently completed European research projects focused on developing (Q)SAR methodology for nanomaterials. The aim was to interpret and expand the guidance for the well-known "OECD Principles for the Validation, for Regulatory Purposes, of (Q)SAR Models", with reference to nano-(Q)SAR, and present our opinions on the criteria to be fulfilled for models developed for nanoparticles.

Keywords: nano-QSAR, QNTR, QNAR, QSAR, validation

1. Introduction

Nanotechnology and the production of nanomaterials have been expanding rapidly in recent years. For definitions of nanotechnology, nanomaterials and nanoparticles, readers are directed to the following references: (Lövestam et al., 2010; Rauscher et al., 2013). Since many types of engineered nanoparticles are suspected to be toxic to living organisms and to have a negative impact on the environment, the process of designing new and safe nanoparticles must be accompanied by a thorough risk analysis (Hassellov et al., 2008; Silva et al., 2015).

Computational techniques, especially (Quantitative) Structure-Activity Relationship ([Q]SAR) modelling, provide hazard estimates which are fundamental for risk assessment. The concept underpinning (Q)SAR (Cherkasov et al., 2014; Dearden, 2016) is that, when the structural characteristics (called "descriptors") are known for a group of compounds, and the experimental activity data are available only for a few of them, it is possible to predict the unknown activities for the remaining compounds directly from the descriptors and a suitable mathematical model derived from algorithmic analysis of the available data. Developed models can be either qualitative (SAR) or quantitative (QSAR) in nature. (It should be noted that variations on these definitions are found in the [Q]SAR literature.) A similar approach can be employed to predict various physicochemical properties; such models are commonly known as Quantitative Structure-Property Relationships (QSPRs). Although the development and validation of computational models are both impossible without utilising high-quality experimental data, the application of (Q)SAR/QSPR may help to significantly reduce the time, cost, and the number of required laboratory tests (Gajewicz et al., 2012; Puzyn et al., 2009; Tantra et al., 2015).

Since the terms (Q)SAR/QSPR have traditionally referred to the modelling of the molecular structures of organic compounds, as opposed to nanoparticles, the abbreviation "nano-(Q)SAR" is often used to highlight the specificity of such models to nano-structures (Gajewicz et al., 2012; Puzyn et al., 2009). Alternatively, the terms Quantitative Nanostructure-Activity Relationships (QNAR) (Fourches et al., 2010), Quantitative Nanostructure-Toxicity Relationships (QNTR) (Le et al., 2013) and nano-SAR (Liu et al., 2013b) are also used in the literature.

In order to address the opportunities offered by modelling the toxicity and properties of nanoparticles, the European Commission funded five modelling projects: MODERN (Brehm et al., 2017), Mod-Enp-Tox (Vriens et al., 2017), PreNanoTox, MembraneNanoPart (Lopez et al., 2017) and NanoPUZZLES (Richarz et al., 2017) from 2013-2015 and a further project to assist in the management of data: eNanoMapper (Jeliazkova et al., 2015), from 2014 – 2017, within the EU NanoSafety Cluster (https://www.nanosafetycluster.eu/). Herein, these projects are collectively referred to as the NanoSafety Modelling Cluster, originally comprising the five projects and later joined by eNanoMapper. The projects within the NanoSafety Modelling Cluster concentrated expertise in the development of (Q)SAR models and other chemoinformatic techniques for nanoparticles. This expertise confirms that nano-QSARs/QNARs/QNTRs can be successfully used to predict the physicochemical properties and bioactivity of nanoparticles as well as to assist in the identification of possible mechanisms of toxicity at different levels of biological organisation. For example, the toxicity of metal oxides to the bacterium *Escherichia coli* (a prokaryotic system) was shown to be related to the release of metal cations from the nanoparticle surface, whereas their toxicity to human keratinocyte cells (an eukaryotic system) is mainly related to the redox properties of the surface of the nanoparticle, thus providing an insight into different mechanisms of action (Gajewicz et al., 2015).

The Organisation for Economic Cooperation and Development (OECD) published a set of principles in 2004 (OECD, 2004) for the validation of QSAR models, along with detailed guidance in 2007 describing the application of the so-called "OECD Principles for the Validation, for Regulatory Purposes, of (Q)SAR Models" (OECD, 2007). The Guidance proposes that five criteria, commonly referred to as "the OECD Principles for Validation", should be considered to evaluate a QSAR: (1) a clear definition of the endpoint; (2) the use of an unambiguous algorithm; (3) the necessity of defining the applicability domain; (4) calculating appropriate measures of goodness-of-fit, robustness and predictive ability; and (5), whenever possible, providing a mechanistic interpretation. The appropriate validation of (Q)SARs, according to these principles, is crucial to demonstrate their true predictive ability and limitations in a regulatory context. Whilst these principles may be considered generally applicable to all (Q)SARs, including nano-QSARs, the fact that the development of the original principles, and their accompanying guidance from the OECD, was based on consideration of QSARs for small organic molecules means there is a need to consider whether the guidance associated with their application requires revision and/or addition to be appropriately applied to the evaluation of nano-QSARs. Some of the considerations which are specific to, or otherwise typical for, nano-QSAR studies that prompt us to ask this question are as follows: (1) the challenge of representing nanomaterial structures, which affects which information should be provided to allow nanomaterial descriptor values to be reproduced (c.f. principle 2); (2) nanomaterial-specific data quality considerations (c.f. principle 1); (3) the typically small size of nano-QSAR datasets (c.f. principle 4). We discuss these issues in detail, along with appropriate literature references, in the next section of our article. This paper is a joint initiative of five of the EU research projects (MODERN, PreNanoTox, MembraneNanoPart, NanoPUZZLES and eNanoMapper), which focused on developing (Q)SAR methodology, and supporting resources, for nanomaterials. The aim of the paper is to

reconsider some of the most important aspects of the evaluation process from the perspective of nano-(Q)SAR models in the context of the existing OECD Principles for Validation. However, whilst they are discussed from a nano-(Q)SAR perspective, it should be noted that various suggestions made herein, as will be indicated in the text, are also applicable to QSAR modelling of conventional (small molecule) chemicals.

Here it should be noted that the aim of this paper is not to be a comprehensive review of nano-(Q)SAR models. A variety of recommended reviews of nano-(Q)SAR models have been published in recent years (Burello, 2017; Chen et al., 2017; Oksel et al., 2015a; Oksel et al., 2017; Oksel et al., 2015b; Tantra et al., 2015; Winkler, 2016; Worth et al., 2017). Some articles discuss, to varying degrees of detail, the application of the OECD principles to nano-(Q)SAR models (Kar et al., 2014a; Oksel et al., 2015a; Oksel et al., 2017; Oksel et al., 2015c; Tantra et al., 2015; Toropov and Toropova, 2015). Our current paper discusses the relevant considerations, which need to be taken into account when evaluating nano-(Q)SAR models according to the OECD principles in greater detail than these previous works. More careful consideration is given to relevant issues such as data quality and reproducibility of computational results. Section 2 presents an in-depth discussion of the issues relating to each of the five principles. Section 3 summarizes some of the key questions which need to be answered, along with more detailed considerations which need to be taken into account, when evaluating nano-(Q)SAR models according to these principles and indicates how these key questions may be used for an initial evaluation of nano-(Q)SAR models developed, in part, by members of the NanoSafety Modelling Cluster. Drawing upon expertise across this cluster of modelling projects, section 4 concludes with some take home messages.

2. Discussion on the criteria for validating Nano-QSARs

2.1. A defined endpoint

The endpoint in a (Q)SAR is defined as "a measure of activity for chemicals made under specific conditions" (OECD, 2007) and refers to "any physicochemical property, biological effect or environmental parameter related to chemical structure that can be measured and modelled" (OECD, 2007). The first "OECD Principle for (Q)SAR validation" (OECD, 2004, 2007) states the need of using "a clear endpoint definition". A well-defined endpoint is key to promote clarity regarding exactly what is being predicted by the (Q)SAR and to ensure that the selected endpoint is relevant for the purpose for which the (Q)SAR was developed and used. The OECD guidance proposes a range of specific considerations for the endpoint evaluation, namely whether (1) the scientific purpose, i.e. the modelled endpoint, is clearly defined; (2) the experimental protocol(s) used to generate the underlying data and other "important experimental conditions" are reported as well as the quality of the assays and sources for experimental error; (3) the units (supposing a numerical endpoint) of measurement are clearly defined; (4) the underlying data have been generated using sufficiently consistent experimental procedures; (5) the model has potential to (partially) address a defined regulatory need (i.e., the predicted endpoint has regulatory relevance) (OECD, 2004, 2007). In the following paragraphs, these points are discussed from the specific perspective of evaluating nano-QSAR models.

Regarding the issue of addressing a regulatory need, for the purposes of this paper, this criterion is broadened to have greater general relevance, i.e. fitness-for-purpose for different possible applications of nano-QSARs, such as initial hazard screening in industry. Thus, what remains important is that the endpoint is relevant for the intended application of the model and, in keeping with this, that the exact endpoint being predicted is clear.

Hence, in keeping with consideration (1) defined above, it is important that the terminology used to define the endpoint is sufficiently precise. For example the term "cytotoxicity" (Kar et al., 2014b; Kleandrova et al., 2014; Luan et al., 2014; Pathakoti et al.,

2014; Puzyn et al., 2011; Toropov et al., 2012; Toropova et al., 2014; Toropova et al., 2015) may be considered to cover the output from a wide range of assays and a variety of cellular effects (Hu et al., 2009; Kroll et al., 2009; Lewinski et al., 2008; Oksel et al., 2015b) such as reduced viability (commonly considered cell death) (Lewinski et al., 2008), apoptosis and necrosis (Oksel et al., 2015b) (mechanisms of cell death) (Jin and El-Deiry, 2005), as well as – by some authors (Lewinski et al., 2008) – sub-lethal effects such as oxidative stress and inflammation or even genotoxicity. Other authors distinguish "cytotoxicity" from apoptosis and oxidative stress etc. (Nel, 2013). This illustrates the fact that the endpoint must be sufficiently described to avoid ambiguity in its interpretation. As a general rule of thumb, endpoint categories need to be distinguished in data and models between those that refer to specific measurements, e.g. different assays or specific ways of counting dead cells, and those that refer to groupings of biological relevance which can be measured in a variety of different ways, e.g. cytotoxicity.

Clarity regarding terminology can be promoted by encouraging nano-(Q)SAR modellers to link their endpoint values to definitions from ontologies (Thomas et al., 2011) e.g. the BioAssay Ontology (BAO) (Visser et al., 2011). While the BAO definition of "percent cytotoxicity" (BAO_0000006) (Netzeva et al., 2005) is not itself 100% specific to a single way of defining cytotoxicity – in the text it lists several different assays which measure arguably different aspects of cytotoxicity including counting dead cells vs. different measures of cellular damage – if one browses down the BAO classification hierarchy, one will notice that there are subclasses to represent "percent apoptotic cells" (BAO_0002006) and "percent dead cells" (BAO_0002046) which are defined more specifically. Wherever possible, data should be annotated to the more specific type of endpoint, which in turn is linked to relevant assays, while the ontology hierarchy should capture the interrelationship between the specific and grouped endpoints. Modellers may then choose to combine data from different specific

endpoints in support of predictions of a general endpoint such as 'cytotoxicity', but this should be appropriately documented, as discussed below. Since the process of building ontologies requires that specific definitions are accepted by the scientific community and universally used, the eNanoMapper project worked towards creating an ontology specifically targeting this area – nanomaterial safety – including terminology for relevant assays, descriptors, and endpoints (Hastings et al., 2015), integrating content from BAO and other sources.

One could argue that a nano-(Q)SAR endpoint, similar to those for (Q)SARs for small molecules, is most precisely defined if it is based on the output of a specific assay. This raises the question of whether or not it is appropriate to mix data obtained from different assays for, ostensibly, the same endpoint. Indeed, this touches upon the wider question of how to ensure minimal experimental heterogeneity, as per consideration (4) defined above, whilst having a sufficiently large number of tested chemicals for training and validating the model (Cronin and Schultz, 2003; Dearden et al., 2009; OECD, 2007). It has been previously argued that nano-(Q)SAR models should be developed using data obtained from a single source, supposing data for a sufficiently large number of nanomaterials were available from that source (Lubinski et al., 2013). However, it is recognised that this is not always a realistic expectation. Nanotoxicology data reported in the public domain are commonly generated according to a variety of different tests or, even when the same test (e.g. the Comet assay for genotoxicity assessment) is applied, according to different experimental systems/conditions (Golbamaki et al., 2015; Krug, 2014).

The view that nano-QSARs should only be developed on data measured according to a single protocol under a single set of conditions is arguably reflected in the typically small datasets employed for nano-QSAR development (Oksel et al., 2015b). However, it is established in both the statistical community more broadly (Kaplan et al., 2014) and in

QSAR/QSPR studies in particular (Hawkins, 2004; Palmer and Mitchell, 2014) that, all else being equal, larger datasets support better predictive modelling and more robust validation statistics (see section 2.4 for more detailed discussion). Indeed, an interesting recent QSPR study indicated that mixing up literature data may yield models which are comparably predictive to models developed using data from a single experimental protocol in the same laboratory (Palmer and Mitchell, 2014). This suggests that it should not be assumed that data which are not perfectly experimentally consistent cannot be combined for modelling, albeit caution is still advisable. Even in the case that the inconsistency in experimental data is sufficient to significantly affect the ability to model merged datasets based on descriptors of nanomaterial structure alone, it may still be possible to merge the data via treating the experimental protocol variables as descriptors. Indeed, this approach has been applied in recent nano-QSAR studies (Cassano et al., 2016; Toropov and Toropova, 2015). Of course, in order to assess whether data are likely to be sufficiently inconsistent for their merging to be inadvisable or so that the inconsistency can be captured via treating the experimental variables as descriptors, it is crucial that the key experimental variables are consistently reported across data sources, which can be supported by use of ontologies (Hastings et al., 2015; Marchese Robinson et al., 2015). Indeed, the wider need for minimum reporting standards is currently a key topic in the nanoscience community (Aberg, 2015; Marchese Robinson et al., 2016; Marquardt et al., 2013; Stefaniak et al., 2013).

Regarding the units of measurement to be reported for numerical endpoints, as per consideration (3) defined above, the following question arises. Which dose (or concentration) units are most appropriate if the endpoint value in question is a dose-response or concentration-response statistic, e.g. an LC_{50} (Netzeva et al., 2005), or a lowest observed adverse effect level, i.e. LOAEL (Lewis et al., 2002)? Whilst the most appropriate dose/concentration units may be scenario-specific (Donaldson and Poland, 2013), it is

generally accepted that mass-based concentrations or doses are least appropriate (Cohen et al., 2014; Donaldson and Poland, 2013; OECD, 2012). For small molecule chemicals, it is generally accepted that concentration response data should be expressed in terms of moles per litre and, indeed, some nano-(Q)SAR studies have reported concentration-response endpoints in moles per litre. However, strictly speaking, the notion of "moles" is not applicable for many nanomaterials as they are not based upon a single-molecule species. Rather, particle number or surface area-based concentrations (or doses) are commonly advocated (Cohen et al., 2014; OECD, 2012). However, converting the mass-based concentrations (or doses) to these, more appropriate units, requires appropriate physicochemical measurements (e.g. specific surface area values or density values depending upon the calculation employed) (Cohen et al., 2014; OECD, 2012) to be made.

The experimental endpoint data need to be of sufficiently quality to create sound models, which also raises some nano-specific issues (Marchese Robinson et al., 2016). Indeed, given that data quality may be considered related to the degree to which the data are clearly defined, including in terms of metadata availability, and are free of errors (Marchese Robinson et al., 2016), considerations (1-3), as defined above, are related to the topic of data quality. One particularly key, nanospecific data quality aspect is the potential for artefacts and misinterpretations that may arise with a variety of nano(eco)toxicology assays (Crist et al., 2013; Handy et al., 2012; Kroll et al., 2009; Krug, 2014; Petersen et al., 2014). Some notable potential problems are as follows: (1) the potential for the nanomaterial to interfere with the assay readout such that this readout does not accurately correspond to the nominal endpoint (Kroll et al., 2009; Petersen et al., 2014); (2) contamination with endotoxin or residual solvent (used for sample preparation) triggering toxicity that is wrongly attributed to the nanomaterial (Crist et al., 2013; Krug, 2014; Petersen et al., 2014). Tests to exclude the possibility of assay interference have been proposed (Petersen et al., 2014), along with suggested test

systems/assays that are expected to be free of nanomaterial interference (Kroll et al., 2009; OECD, 2010) and assessment of endotoxin contamination and its potential significance for the test system are recommended (Crist et al., 2013; Petersen et al., 2014). Whether or not these issues were accounted for should be reported.

The precise identification of the test substance is also a requirement for high quality data (Marchese Robinson et al., 2016; Przybylak et al., 2012). In the case of nanomaterials, this typically, but not necessarily (Marchese Robinson et al., 2016), requires that essential key physicochemical properties are reported (Marchese Robinson et al., 2016; Powers et al., 2006; Stefaniak et al., 2013). (Under certain circumstances, suitable identifiers might be sufficient to determine the nanomaterial being tested (Marchese Robinson et al., 2016)). For example, in contrast to small molecule chemicals, nanomaterials are typically polydisperse, hence the size distribution needs to be reported if physicochemical characterisation is used to identify the nanomaterials. N.B. It should be noted that the characterisation of the nanomaterials, in terms of their key physico-chemical properties, may also be considered relevant to definition of the applicability domain (see section 2.3). High variability of measured data depending on the experimental conditions is an issue particularly for nanomaterials, impacting on the characterisation of the materials (Worth et al., 2017).

Another consideration typically argued to be related to data quality is the adherence to a standardised experimental protocol. However, the most appropriate way to adapt the existing standardised test procedures to the specific nanomaterial issues was still under discussion at the time of writing (OECD, 2009, 2013a, 2014a, b). A more detailed discussion of data quality and completeness considerations for nanomaterials was recently published (Marchese Robinson et al., 2016).

When considering all these issues, one must bear in mind that nanoparticle safety assessment is a relatively young discipline; hence the requirements regarding endpoints and

characterisation of data and experimental protocols, which are specific to nanomaterials, are still being discussed in the community. Drawing upon the preceding discussion, we propose that any evaluation of a nano-QSAR according to the OECD requirement for a defined endpoint should entail consideration of the following key questions. Is a precise definition of the endpoint provided (e.g. an ontology annotation)? Are the test methods / assays, along with the key experimental variables, used to generate the endpoint data documented and relevant for nanomaterials? Are the data "reasonably" experimentally consistent? Are units provided for numerical endpoints? Have concentration / dose related units been converted from mass based units (e.g. into surface area based units)? Has the potential for nanomaterial interference with the assays been excluded? Has endotoxin or residual solvent contamination been assessed? Is the endpoint relevant for the intended application of the model? In summary, the most important requirement for providing "a defined endpoint" is the transparent and clear description of all relevant key parameters as discussed above, in order for the potential user of the nano-QSAR model to be able to judge whether the nano-QSAR model can be used for the intended purpose.

2.2 An unambiguous algorithm

The "unambiguous algorithm" principle requires that the complete structure of the model, as well as the exact values of all the model parameters, should be made explicit. This includes adequate description of the mathematical method employed for defining the relationship between the descriptors of the structure and the "activity" endpoint of interest, as well as how descriptors are calculated (or measured). The complete dataset of substances, end-point and descriptor values, together with clearly defined training and test sets should be provided to the user. The remainder of this section is divided into subsections concerned with critical considerations which need to be taken into account when assessing compliance with

the principle of an "unambiguous algorithm": modelling algorithms (section 2.2.1), descriptors (section 2.2.2), variable selection (section 2.2.3) and model reproducibility (section 2.2.4).

2.2.1. Modelling algorithms

A (Q)SAR model is built via applying a statistical or machine learning algorithm to a training set, comprising a matrix of descriptor values with associated endpoint values. Certain algorithmic parameters, sometimes known as "hyperparameters", which determine how the algorithm generates a model from this training set matrix may be tuned using internal validation data. With regard to data mining and machine-learning methods used to derive nano-(Q)SARs, there are no major differences compared to classical (Q)SAR analysis, therefore all the well-established algorithms (including but not limited to multiparametric linear regression, partial least squares, different types of neural network architectures, support vector machines, decision trees etc.) can also be employed (Fourches et al., 2010; Mitchell, 2014). Recently, there has been renewed interest in the QSAR community regarding artificial neural networks, especially in "deep learning" (Baskin et al., 2016). However, given that it has been suggested that "deep learning" is particularly suited for analysis of large datasets (Baskin et al., 2016), whether it is suitable for the small datasets typically considered in nano-QSAR analyses (Oksel et al., 2015b) is perhaps questionable.

2.2.2. Descriptors

The issue of the descriptors to be employed for nano-QSARs and the representation of the nanomaterial structural information from which these are derived represents the major difference between nano-QSARs and most QSARs, which are designed for small molecules (Cherkasov et al., 2014; Dearden, 2016). Although a large number (in the order of thousands) of (Q)SAR molecular descriptors have been defined so far (Consonni and Todeschini, 2010),

they are often inadequate to express the supra-molecular pattern governing the activity and properties of nanomaterials. However, here it should be noted that molecular descriptors may be used to build nano-(Q)SAR models under certain circumstances. For example, they have been successfully applied for the scenario in which all modelled nanoparticles comprise exactly the same core and variation arises from the molecular structure of the small molecule surface modifier, with descriptors calculated solely for this modifier (Kar et al., 2014a). Since nanomaterials based on fullerenes (e.g. C_{60}) and their derivatives are based on well-defined molecular structures, it could be argued that molecular descriptors are adequate for modelling their effects. However, these nanomaterials may exist as agglomerates of fullerene molecules (e.g. "nC_{60}" particles), with characteristics likely to depend upon sample preparation, including the dispersion protocol, rather than being solely determined by the fullerene molecular structures (Astefanei et al., 2015). It should be noted that molecular descriptors, calculated from SMILES representations (Weininger, 1988), have been applied for modelling of a variety of kinds of nanomaterials (Singh and Gupta, 2014), even though the nanomaterials in question cannot truly be represented by a SMILES string.

One of the major challenges in the derivation of nano-(Q)SAR models is the fact that nanomaterials are typically not distinct chemical structures, but rather substances of more complicated structures, which often comprise core chemistries, multiple coatings and linkages between components (encapsulation, entrapments, amide linkages, etc.) of varying size distributions (Thomas et al., 2013). The extension of the "similar compounds have similar properties" principle (Johnson and Maggiora, 1990), which underpins QSAR analyses (Golbraikh et al., 2014; Hansch and Fujita, 1964), to nanostructures is not trivial. Similarity of nanoparticles must accommodate many aspects other than chemical similarity, such as structural similarity including primary size, size distribution, shape, porosity and crystal structure (Burello and Worth, 2011; Gajewicz et al., 2012; Le et al., 2013; Puzyn et al., 2009),

whilst the potential presence of multiple coatings can further modify the material properties and should be taken into account when addressing nanoparticle similarity or defining their identity. Furthermore, the exchange between agglomerated and dispersed forms, as well as the formation of lipid and protein coronas, whose composition can be considered as a biological fingerprint of a nanoparticle, are also factors that define the identity of the nanomaterial for the toxicity assessment purposes (Walkey et al., 2014). Recent reviews discuss the relationship between nanomaterial structures / physicochemical characteristics and their biological effects in greater detail (Bai et al., 2017; Oksel et al., 2017).

By definition, the nanomaterial is presented on a scale of at least one nanometer (Lövestam et al., 2010; Rauscher et al., 2013), which implies that a nanoparticle contains from dozens to billions of atoms. One can mention at least five features that characterize the nanomaterial and are not present at the level of single molecule or pure chemical. Firstly, the processing of nanomaterials often involves surface modification of different chemical nature to improve the dispersion stability – i.e., to prevent nanoparticles aggregation and deposition on the walls of the sample cells or container. Therefore, the molecular structure of the material that interacts with biomolecules is often different to the reported core nanomaterial. Secondly, whilst being transported to the tissue, the nanoparticles may come in contact with biomolecules of the dispersion medium, as the nanoparticles are often dispersed in surfactant or proteins solutions. Therefore, the properties of the core nanomaterial itself may be irrelevant for the (Q)SAR, if the co-solutes are not specified. Thirdly, the structure of a nanomaterial is not fully defined by chemistry alone. The nano-bio interactions may depend on molecular packing (e.g., crystalline or amorphous phase), size and shape etc. Fourthly, even when the main material is defined, a single nanoparticle might contain a macroscopic number of impurities or dopants. Fifthly, nanomaterials are typically not a single entity or even a small set of discrete entities. Rather, they may exhibit “polydispersity” in terms of a

range of properties (Miller and Hobbie, 2013). For example, a nanomaterial may comprise a number of particles with the same chemical composition but distributed across a range of different sizes (Powers et al., 2007).

Thus, a meaningful nano-(Q)SAR should arguably include the information about the surface chemistry, surface charge, crystalline structure, particle size and shape, the delivery vehicle or co-solute, and the dopants/impurities in addition to the chemical content of the particle. Ideally, information about the particle size distribution and, if relevant, other forms of "polydispersity" should also be captured. This means, in practice, that more than one type/system of descriptors is often required for nano-(Q)SAR modelling. Moreover, the descriptors can be divided into two classes: "intrinsic" and "extrinsic" properties. The first class contains properties that are independent on the external conditions (e.g., on pH, the presence of proteins), whereas the second class describes changes in the structure dependent on the changing environment (Mikolajczyk et al., 2015). Taken together, these parameters make an analogue of a pure chemical in a standard molecular (Q)SAR. By this we mean that the combined set of "intrinsic" and "extrinsic" properties give rise to biological effects and, if suitably characterised using descriptors, can serve as the basis of a nano-QSAR.

Given this structural complexity, it follows that representation formats for nanomaterials are definitely more complicated than those of single-molecule chemicals, and consequently this subject deserves separate and careful assessment. One such proposed format is the ISA-TAB-Nano Material file and its associated data files (Marchese Robinson et al., 2015; Thomas et al., 2013). The format used to represent the nanomaterial, even before starting any descriptor calculation, may have huge impact on the final model and therefore should be carefully specified.

Novel and more appropriate types of nano-descriptors are being developed, such as features derived from quasi-SMILES, encoding physicochemical properties and/or

experimental conditions using string labels (Cassano et al., 2016; Toropov et al., 2015; Toropov and Toropova, 2015; Toropova et al., 2011; Toropova et al., 2015), descriptors derived from molecular graph or the graph of atomic orbitals theory (Puzyn et al., 2009) and features derived from quantum-mechanical (QM) calculations (e.g. electron distribution, ionisation potential, electron affinity, surface reactivity, band gap, electronegativity and enthalpy of formation) (Gajewicz et al., 2015; Puzyn et al., 2011). Mechanistic justification for these QM descriptors is provided in the cited references (Gajewicz et al., 2015; Puzyn et al., 2011). Other relevant descriptors that cover the biological activity of nanomaterials, such as protein adsorption energies, can be calculated using coarse-grain modelling (Lopez and Lobaskin, 2015).

The exact method and the software used to derive these descriptors should be stated explicitly. For example, MOPAC (http://openmopac.net/) is a popular semi-empirical quantum chemistry program for deriving QM descriptors. However, in order to reproduce the results, the complete crystal structure and the optimised geometry of the nanoparticle, or representative cluster built from the bulk crystal structure, must be available (Gajewicz et al., 2015). Additionally, the exact QM method used to perform the calculations should be provided, since different methods (e.g., PM3, PM6, PM7) may lead to substantially different results. Furthermore, semi-empirical models, whilst allowing for descriptors to be calculated relatively quickly, may lead to serious errors, in particular if the considered derivatives include metals. (Nonetheless, nano-QSAR investigations suggest descriptors derived from semi-empirical calculations can still yield reasonable models (Gajewicz et al., 2015; Puzyn et al., 2011)). Thus, it is advisable to validate the semi-empirical results by using rigorous *ab initio* methods (e.g. CCSDT(2), CASSCF/CASPT2) and reasonably extended basis sets (Coe et al., 2013; Serrano-Andres et al., 2009).

Experimentally derived properties also serve as suitable descriptors for developing nano-(Q)SARs, as they are especially useful for expressing size distribution, agglomeration state, shape, porosity and irregularity of the surface. A potentially useful source for deriving structural information of nanoparticles is images taken by scanning microscopy (SEM), transmission electron microscopy (TEM), or atomic force microscopy, where descriptors are calculated using image analysis techniques (Gajewicz et al., 2015). Other experimental data which may provide descriptors for nano-QSAR include zeta potential measurements (Cassano et al., 2016). To ensure consistency with the requirements of the second OECD Principle, the experimental protocol, the laboratory conditions and the algorithms used to calculate those descriptors should be provided in detail when experimentally derived descriptors are used in nano-(Q)SARs. This is demonstrated in the following examples:

- The size of nanomaterials can be determined by different experimental methods. TEM typically measures the primary size of the nanoparticles (Murdock et al., 2008), but Dynamic Light Scattering (DLS) measures the "hydrodynamic diameter" which is an effective diameter calculated based on the assumption that the particles are spherical (Baalousha and Lead, 2012; DLA, 2011). Indeed, the size measured via DLS may also reflect the presence of aggregates/agglomerates (DLA, 2011) hence may be significantly larger than primary size values.
- Zeta potential is an important characteristic of the nanoparticle surface that determines the long-term stability of the nanoparticle dispersions. Zeta potential measurement techniques (e.g., light scattering and agarose gel electrophoresis) allow one to evaluate the surface charge of nanoparticles in the medium. The exact conditions under which zeta potential is measured should be available, as various factors (notably medium composition, sample pH, serum incubation) can significantly affect the measured values (Cho et al., 2012; Walkey et al., 2014; Worth et al., 2017).

- The composition of the protein corona may be modified when the nanoparticles move from one compartment to the other. The biological fluid, in which the corona composition is measured, should be exactly known (Le et al., 2013).

Interestingly, experimentally-derived descriptors can be also used for calculating new types of theoretical descriptors, such as descriptors based on the "Liquid-Drop Model" (LDM descriptors) (Sizochenko et al., 2014). The model can be used in order to encode nanoparticle aggregates in a solution. In LDM, the aggregate is represented as a spherical drop, where elementary nanoparticles are densely packed and the density of the aggregate is equal to the mass density. Then, the five following descriptors can be easily calculated: Wigner-Seitz radius (the minimum radius of the interactions between the elementary nanoparticles), the number of nanoparticles in the aggregate, the number of nanoparticles on the surface, the surface-to-volume ratio and the aggregation parameter. It is worth mentioning that the same scheme can be used to describe volume-related features of single nanoparticles. In such a case, a nanoparticle is treated as a cluster ("aggregate") of atoms (i.e. LDM represents a single nanoparticle built up from atoms as the basic elements) (Sizochenko et al., 2014).

2.2.3. Variable selection

In order to select the descriptors employed for building a nano-QSAR model, a combination of mechanistic expertise and statistical variable selection may be employed, with the latter commonly referred to as "feature selection". Feature selection may be appropriate for two key reasons: (1) reducing the number of descriptors *might* avoid overfitting the training data, a scenario in which the model parameters are adjusted to predict the training set endpoint values well at the expense of true predictivity on data not seen during training; (2) a smaller number of descriptors makes expert assessment of the mechanistic basis for the model (see section 2.5) more manageable (Cherkasov et al., 2014).

A commonly quoted heuristic in the QSAR literature, attributed to the work of Topliss and Costello (Topliss and Costello, 1972), is that the ratio of compounds : descriptors should be greater than or equal to 5:1, at least when using simple linear regression methods (see section 2.2.1) (Cherkasov et al., 2014; Dearden et al., 2009). In the case of the typically small datasets used in nano-QSAR studies (Oksel et al., 2015b), sometimes of the order of 20 instances or less, this would imply very few descriptors should be considered. Indeed, according to the analysis of Topliss and co-workers, the probability of finding chance correlations in traditional multiple linear regression models is related to the total number of descriptors evaluated, via statistical procedures, rather than the final number included in the model (Topliss and Costello, 1972; Topliss and Edwards, 1979), so it cannot reasonably be claimed that a nano-QSAR model is compliant with the so-called Topliss and Costello rule if the ratio is reached via statistically evaluating a larger set of descriptors for inclusion in the model. However, it should be noted that the seminal work of Topliss and co-workers (Topliss and Costello, 1972; Topliss and Edwards, 1979) was based on specific variable selection approaches in the context of traditional multiple linear regression analysis and it cannot be assumed that their findings are necessarily applicable to all other techniques. Furthermore, these authors did not actually propose a hard and fast rule. Hence, rather than condemn a nano-QSAR model as worthless if it violates the so-called Topliss and Costello rule, it is more appropriate to be judicious regarding the descriptors considered for evaluation on mechanistic grounds, where possible, and employ rigorous statistical techniques, e.g. ‘external cross-validation’ (Hawkins, 2004; Low et al., 2011) and y-scrambling (Rucker et al., 2007) discussed in section 2.4, in combination with assessment of the mechanistic plausibility of obtained correlations (see section 2.5) to ascertain whether a nano-QSAR model really is based on chance correlations and cannot be trusted to make reliable predictions for unseen data.

A variety of approaches for statistical feature selection exist (Ferreira and Figueiredo, 2012; Guyon and Elisseeff, 2013). A number of these are based on the so-called "relevance and redundancy" criteria (Ferreira and Figueiredo, 2012) that, in the terminology of QSAR, a good set of descriptors should be well correlated with the modelled endpoint and poorly correlated with each other. However, different measures of variable association might give different results and descriptors which appear highly correlated according to some measures may not be truly redundant (Guyon and Elisseeff, 2013). Statistical feature selection remains an active area of research.

If feature selection algorithms are applied, both the original and the reduced sets of data should be made available to the user, including the rules and mathematical formulae used to select, prioritise or cluster the data. Moreover, values of the steering parameters for the algorithms (if any) should be provided in detail.

Feature ranking methods, which rank descriptors according to their relevance to modelling the endpoint as may be estimated using a variety of techniques (Ferreira and Figueiredo, 2012), might be used for the mechanistic interpretation of models. For example, a mechanistically comprehensible descriptor found to show high (positive) association with an endpoint could yield insights into the structural factors driving the modelled endpoint.

2.2.4. Model reproducibility

Reproducibility of the models, including easy transfer and exchange across different platforms, is an important issue in QSAR modelling (Tetko et al., 2017), which also applies to the case of nano-QSAR analysis (Helma et al., 2017). Indeed, the reproducibility of computational science more generally has been a key concern in the recent scientific literature (Editorial, 2014).

A variety of formats exist for documenting QSAR models, for the purpose of creating model-specific records in online or in-house repositories, in order to facilitate their reuse.

These formats and repositories were recently reviewed (Tetko et al., 2017). One such format, established over a decade ago by the European Commission's Joint Research Centre (JRC), is the (Q)SAR Model Reporting Format (QMRF), designed to document QSAR models in keeping with the OECD principles [(Pavan and Worth, 2008); http://qsardb.jrc.it/qmrf]. In a recent project by JRC it has been used in a format extended for nanomaterials to compile an inventory of currently existing nano-QSARs/QSPRs (Worth et al., 2017). The eNanoMapper project developed a resource for generating QMRF reports for nano-QSAR models (Drakakis et al., 2016). However, this format only supports documentation of human readable, free text descriptions of the relevant details required to reproduce the models, along with encouraging links to relevant software and structural files. In practice, this might not be sufficient to (easily) reproduce the model. The QsarDB format [(Ruusmann et al., 2014, 2015; Tetko et al., 2017); http://qsardb.org], in contrast, seeks to document the QSAR models in a machine-readable fashion, designed to be readily parsed by appropriate software tools to reproduce the models. Notably, the QsarDB format uses the Predictive Model Markup Language [(Tetko et al., 2017); http://dmg.org/pmml/v4-3/GeneralStructure.html], where applicable, to represent the structure of the model itself in a software independent machine-readable fashion. This is complemented with information about the descriptors, endpoint values and chemicals in the training and test sets. For example, the nano-QSAR model of Puzyn and co-workers for inorganic nanoparticle cytotoxicity (Puzyn et al., 2011) has been recorded within the QsarDB database (Piir, 2014), with the structure of the final derived regression model documented using PMML.

However, this example (Piir, 2014) also highlights potential challenges for representing nano-QSAR models in this fashion. It is implied that the nanoparticles for which descriptors are calculated can be represented using a CAS number, which is clearly inadequate for describing nanomaterials, as opposed to small molecule chemicals. Future

work should consider whether the existing model reporting formats can be supplemented with links to suitable representations of nanomaterial structures, e.g. based on the Material file of ISA-TAB-Nano and linked data files (Burello, 2017; Marchese Robinson et al., 2015; Thomas et al., 2013), to better support the reproducibility and re-use of nano-QSAR models.

The JaqPot Quattro (JQ) nanomaterial web modelling platform (Chomenidis et al., 2017) , developed in the context of the eNanoMapper project, is linked to the eNanoMapper database (Jeliazkova et al., 2015), which integrates diverse and heterogeneous information to adequately represent complex nanomaterial structures. Among other functionalities, JQ facilitates the automatic creation of reproducible nanoQSAR models in the form of ready-to-use web resources, that can be accessed either through the system API or through a user-friendly interface. Additionally, JQ creates PMML representations of the produced nanoQSAR models. Model validations and end-point predictions are performed either by feeding data from the eNanoMapper database or by manually entering the descriptor values. For example, the model developed by Gajewicz co-workers for predicting $\log(1/LC_{50})$ toxicity to the human keratinocyte cell line (HaCaT) cell line (Gajewicz et al., 2015) has been generated using the JQ functionalities and is offered to the community as a web resource and application (http://www.jaqpot.org/m_detail?name=gaj-18-linear).

2.3. A defined domain of applicability

The OECD guidance defines the "applicability domain" as follows (OECD, 2007):

"The applicability domain of a (Q)SAR model is the response and chemical structure space in which the model makes predictions with a given reliability". In this definition, chemical structure can be expressed by physicochemical and/or fragmental information, and response can be any physicochemical, biological or environmental effect that is being predicted".

Thus, the ultimate purpose of defining the "domain" is to ensure the model is not used to make predictions for chemicals for which the predictions are expected to be unacceptably unreliable (Gadaleta et al., 2016; Hanser et al., 2016). However, what this is interpreted to mean in practice varies considerably and there are a plethora of different approaches to characterising the applicability domain (Hanser et al., 2016). Since QSAR models are obtained using statistical algorithms based on the observed trends within the training set, there is no reason to expect the relationships they capture to be applicable to regions of "chemical space" lying outside the training set chemicals. Hence, various approaches to defining the applicability domain are based on ensuring the model does not extrapolate outside the chemical space of the training set in terms of relevant descriptors (Hanser et al., 2016). This might be quantified in terms of a "leverage" threshold, as employed for applicability domain characterization for one of the nano-QSAR models of Puzyn and co-workers (Puzyn et al., 2011). However, a variety of different kinds of information might be considered when evaluating the applicability domain, including direct estimations of the uncertainty in individual predictions, coupled with a suitable threshold for rejecting a prediction as being unacceptably unreliable (Hanser et al., 2016). It should be understood that these seemingly disparate approaches are all concerned with the central question underpinning the concept of an applicability domain: is the prediction returned for this new chemical (or, in the context of nano-QSARs, nanomaterial) of acceptable reliability, or should it be rejected? Nonetheless, it should be noted that Hanser et al. (Hanser et al., 2016) have proposed breaking down this assessment into a multi-step decision framework, based upon consideration of different kinds of information at each stage: firstly, decide whether the model can be applied at all to make a prediction for the current use case (applicability assessment); secondly, decide whether the prediction is reliable enough for the current use case (reliability assessment); finally, decide whether a clear decision can be made based on the prediction (decidability assessment).

The recent QSAR literature has increasingly focused on approaches which seek to directly estimate the reliability of predictions for new chemicals (Dragos et al., 2009; Hanser et al., 2016; Lindh et al., 2017; Sushko et al., 2010), investigating a variety of measures reflecting the predictive error: standard deviation calculated from the ensemble of models (Tetko et al., 2008), bagged variance (directly available from Random forest models), estimating the error model of a machine learning algorithm via another machine learning algorithm (Sheridan, 2013), Kullback-Leibler (K-L) divergence probability distributions (Wood et al., 2013), local estimates of error (Clark, 2009; Sahlin et al., 2014; Sheridan, 2013) or confidence (Helma, 2006), sensitivity analysis and most recently, conformal prediction (Lindh et al., 2017; Norinder et al., 2014). An overview of the importance of and methods to characterize uncertainty is provided in a series of recently published contributions (Dragos et al., 2009; Hanser et al., 2016; Iqbal et al., 2013; Lindh et al., 2017; Sahlin, 2014; Sahlin et al., 2013; Sushko et al., 2010). The current trend of moving away from similarity as an assumed measure of applicability domain and predictivity (Sheridan et al., 2004) through adding additional metrics (Keefer et al., 2013; Sheridan, 2012) and finally realising that similarity is redundant (Sheridan, 2013), reflects a view that the essential need is not delineating the "domain", but being able to estimate the uncertainty of predictions.

In keeping with the preceding comments regarding the potential need to take account of a variety of different kinds of information when assessing the applicability domain, another key question to ask is the following one: is the compound in the applicability domain of the model in terms of its mechanistic and metabolic profile? This shifts the weight of the "domain" problem from purely statistical analysis to the mechanistic / metabolic definition of applicability domain. These questions are also directly related to the mechanistic interpretation of the model; the modeller assumes that a given combination of descriptors has a concrete meaning in the context of the studied toxicity mechanism. Thus, the assumed

mechanism, or set of mechanisms in the case of a non-linear modelling technique which is able to model toxicity data arising from multiple mechanisms, should be the same for training nanoparticles and nanoparticles for which the predictions are made. Regarding whether the metabolic profile, which is linked in turn to the mechanism of action, for a new nanomaterial reflects the metabolic profile of the nanomaterials in the training set, it should be acknowledged that this information may not be available in practice, yet may become increasingly available in the future (Lv et al., 2015).

Current (Q)SARs for nanomaterials mostly use descriptor-based approaches to applicability domain estimation (Gajewicz et al., 2012; Gajewicz et al., 2015; Liu et al., 2013b; Puzyn et al., 2009; Puzyn et al., 2011). However, we can make the following key recommendations. Firstly, we recommend statistical approaches to directly estimating the prediction reliability which were discussed above (e.g. a conformal prediction framework (Lindh et al., 2017)). Secondly, we make the following key recommendations which are specific to nano-QSAR modelling:

- If a nano-(Q)SAR model is specifically developed for a certain type of nanostructure (for example fullerenes, carbon nanotubes, metal oxides etc.), the AD of the model may be considered limited to this type of nanomaterial.
- A single nanostructure is often composed of multiple different components (e.g. core and coating). Theoretical features or molecular fragments can be obtained in this case for all different components and used as descriptors in a nano-(Q)SAR model. In the process of examining the similarity of a query substance to the training set, comparisons should be made among equivalent components, i.e. the core of the new substance should be compared against core components in the training set and the same should happen for the coatings, taking into account the order of coatings in the formation of the nanomaterials.

- When experimental measurements are used as descriptors, the experimental protocols and conditions for measuring the descriptors in the query nanostructure and the training set should be sufficiently consistent.
- Data on the regulation of genes, production of metabolites, interactions with all relevant proteins, such as information on the protein corona (i.e., proteins that are adsorbed onto the surface of nanoparticles), etc. (collectively known as "omics" data) can be used as quantitative experimental descriptors. Furthermore, they can uncover fundamental mechanisms of nano-bio interactions and be used for defining the mechanistic domain of the model (Walkey et al., 2014). Again, experimental procedures and biological fluids for obtaining "omics" data should not differ significantly between query substances and the training set.

Compared to classical (Q)SAR modelling, it seems more important to achieve a balance between the level of confidence in the predictions of a nano-(Q)SAR and its scope. For example, should a substance with similar core, but dissimilar coatings to those present in the training set (or *vice versa*) automatically be excluded from the applicability domain? Is an *in silico* prediction totally unreliable if experimental structural descriptors (such as size or shape) are measured under different experimental protocols? (A cautious response would to be say, "yes", however it is not *necessarily* the case that the differences in, say, size measured via two slightly different protocols are sufficiently large to significantly affect the model predictions. Hence, this question should ideally be evaluated empirically for different scenarios.) The challenge for the nano-(Q)SAR research community is to develop new methods for defining the applicability of a model, taking into account the above considerations, in order to broaden the applicability domain without compromising the reliability of predictions.

2.4. Appropriate measures of goodness-of-fit, robustness and predictivity for nanomaterial models

The fourth OECD Validation Principle (OECD, 2007) expresses the need to perform statistical analysis to establish the performance of a model, which consists of an internal validation process (i.e., measures of *goodness-of-fit* and *robustness*) followed by the external validation (i.e., measures of *predictivity*). The statistical validation techniques described in this subsection should be considered in combination with any knowledge about the applicability domain of the model, since the choice of nanomaterials during model development and validation strongly affects the assessment of performance.

Predictions for nanomaterials in the training set, made using the final model derived from the whole training set, can be used to assess the *goodness-of-fit* of the model, which is a measure of how well the model accounts for the variance of the response in the training set and, most importantly, whether the model is statistically significant. This type of error estimate is known as the *apparent error rate*. It will generally suffer from substantial optimistic bias, since many algorithms can fit a given training set perfectly or near perfectly, and thus yield an apparent error of zero or near zero. A model that is not statistically significant, or that is significant but of poor fit, cannot be expected to be useful for predictive purposes.

The *robustness* of a model refers to the stability of its parameters and consequently the stability of its predictions when a perturbation (e.g. deletion of one or more nanomaterials) is applied to the training set, and the model is regenerated from the perturbed training set. If the model is not robust to small perturbations in the training set, it is unlikely to be useful for predictive purposes.

Predictions for nanomaterials in the test set(s) are used to assess the *predictivity* (predictive ability, predictive capacity, or predictive power) of a model, which is a measure of

how well the model can predict new data not used during model development. N.B. Certain training set resampling techniques, discussed in section 2.4.1, may also be considered as internal validation, albeit if applied with care may avoid optimistically biased performance estimates, e.g. so-called 'external cross-validation' (Hawkins, 2004; Low et al., 2011)

In order for a statistical model to be useful for predictive purposes, it should be developed from a sufficiently large and representative amount of information regarding the biological activity and should only contain relevant variables. A variety of statistical validation techniques are available for assessing the goodness-of-fit, robustness and predictivity of models, and a variety of statistics are routinely used to express these aspects of model performance. However, information related to nanosafety is usually found in the form of small (i.e. limited number of samples) datasets (Oksel et al., 2015b). It is commonly known that the smaller the number of samples available in the training/test data sets, the less reliable the error rate estimate will be.

2.4.1. Statistical validation techniques

A number of statistical validation techniques can be used to assess the predictive ability of a QSAR model based upon resampling or permuting the available data used to build the final model. (Resampling techniques entail rebuilding the model on subsets of the data and evaluating on the remaining data, as explained below for different techniques.) In principle, these techniques are equally applicable to nano-QSAR models although the typically small size (Oksel et al., 2015b) of nano-QSAR datasets makes certain considerations particularly pertinent, as will be explained below. The most popular ones are described next; more extensive description of the statistics and validation techniques can be found elsewhere (Alexander et al., 2015; EChA, 2008; Worth et al., 2005).

The *training/test set splitting* is a validation technique based on the splitting of the dataset into a *single* training set and a test set, unlike the resampling techniques described

below. The model is derived from the training set and the predictive power is estimated by applying the model to the test set. The splitting is performed by either randomly or systematically (e.g., taking every third compound from the set sorted according to the descending endpoint value) selecting the instances (i.e. nanomaterials in the current context) belonging to the two sets.

Cross-validation (CV) is the most common resampling validation technique, where a number of modified datasets are created by deleting one or a small group of nanomaterials from the data in such a way that each nanomaterial is removed away once and only once. From the original dataset, a reduced dataset (i.e., training set) is used to develop a partial model, while the remaining data (i.e., test set) are used to evaluate model predictivity (Efron, 1983). The simplest cross-validation procedure is the leave-one-out (LOO) technique, where each nanomaterial is removed, one at a time. Using this method, given *n* nanomaterials, *n* reduced models are calculated, each of these models is developed with the remaining *n*-1 nanomaterials and used to predict the response of the deleted nanomaterial. The averaged (or pooled, i.e. aggregated across all *n* predictions) performance of the *n* reduced models is reported as the LOO cross-validated performance. Other cross-validation procedures, such as the leave-many-out (LMO) technique, try to introduce a larger perturbation in the dataset by removing more than one nanomaterial at each step. In K-fold cross-validation, the training set is divided into K sets of approximately equal size, each one being used in turn as a test set for evaluation of models build on the remaining data.

Bootstrap resampling or *bootstrapping* is another technique to perform statistical validation (Braga-Neto and Dougherty, 2004; Wehrens et al., 2000). In a typical bootstrap validation, *k* groups of size *n* are generated by a repeated random selection, with replacement, of *n* nanoparticles from the original dataset. For any one of the *k* groups, some of these nanoparticles can be included in the same random sample several times, while other

nanoparticles will never be selected. Each bootstrap sample is treated as a training set, used to build the model, whilst the corresponding test set is the set of nanoparticles not selected for the bootstrap sample. This procedure of building training sets and test sets is repeated many times to obtain significant performance statistics. It should be noted that different variations, meaning different ways of estimating predictive performance, of the basic premise exist (Braga-Neto and Dougherty, 2004).

Regarding which resampling techniques, whether a variant of cross-validation or bootstrap resampling, are most appropriate for evaluation of the typically small (Oksel et al., 2015b) datasets used in nano-QSAR studies, the discussed techniques may be considered to have various strengths and weaknesses. It has been argued that, for small datasets, cross-validation is preferable to using a single partition into training and test sets, as the cross-validation estimate of model performance is likely to be more reliable (Hawkins, 2004; Hawkins et al., 2003). It has further been argued that cross-validation may even underestimate predictive performance (Hawkins et al., 2003). Regarding the suitability of different cross-validation schemes for small datasets, cross-validation schemes, especially the LOO approach, have been suggested to suffer from considerable variance and can yield outliers which are highly misleading as to the true predictive performance (Braga-Neto and Dougherty, 2004). (However, it has elsewhere been suggested that LOO estimators are preferrable for small datasets (Hawkins, 2004)). Bootstrap estimators are suggested to be more robust, albeit to produce more biased results (Braga-Neto and Dougherty, 2004). Finally, it should be noted that heuristic guides to the minimum size of the datasets required for reliable modelling results, especially for nanomaterials, have been proposed elsewhere (Lubinski et al., 2013).

Y-scrambling or *response permutation testing* is another widely used technique to check the robustness of a model, and to identify models based on chance correlation, *i.e.* models where the independent variables (descriptors in this context) are correlated to the

response variable (the endpoint in this context) by chance, within the available data. The test is performed by calculating the quality of the model obtained after randomly modifying the sequence of the response vector (e.g. the vector of measured biological activities for each nanomaterial), *i.e.* by assigning to each nanomaterial a response randomly selected from the true set of responses (Lindgren et al., 1996). If the original model has no chance correlation, there is a significant difference in the quality of the original model and that associated with a model obtained with random responses. The procedure is repeated many times. In order for Y-scrambling to be valid, the 'entire' modelling protocol (including statistical selection of descriptors and algorithmic parameters, if applicable) should be repeated each time (Lindgren et al., 1996; Rucker et al., 2007)

Finally, it should be noted that the techniques described in this section (2.4.1) are also advocated as means of assessing model robustness (Chirico and Gramatica, 2011, 2012; Golbraikh and Tropsha, 2002; Gramatica, 2007, 2013).

2.4.2. Statistics for assessing goodness-of-fit and predictivity of regression models

Regression models are mathematical models, linear or non-linear, that attempt to numerically explain the observed values of a (biological activity) endpoint variable in terms of several independent or predictor variables.

To assess goodness-of-fit, the coefficient of multiple determination, R^2, is calculated for the training set. R^2 estimates the proportion of the variation of the endpoint variable that is explained by the model. However, the value of R^2, for the training set, can generally be increased by adding additional predictor variables to the model, even if the added variable does not contribute to reduce the unexplained variance of the dependent variable. This can be avoided by using other statistical parameters such as the adjusted R^2 adjusted for degrees of freedom, R^2_{adj}, or the explained variance in prediction, Q^2. In contrast to R^2 for the training set,

the values of R^2_{adj} and Q^2 do not increase if an added variable to the equation does not reduce the unexplained variance (Massart et al., 1997).

The strengths and weaknesses of R^2 as a measure of goodness-of-fit and of assessing predictive power for QSAR models were recently discussed (Alexander et al., 2015). It should be noted that, when applied as a measure of true predictive performance on an external test set (section 2.4.4), there is no need to adjust the statistic according to the number of descriptors used.

From the calculated and observed dependent variable values, the standard error of estimate, *s*, can be obtained. The standard error of estimate measures the dispersion of the observed values, and a smaller value of *s* may indicate a higher reliability of the prediction. However, a standard error of estimate smaller than the experimental error of the biological data is an indication of an overfitted model (Wold et al., 1984).

Other recent articles also present detailed evaluations of a range of statistics for characterising the predictive power of QSAR models (Gramatica and Sangion, 2016; Roy et al., 2017; Roy et al., 2016; Todeschini et al., 2016). Likewise, it has elsewhere been suggested that an analysis of residuals should be performed, with a view to detecting bias (systematic error) in regression predictions (Roy et al., 2017).

For the avoidance of doubt, it should be reiterated that whether these statistics represent unbiased estimates of the predictive power of a model is related to the set of instances (i.e. nanomaterials in the current context) for which they are calculated. If they are estimated by comparing the predictions of the model to the training set endpoint values, they will be optimistically biased. If they are used to compare endpoint values to predictions obtained from suitable resampling schemes (section 2.4.1) or, ideally, robust external validation (section 2.4.4), they may more reasonably be considered to quantify the predictive performance of a model.

2.4.3. Statistics for assessing goodness-of-fit and predictivity of classification models

Classification models assign nanoparticles into two or more pre-defined categories. In a classification model, the results of the classification can be arranged in the so-called confusion or contingency matrix, where the rows represent the experimentally determined classes, while the columns represent the predicted classes assigned by the model. The goodness-of-fit of a classification model can be assessed in terms of its Cooper statistics (Cooper et al., 1979).

The classification ability of a classification model depends on the particular dataset of nanomaterials used, especially if it is a small one. It is therefore useful to report some measure of the variability associated with the classifications, which indicates whether the classification performance of the model would vary significantly if it had been assessed with a different set of nanomaterials. To estimate confidence intervals for the Cooper statistics, the bootstrap re-sampling technique can be used (Braga-Neto and Dougherty, 2004; Wehrens et al., 2000; Worth and Cronin, 2001). To compare the performances of a number of classification models, the Receiver Operating Characteristic (ROC) curve can be used (Lusted, 1971). In the ROC graph, the X-axis is 1-specificity (false positive rate) and the Y-axis is the sensitivity (true positive rate). An index of the performance of a classification model is the area under the curve. The predictive performance of a classification model can be evaluated by the proportion of misclassifications, e.g. as estimated with the leave-one-out method or with confidence intervals for a binomial proportion (Newcombe, 1998; Ross, 2003). However, estimating the overall percentage of misclassifications can be a poor measure of model performance in the case that the considered data are drawn disproportionately from one class (Baldi et al., 2000; Gorodkin, 2004).

Reporting the observed error rate without estimated confidence intervals is of limited value. Given the study of a particular population, the true error rate constitutes an inherent

property of the model. The observed error rate is only an estimate and depends largely on the adopted sampling strategy. Confidence intervals for the statistic of interest are of particular importance in scenarios comprising small data sets such as nanomaterials ones (Berrar et al., 2006a, b; Oksel et al., 2015b).

The same comments, made at the end of section 2.4.2, regarding whether the calculated statistics for regression models quantify goodness-of-fit or predictivity also apply for the statistics calculated for classification models.

2.4.4. Evaluating the predictivity of models

One of the most important characteristics of a QSAR model is its predictive power, *i.e.* the ability, in the current context, of the model to predict accurately the (biological) activity of nanomaterials that were not used for model development. The resampling techniques discussed in section 2.4.1 are commonly described as internal validation techniques that cannot directly estimate the true predictive ability of a model on unseen data (Chirico and Gramatica, 2011, 2012; Golbraikh and Tropsha, 2002; Gramatica, 2007, 2013).However, it should be noted that some so-called internal validation procedures, such as cross-validation, need not necessarily yield optimistically biased estimates of model performance, if none of the selection of model parameters and descriptors is based on the nanoparticles used to assess model performance (Hawkins, 2004). Cross-validation carried out in this fashion may be termed 'external cross-validation' (Low et al., 2011).

Nonetheless, true external validation (Chirico and Gramatica, 2011, 2012; Gramatica, 2007, 2013; Tropsha, 2010; Tropsha et al., 2003) of the final model on data not seen during model development should be the ultimate aim. In principle, the predictivity of a model should be estimated by comparing the predicted and observed values/classes of a sufficiently large and representative external test set of nanomaterials that were not used in the development of the model. However, external validation can be difficult to assess in a

meaningful way when data of sufficient quality are scarce. For this reason, a common practice is to split the available dataset into a training set, used to develop the model, and an external test set, used to assess the predictive capability of the model. (Indeed, if data are only available for a very small number of nanoparticles, the use of cross-validation protocols, suitably adapted to avoid optimistic bias, may yield a more reliable estimate of predictive power than splitting into a single training and test set (Hawkins, 2004).) One view is that an ideal splitting leads to a test set such that each of its members is close to at least one member of the training set (Tropsha et al., 2003), although it can be argued that this leads to an unrealistic test set which overestimates the performance of the model as a predictive tool. Approaches for the selection of training and test sets range from the straightforward random selection (Yasri and Hartsough, 2001), through activity sampling and various systematic clustering techniques (Potter and Matter, 1998; Taylor, 1995), to the methods of self-organising maps (Gasteiger and Zupan, 1993), Kennard and Stone algorithm (Kennard and Stone, 1969), formal statistical experimental design (Eriksson and Johansson, 1996), and the modified sphere exclusion algorithm (Golbraikh et al., 2003). However, it is recommended that, whenever possible, the external predictivity of the model should be assessed based on a test set drawn from the studied general population (domain) of nanoparticles independently from sampling the training set nanoparticles. This strategy results in more realistic assessment of the external predictive ability of the (Q)SAR model than just a single sampling and then splitting the samples into the training and test sets (Esbensen and Geladi, 2010).

To conclude this section (2.4), the above statistical approaches originally developed for “classic” QSARs are equally appropriate for nano-QSARs; there is no need to invent specific statistical measures of goodness-of-fit, robustness and predictivity. However, the typically small size of nano-QSAR datasets means certain considerations related to estimating model performance from data resampling are particularly pertinent, as discussed above.

2.5. Mechanistic interpretation, if possible

According to the fifth OECD Principle (OECD, 2007), a (Q)SAR should be associated with a "mechanistic interpretation", wherever such an interpretation can be made. The purpose of this was to allow for interpretation of QSAR in a manner that would increase confidence and reduce the possibility of models based on chance correlations. Obviously, it is not always possible to provide a mechanistic interpretation of a given (Q)SAR. The intent of this principle is therefore to ensure that there is an assessment of the mechanistic associations between the descriptors used in a model, if any, and the endpoint being predicted, and that any assessment is documented. Where a mechanistic interpretation is possible, it can also form part of the defined applicability domain (section 2.3).

Consequently, a useful (Q)SAR model may lack mechanistic interpretation because a model is in its early stages of evolution, because the mechanistic elements of the application domain have not been compiled from the literature, or because the underlying toxicity mechanisms are still not fully scientifically understood. Principle 5 encourages the validation process to include mechanistic interpretations, which can add to the understanding of the statistical validity and the domain of application.

The mechanistic interpretation of (Q)SARs for nanomaterials represents a major challenge because it should be based on (i) an understanding of the mechanism of action within an appropriate Adverse Outcome Pathway (AOP) (OECD, 2013b) triggered by the nanomaterial at different stages of systemic transport, (ii) knowledge of the Molecular Initiating Events (MIE) for the AOP, and once (i) and (ii) are known, (iii) quantitative understanding of interactions at the nano-bio interface. This possibility depends on the availability of the relevant nanomaterials' structural characteristics, from physicochemical characterisation, and mechanistic information.

We should state that the majority of interactions at the nano-bio interface cannot be currently understood and interpreted (quantitatively) due to the lack of adequate molecular models (Boulos et al., 2013). Therefore, the mechanistic interpretations of pathways leading to e.g. genotoxicity, protein misfolding (Khan et al., 2013), and promotion of protein aggregation (e.g. amyloid fibre formation) cannot be directly made (but may be assumed in certain cases). In contrast, the pathways involving more basic interactions leading to oxidative stress, inflammation, due to ion release of dissociating ions etc. can be directly assessed (Fahmy and Cormier, 2009; Manke et al., 2013; Xia et al., 2008; Yang et al., 2009). However, it remains the case that full understanding of the biological significance of key nanomaterial properties, such as particle shape, has still not been achieved (Bai et al., 2017).

In summary, the evolving state of understanding regarding the mechanistic basis of nanomaterial toxicity means that it is understandable if the mechanistic basis for nano-QSAR models is somewhat speculative. On the other hand, nano-QSAR studies may themselves help to advance mechanistic understanding (Bai et al., 2017; Gajewicz et al., 2015).

3. Applying the OECD principles to nano-QSAR models: examples from the NanoSafety Modelling Cluster

In order to illustrate how the OECD principles may be applied, in keeping with the detailed discussion presented in section 2, to nano-QSAR models, we critically evaluated four examples taken from the literature. Specifically, these models were developed with the support of projects in the NanoSafety Modelling Cluster, as defined in the Introduction, sometimes in collaboration with external researchers and initiatives. The models are listed in Table 1 and the evaluations are presented in Table 2.

Importantly, in keeping with Burello (Burello, 2017), we do not apply a simplistic, tick box “reject vs. accept” approach to evaluation according to the five principles. Rather, we

advocate that careful evaluation of nano-QSAR models be performed by experts in QSAR modelling of nanomaterials in consultation with experts in the relevant area of experimental nanoscience and the intended application area (e.g. regulatory decision making), as applicable.

For brevity, we do not raise every single possible question which experts in these areas would need to pose. Rather, the key questions posed in Table 2 reflect important considerations building upon our discussion under section 2. Other relevant questions, which would require more detailed critical examination of the models, should also be considered. For example, additional aspects related to data quality and completeness, e.g. in terms of the physicochemical characterization and experimental metadata, would be relevant for assessing compliance with principles 1 (a defined endpoint, discussed in section 2.1) and 3 (a defined domain of applicability, discussed in section 2.3). A detailed discussion of assessing the completeness and quality of nanomaterial data was recently presented elsewhere (Marchese Robinson et al., 2016). In addition to descriptive documentation of the modelling workflow, full compliance with principle 2 (discussed in section 2.2), may require the relevant source code and, where relevant, random number generator seeds and other computational details to be documented in order to fully reproduce the models, and/or for the structure of the models to be encoded in the Predictive Model Markup Language (PMML) along with other necessary information required to calculate the descriptors etc. (Editorial, 2014; Helma et al., 2017; Tetko et al., 2017). (The "other necessary information" might be documented using the QMRF or QSARdb formats (Tetko et al., 2017), possibly with adaptations as discussed in section 2.2.) Likewise, careful consideration is required of whether the performance statistics were suitable (Alexander et al., 2015) and were obtained from reliable, truly external, or otherwise unbiased, validation protocols, as is discussed in section 2.4 and in key literature references (Braga-Neto and Dougherty, 2004; Hawkins, 2004; Hawkins et al., 2003; Tropsha, 2010; Tropsha et al., 2003). A relevant consideration, regarding the reliability of the

validation statistics, relates to the size of the nanomaterial datasets, which are typically small, as discussed under section 2.4. Regarding the principle of a mechanistic interpretation, as discussed in section 2.5, the fact that our understanding of the mechanisms of action of nanomaterial effects may be expected to continue to evolve, means that assessing the mechanistic plausibility of a nano-QSAR model remains challenging. Finally, regarding the question of whether the endpoint is relevant for the intended application of the model, as discussed in section 2.1, the evaluated models (Table 1) might be considered for use in a variety of different contexts (e.g. initial screening or safety-by-design by manufacturers, as compared to decision support for regulators) and, outside of a specific context, this question cannot be directly answered.

Table 1. Summary of models evaluated in Table 2

Model Label	Model Title and Reference	Supporting Project(s)
Model 1	Random Forest nanosilica cytotoxicity model (Cassano et al., 2016)	NanoPUZZLES
Model 2	SVM metal oxide nanoparticle toxicity classifier (Liu et al., 2013a)	MODERN
Model 3	CORAL carbon nanotube mutagenicity model (Toropov and Toropova, 2015)	PreNanoTox & NanoPUZZLES
Model 4	Local weighted Random Forest (proteomics descriptors) gold and silver nanoparticle cell association model (Helma et al., 2017)	eNanoMapper

Table 2. Minimal evaluation of selected nano-QSAR models (Table 1) according to the OECD principles

OECD Principle	Key Question	Model 1	Model 2	Model 3	Model 4
A defined endpoint (section 2.1)	Is a precise definition of the endpoint provided (e.g. an ontology definition)?	Yes. The endpoint is –log(EC25), fully described, for the *in vitro* WST-1 assay.	The definition of "toxic" vs. "non-toxic" classes is clearly described. However, as this is based upon clustering analysis of data from a variety of toxicity assays, the interpretation of this endpoint is arguably complicated.	Yes. The endpoint is defined as –log(TA100), where TA100 is defined as the mean mutant counts from the Salmonella microsome test in the TA100 strain tested at one, out of many, doses.	Yes. The endpoint is log2[net cell association], where net cell association was described herein (Helma et al., 2017) and precisely defined in the cited original data source (Walkey et al., 2014).
	Are the test methods / assays, along with the key	Yes, to some extent. The assay, cell type and treatment type are	Yes	Yes	Yes, in the original source of the dataset (Walkey et

	experimental variables, used to generate the endpoint data documented?	documented. Other relevant details, e.g. exposure medium details, are not provided.			al., 2014).
	Are the data "reasonably" experimentally consistent?	Data were combined from diverse protocols, e.g. different cell types. However, this variability is accounted for via considering the varied experimental conditions as descriptors.	Whilst data from multiple assays were combined to provide the "toxic" vs. "non-toxic" endpoint, each individual assay was performed according to a consistent protocol as described in a single primary literature citation (Zhang et al., 2012).	Data were generated according to different conditions, yet in the same lab according to a single experimental protocol described in the cited primary literature reference (Wirnitzer et al., 2009). However, this variability is accounted for via considering the varied experimental conditions as descriptors.	Data were generated according to a single experimental protocol (Walkey et al., 2014).
	Are units provided for numerical endpoints?	Yes. The units for the EC25 values are provided.	N/A	N/A	Yes
	Have concentration / dose related units been converted from mass based units (e.g. into surface area based units)?	Yes. The EC25 values are expressed in surface area concentration units.	No (Zhang et al., 2012)	No (Wirnitzer et al., 2009)	No
	Has the potential for nanomaterial interference with the assays been excluded?	This is unclear. It has been suggested that the WST-1 assay avoids certain interference problems with some carbon based nanomaterials, compared to some other cytotoxicity assays (Domey et al., 2013).	This is unclear.	This is unclear.	This is unclear.
	Has endotoxin or residual solvent contamination been assessed?	This was not addressed.	This was not addressed.	This was not addressed.	This was not addressed.
An unambiguous algorithm (section 2.2)	Which descriptors were used?	An expert selected combination of physicochemical measurement and experimental variables, all converted to binary variables, were	A pool of thirty descriptors was initially considered, including simple atom counts, experimental properties of the constituents, and	Different experimental conditions were converted to labels incorporated into "quasi-SMILES".	Experimental measurements characterizing interactions with proteins in human serum, as fully described in the original source of

		used.	conduction band energy estimated from the measured band gap (Zhang et al., 2012). This initial pool of descriptors was reduced to two via statistical feature selection: the conduction band energy and the ionic index of the metal cation.		the dataset (Walkey et al., 2014), were reduced via feature selection.
	Which statistical / machine learning algorithm was used?	Random Forest	SVM	The CORAL software was employed (http://www.insilico.eu/coral).	Weighted local average Random Forest modelling, trained on the nearest neighbours identified via descriptor similarity.
A defined domain of applicability (section 2.3)	Have the modelled nanomaterials been adequately characterized in terms of their physico-chemical characteristics?	To some extent. The composition of the core and, to some extent, details of the coatings are provided in the Supporting Information. Some additional physicochemical characterization data are provided, with partial description of experimental protocols.	To a considerable extent. The chemical composition, along with experimentally measured sizes, zeta potential, dissolution profile and crystalline structure are provided, along with experimental details, in the cited primary literature (Zhang et al., 2012).	To some extent. The tested nanomaterials were described as the commercially available Baytubes® (multi-walled carbon nanotubes) and a characterization of size and shape distribution for the prepared samples is provided in the primary literature reference (Wirnitzer et al., 2009). N.B. Since only experimental conditions were incorporated into the model, based on results for a single nanomaterial, it could not be expected to be applicable to anything other than these same nanomaterials.	To a considerable extent. Detailed descriptions of core and surface chemical composition are provided, alongside detailed measurement of size distributions, according to a variety of techniques, and zeta potential in the original source of the dataset (Walkey et al., 2014).
	Has an applicability domain been defined?	No	Yes	Yes, based on analysis of the “quasi-SMILES”.	Yes
	Are uncertainty estimates provided for predictions?	No	In some sense: the model estimates the probability of being toxic.	No	Yes (95% prediction intervals)
Appropriate measures of goodness-of-fit, robustness and predictivity (section 2.4)	How was the ability of the model to make predictions for untested nanomaterials assessed?	“External” LOO cross-validation	Via the “0.632 estimator”, based on bootstrapping.	LMO cross-validation	5 repetitions of 10-fold cross-validation, with feature selection being repeated for each training set to avoid overfitting

	What were the overall performance statistics and their values?	R^2 = 0.78	Balanced classification accuracy = 93.74%	r^2 (correlation coefficient) = 0.74 – 0.80 (across three validation splits); 0.75 – 0.82 (only considering instances inside the domain of applicability)	r^2 (explained model variance) = 0.55 – 0.68 across all five repetitions of cross-validation; RMSE = 1.51 – 1.8 log units; 87 – 92% of predictions within the 95% prediction intervals
Mechanistic interpretation, if possible (section 2.5)	Was a mechanistic interpretation proposed by the study authors?	This issue is touched upon, but a full mechanistic interpretation is not provided.	Yes	Yes	No
	How do the study authors arrive at this interpretation?	They evaluate "feature contributions" for Random Forest.	They evaluate the range of descriptor values associated with toxic vs. non-toxic classifications, according to the model.	They consider variables which have a "stable positive correlation weight" with the endpoint.	N/A
	Do the study authors consider whether this mechanistic interpretation is consistent with current understanding?	Yes	Yes	No	N/A

4. Conclusions

(Quantitative) Structure-Activity Relationship ([Q]SAR) modelling is one possibility for estimating hazard or exposure related characteristics, either in the context of a risk assessment or as part of a "safe-by-design" approach. It might be used for predicting a variety of properties, including toxicity of newly designed nanoparticles. However, every nano-(Q)SAR must be appropriately validated, which is crucial for ensuring its predictive accuracy. In this contribution, we proposed an interpretation of the well-known "OECD Principles for the Validation, for Regulatory Purposes, of (Q)SAR Models" in the context of nano-(Q)SAR and presented our opinion on the criteria to be fulfilled by every model developed for nanoparticles, whether employed in a regulatory context or otherwise.

In general, we agree that the OECD principles create an appropriate framework for validating nano-(Q)SARs as well. However, special attention is required for some issues specific to nanoparticles. The most important conclusions are as follows:

- Careful consideration is required as to whether or not the activity data are reliable, due to various potential pitfalls associated with experimental assessment of nanomaterial activity, and whether or not appropriate concentration/dose units have been used.
- Classic molecular descriptors are typically, but not always, inappropriate for modelling nanoparticles. On the other hand, newly developed descriptors should be validated and reported by providing the details necessary for anyone interested to calculate them or, where applicable, obtain them from experimental characterization of the nanomaterials.
- It is highly recommended that the models are presented and documented in a format such as the descriptive (Q)SAR Model Reporting Format (QMRF), or the machine readable QsarDB format – including model representation using Predictive Model Markup Language (PMML) and, whenever possible, be placed in public repositories and exposed as web applications to ensure their transparency and reproducibility. N.B. Guidance on suitable linked data files, or possibly other adaptations of these formats, is required to ensure their suitability for nano-QSAR, as opposed to classical (Q)SAR, models.
- Compared to classical (Q)SAR modelling, it seems much more important to achieve a balance between the level of confidence in the predictions of a nano-(Q)SAR and its scope (i.e., applicability domain). It is essential to reach a balance between "local" and "global" models.
- Nano-bio interactions involving nanoparticles are not fully determined by the particle chemistry alone. Because of that, mechanistic interpretation of nano-(Q)SARs can be

more problematic than those for classic (Q)SAR models; very often a wider context is required.

Acknowledgements

TP, RLMR, ANR, AG, MGP, MTDC and EB acknowledge the funding received from the European Union Seventh Framework Programme (FP7/2007- 2013) under grant agreement n° 309837 (NanoPUZZLES project) and TP and AG acknowledge the financial support of the Foundation for Polish Science (FOCUS Programme). RR and AF acknowledge the funding received from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 309314 (MODERN project) and from the Generalitat de Catalunya (2014 SGR 1352). The eNanoMapper project is funded by the European Union's Seventh Framework Programme for research, technological development and demonstration (FP7-NMP-2013-SMALL-7) under grant agreement no. 604134. VL acknowledges the funding received from the European Union Seventh Framework Programme under grant agreement n° 310465 (MembraneNanoPart project).

References

Aberg, C., 2015. NanoSafety Cluster Databases Working Group. Overview and recommendation of data quality: Working draft.

Alexander, D.L.J., Tropsha, A., Winkler, D.A., 2015. Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. J Chem Inf Model 55, 1316-1322.

Astefanei, A., Nunez, O., Galceran, M.T., 2015. Characterisation and determination of fullerenes: A critical review. Anal Chim Acta 882, 1-21.

Baalousha, M., Lead, J.R., 2012. Rationalizing Nanomaterial Sizes Measured by Atomic Force Microscopy, Flow Field-Flow Fractionation, and Dynamic Light Scattering: Sample Preparation, Polydispersity, and Particle Structure. Environ Sci Technol 46, 6134-6142.

Bai, X., Liu, F., Liu, Y., Li, C., Wang, S.Q., Zhou, H.Y., Wang, W.Y., Zhu, H., Winkler, D.A., Yan, B., 2017. Toward a systematic exploration of nano-bio interactions. Toxicol Appl Pharm 323, 66-73.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F., Nielsen, H., 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16, 412-424.

Baskin, I.I., Winkler, D., Tetko, I.V., 2016. A renaissance of neural networks in drug discovery. Expert Opin Drug Dis 11, 785-795.

Berrar, D., Bradbury, I., Dubitzky, W., 2006a. Avoiding model selection bias in small-sample genomic datasets. Bioinformatics 22, 1245-1250.

Berrar, D., Bradbury, I., Dubitzky, W., 2006b. Avoiding model selection bias in small-sample genomic datasets (vol 22, pg 1245, 2006). Bioinformatics 22, 2453-2453.

Boulos, S.P., Davis, T.A., Yang, J.A., Lohse, S.E., Alkilany, A.M., Holland, L.A., Murphy, C.J., 2013. Nanoparticle-Protein Interactions: A Thermodynamic and Kinetic Study of the Adsorption of Bovine Serum Albumin to Gold Nanoparticle Surfaces. Langmuir 29, 14984-14996.

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

Brehm, M., Kafka, A., Bamler, M., Kuhne, R., Schuurmann, G., Sikk, L., Burk, J., Burk, P., Tamm, T., Tamm, K., Pokhrel, S., Madler, L., Kahru, A., Aruoja, V., Sihtmae, M., Scott-Fordsmand, J., Sorensen, P.B., Escorihuela, L., Roca, C.P., Fernandez, A., Giralt, F., Rallo, R., 2017. An Integrated Data-Driven Strategy for Safe-by-Design Nanoparticles: The FP7 MODERN Project. Adv Exp Med Biol 947, 257-301.

Burello, E., 2017. Review of (Q)SAR models for regulatory assessment of nanomaterials risks. NanoImpact 8, 48-58.

Burello, E., Worth, A.P., 2011. A theoretical framework for predicting the oxidative stress potential of oxide nanoparticles. Nanotoxicology 5, 228-235.

Cassano, A., Robinson, R.L.M., Palczewska, A., Puzyn, T., Gajewicz, A., Tran, L., Manganelli, S., Cronin, M.T.D., 2016. Comparing the CORAL and Random Forest Approaches for Modelling the In Vitro Cytotoxicity of Silica Nanomaterials. Atla-Altern Lab Anim 44, 533-556.

Chen, G., Peijnenburg, W., Xiao, Y., Vijver, M.G., 2017. Current Knowledge on the Use of Computational Toxicology in Hazard Assessment of Metallic Engineered Nanomaterials. Int J Mol Sci 18.

Cherkasov, A., Muratov, E.N., Fourches, D., Varnek, A., Baskin, II, Cronin, M., Dearden, J., Gramatica, P., Martin, Y.C., Todeschini, R., Consonni, V., Kuz'min, V.E., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A., Tropsha, A., 2014. QSAR modeling: where have you been? Where are you going to? Journal of medicinal chemistry 57, 4977-5010.

Chirico, N., Gramatica, P., 2011. Real External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient. J Chem Inf Model 51, 2320-2335.

Chirico, N., Gramatica, P., 2012. Real External Predictivity of QSAR Models. Part 2. New Intercomparable Thresholds for Different Validation Criteria and the Need for Scatter Plot Inspection. J Chem Inf Model 52, 2044-2058.

Cho, W.S., Duffin, R., Thielbeer, F., Bradley, M., Megson, I.L., MacNee, W., Poland, C.A., Tran, C.L., Donaldson, K., 2012. Zeta Potential and Solubility to Toxic Ions as Mechanisms of Lung Inflammation Caused by Metal/Metal Oxide Nanoparticles. Toxicol Sci 126, 469-477.

Chomenidis, C., Drakakis, G., Tsiliki, G., Anagnostopoulou, E., Valsamis, A., Doganis, P., Sopasakis, P., Sarimveis, H., 2017. Jaqpot Quattro: A novel computational web platform for modelling and analysis in nanoinformatics. J Chem Inf Model Just Accepted Manuscript.

Clark, R.D., 2009. DPRESS: Localizing estimates of predictive uncertainty. J Cheminformatics 1.

Coe, B.J., Avramopoulos, A., Papadopoulos, M.G., Pierloot, K., Vancoillie, S., Reis, H., 2013. Theoretical Modelling of Photoswitching of Hyperpolarisabilities in Ruthenium Complexes. Chem-Eur J 19, 15955-15963.

Cohen, J.M., Teeguarden, J.G., Demokritou, P., 2014. An integrated approach for the in vitro dosimetry of engineered nanomaterials. Part Fibre Toxicol 11.

Consonni, V., Todeschini, R., 2010. Molecular descriptors, in: T. Puzyn, M.T.D. Cronin, J. Leszczynski (Eds.), Recent advances in QSAR studies: Methods and applications. Springer, Dordrecht, Heidelberg, London, New York.

Cooper, J.A., Saracci, R., Cole, P., 1979. Describing the Validity of Carcinogen Screening-Tests. Brit J Cancer 39, 87-89.

Crist, R.M., Grossman, J.H., Patri, A.K., Stern, S.T., Dobrovolskaia, M.A., Adiseshaiah, P.P., Clogston, J.D., McNeil, S.E., 2013. Common pitfalls in nanotechnology: lessons learned from NCI's Nanotechnology Characterization Laboratory. Integr Biol-Uk 5, 66-73.

Cronin, M.T.D., Schultz, T.W., 2003. Pitfalls in QSAR. J Mol Struc-Theochem 622, 39-51.

Dearden, J.C., 2016. The History and Development of Quantitative Structure-Activity Relationships (QSARs). International Journal of Quantitative Structure-Property Relationships 1, 1-44.

Dearden, J.C., Cronin, M.T.D., Kaiser, K.L.E., 2009. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). Sar Qsar Environ Res 20, 241-266.

DLA, 2011. Dynamic light scattering - common terms defined (White paper). Malvern Instruments Limited.

Domey, J., Haslauer, L., Grau, I., Strobel, C., Kettering, M., Hilger, I., 2013. Probing the cytotoxicity of nanoparticles: Experimental pitfalls and artifacts, in: J. Wegner (Ed.), Measuring Biological Impacts of Nanomaterials. Springer, Cham, Switzerland, pp. pp. 31–34.

Donaldson, K., Poland, C.A., 2013. Nanotoxicity: Challenging the myth of nano-specific toxicity. Curr. Opin. Biotechnol. 24, 724-734.

Dragos, H., Gilles, M., Alexandre, V., 2009. Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models. J Chem Inf Model 49, 1762-1776.

Drakakis, G., Chomenidis, C., Tsiliki, C., Doganis, P., Anagnostopoulou, E., Sarimveis, H., Rautenberg, M., Gebele, D., Helma, C., Jeliazkova, N., Jeliazkov, V., Hardy, B., 2016. Deliverable Raport D4.6 Tools for generating QMRF and QPRF reports. Zenodo. http://doi.org/10.5281/zenodo.375619.

EChA, 2008. Guidance on information requirements and chemical safety assessment. Chapter R.6: QSARs and grouping of chemicals. In: Guidance for the implementation of REACH. European Chemical Agency.

Editorial, N.M., 2014. Software with impact. Nature Methods 11, 211.

Eriksson, L., Johansson, E., 1996. Multivariate design and modeling in QSAR. Chemometr Intell Lab 34, 1-19.

Esbensen, K.H., Geladi, P., 2010. Principles of Proper Validation: use and abuse of re-sampling for validation. J Chemometr 24, 168-187.

Fahmy, B., Cormier, S.A., 2009. Copper oxide nanoparticles induce oxidative stress and cytotoxicity in airway epithelial cells. Toxicol in Vitro 23, 1365-1371.

Ferreira, A.J., Figueiredo, M.A.T., 2012. Efficient feature selection filters for high-dimensional data. Pattern Recogn Lett 33, 1794-1804.

Fourches, D., Pu, D.Q.Y., Tassa, C., Weissleder, R., Shaw, S.Y., Mumper, R.J., Tropsha, A., 2010. Quantitative Nanostructure-Activity Relationship Modeling. Acs Nano 4, 5703-5712.

Gadaleta, D., Mangiatordi, G.F., Catto, M., Carotti, A., Nicolotti, O., 2016. Applicability Domain for QSAR Models: Where Theory Meets Reality. International Journal of Quantitative Structure-Property Relationships 1, 45-63.

Gajewicz, A., Rasulev, B., Dinadayalane, T.C., Urbaszek, P., Puzyn, T., Leszczynska, D., Leszczynski, J., 2012. Advancing risk assessment of engineered nanomaterials: Application of computational approaches. Adv Drug Deliver Rev 64, 1663-1693.

Gajewicz, A., Schaeublin, N., Rasulev, B., Hussain, S., Leszczynska, D., Puzyn, T., Leszczynski, J., 2015. Towards understanding mechanisms governing cytotoxicity of metal oxides nanoparticles: Hints from nano-QSAR studies. Nanotoxicology 9, 313-325.

Gasteiger, J., Zupan, J., 1993. Neural Networks in Chemistry. Angew Chem Int Edit 32, 503-527.

Golbamaki, N., Rasulev, B., Cassano, A., Robinson, R.L.M., Benfenati, E., Leszczynski, J., Cronin, M.T.D., 2015. Genotoxicity of metal oxide nanomaterials: review of recent data and discussion of possible mechanisms. Nanoscale 7, 2154-2198.

Golbraikh, A., Muratov, E., Fourches, D., Tropsha, A., 2014. Data Set Modelability by QSAR. J Chem Inf Model 54, 1-4.

Golbraikh, A., Shen, M., Xiao, Z.Y., Xiao, Y.D., Lee, K.H., Tropsha, A., 2003. Rational selection of training and test sets for the development of validated QSAR models. J Comput Aid Mol Des 17, 241-253.

Golbraikh, A., Tropsha, A., 2002. Beware of q(2)! J Mol Graph Model 20, 269-276.

Gorodkin, J., 2004. Comparing two K-category assignments by a K-category correlation coefficient. Comput Biol Chem 28, 367-374.

Gramatica, P., 2007. Principles of QSAR models validation: internal and external. Qsar Comb Sci 26, 694-701.

Gramatica, P., 2013. On the development and validation of QSAR models. Methods in molecular biology 930, 499-526.

Gramatica, P., Sangion, A., 2016. A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics and Terminology. J Chem Inf Model 56, 1127-1131.

Guyon, I., Elisseeff, A., 2013. An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 3, 1157-1182.

Handy, R.D., van den Brink, N., Chappell, M., Muhling, M., Behra, R., Dusinska, M., Simpson, P., Ahtiainen, J., Jha, A.N., Seiter, J., Bednar, A., Kennedy, A., Fernandes, T.F., Riediker, M., 2012. Practical considerations for conducting ecotoxicity test methods with manufactured nanomaterials: what have we learnt so far? Ecotoxicology 21, 933-972.

Hansch, C., Fujita, T., 1964. Rho-Sigma-Pi Analysis . Method for Correlation of Biological Activity + Chemical Structure. J Am Chem Soc 86, 1616-&.

Hanser, T., Barber, C., Marchaland, J.F., Werner, S., 2016. Applicability domain: towards a more formal definition. Sar Qsar Environ Res 27, 865-881.

Hassellov, M., Readman, J.W., Ranville, J.F., Tiede, K., 2008. Nanoparticle analysis and characterization methodologies in environmental risk assessment of engineered nanoparticles. Ecotoxicology 17, 344-361.

Hastings, J., Jeliazkova, N., Owen, G., Tsiliki, G., Munteanu, C.R., Steinbeck, C., Willighagen, E., 2015. eNanoMapper: harnessing ontologies to enable data integration for nanomaterial risk assessment. J Biomed Semant 6.

Hawkins, D.M., 2004. The problem of overfitting. J Chem Inf Comp Sci 44, 1-12.

Hawkins, D.M., Basak, S.C., Mills, D., 2003. Assessing Model Fit by Cross-Validation. J Chem Inf Comp Sci 43, 579-586.

Helma, C., 2006. Lazy structure-activity relationships (lazar) for the prediction of rodent carcinogenicity and Salmonella mutagenicity. Mol Divers 10, 147-158.

Helma, C., Rautenberg, M., Gebele, D., 2017. Nano-Lazar: Read across Predictions for Nanoparticle Toxicities with Calculated and Measured Properties. Front Pharmacol 8.

Hu, X.K., Cook, S., Wang, P., Hwang, H.M., 2009. In vitro evaluation of cytotoxicity of engineered metal oxide nanoparticles. Sci Total Environ 407, 3070-3072.

Iqbal, M.S., Golsteijn, L., Oberg, T., Sahlin, U., Papa, E., Kovarich, S., Huijbregts, M.A.J., 2013. Understanding Quantitative Structure-Property Relationships Uncertainty in Environmental Fate Modeling. Environ Toxicol Chem 32, 1069-1076.

Jeliazkova, N., Chomenidis, C., Doganis, P., Fadeel, B., Grafstrom, R., Hardy, B., Hastings, J., Hegi, M., Jeliazkov, V., Kochev, N., Kohonen, P., Munteanu, C.R., Sarimveis, H., Smeets, B., Sopasakis, P., Tsiliki, G., Vorgrimmler, D., Willighagen, E., 2015. The eNanoMapper database for nanomaterial safety information. Beilstein J Nanotechnol 6, 1609-1634.

Jin, Z.Y., El-Deiry, W.S., 2005. Overview of cell death signaling pathways. Cancer Biol Ther 4, 139-163.

Johnson, A.M., Maggiora, G.M., 1990. Concepts and Applications of Molecular Similarity. John Willey & Sons, New York.

Kaplan, R.M., Chambers, D.A., Glasgow, R.E., 2014. Big data and large sample size: a cautionary note on the potential for bias. Clin Transl Sci 7, 342-346.

Kar, S., Gajewicz, A., Puzyn, T., Roy, K., 2014a. 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 28, 600-606.

Kar, S., Gajewicz, A., Puzyn, T., Roy, K., Leszczynski, J., 2014b. Periodic table-based descriptors to encode cytotoxicity profile of metal oxide nanoparticles: a mechanistic QSTR approach. Ecotoxicology and environmental safety 107, 162-169.

Keefer, C.E., Kauffman, G.W., Gupta, R.R., 2013. Interpretable, Probability-Based Confidence Metric for Continuous Quantitative Structure-Activity Relationship Models. J Chem Inf Model 53, 368-383.

Kennard, R.W., Stone, L.A., 1969. Computer Aided Design of Experiments. Technometrics 11, 137-&.

Khan, T.A., Mazid, M., Ansari, S.A., Azam, A., Naeem, A., 2013. Zinc Oxide Nanoparticles Promote the Aggregation of Concanavalin A. Int J Pept Res Ther 19, 135-146.

Kleandrova, V.V., Luan, F., Gonzalez-Diaz, H., Ruso, J.M., Speck-Planche, A., Cordeiro, M.N.D.S., 2014. Computational Tool for Risk Assessment of Nanomaterials: Novel QSTR-Perturbation Model for Simultaneous Prediction of Ecotoxicity and Cytotoxicity of Uncoated and Coated Nanoparticles under Multiple Experimental Conditions. Environ Sci Technol 48, 14686-14694.

Kroll, A., Pillukat, M.H., Hahn, D., Schnekenburger, J., 2009. Current in vitro methods in nanoparticle risk assessment: Limitations and challenges. Eur J Pharm Biopharm 72, 370-377.

Krug, H.F., 2014. Nanosafety Research-Are We on the Right Track? Angew Chem Int Edit 53, 12304-12319.

Le, T.C., Mulet, X., Burden, F.R., Winkler, D.A., 2013. Predicting the complex phase behavior of self-assembling drug delivery nanoparticles. Mol Pharm 10, 1368-1377.

Lewinski, N., Colvin, V., Drezek, R., 2008. Cytotoxicity of nanoparticles. Small 4, 26-49.

Lewis, R.W., Billington, R., Debryune, E., Gamer, A., Lang, B., Carpanini, F., 2002. Recognition of adverse and nonadverse effects in toxicity studies. Toxicol Pathol 30, 66-74.

Lindgren, F., Hansen, B., Karcher, W., Sjostrom, M., Eriksson, L., 1996. Model validation by permutation tests: Applications to variable selection. J Chemometr 10, 521-532.

Lindh, M., Karlen, A., Norinder, U., 2017. Predicting the Rate of Skin Penetration Using an Aggregated Conformal Prediction Framework. Mol Pharmaceut 14, 1571-1576.

Liu, R., Rallo, R., Weissleder, R., Tassa, C., Shaw, S., Cohen, Y., 2013a. Nano-SAR Development for Bioactivity of Nanoparticles with Considerations of Decision Boundaries. Small 9, 1842-1852.

Liu, R., Zhang, H.Y., Ji, Z.X., Rallo, R., Xia, T., Chang, C.H., Nel, A., Cohen, Y., 2013b. Development of structure-activity relationship for metal oxide nanoparticles. Nanoscale 5, 5644-5653.

Lopez, H., Brandt, E.G., Mirzoev, A., Zhurkin, D., Lyubartsev, A., Lobaskin, V., 2017. Multiscale Modelling of Bionano Interface. Adv Exp Med Biol 947, 173-206.

Lopez, H., Lobaskin, V., 2015. Coarse-grained model of adsorption of blood plasma proteins onto nanoparticles. J Chem Phys 143.

Lövestam, G., Rauscher, H., Roebben, G., Sokull Klüttgen, B., Gibson, N., Putaud, J.-P., Stamm, H., 2010. Considerations on a Definition of Nanomaterial for Regulatory Purposes, JRC Reference Reports. European Commission Joint Research Centre.

Low, Y., Uehara, T., Minowa, Y., Yamada, H., Ohno, Y., Urushidani, T., Sedykh, A., Muratov, E., Kuz'min, V., Fourches, D., Zhu, H., Rusyn, I., Tropsha, A., 2011. Predicting Drug-Induced Hepatotoxicity Using QSAR and Toxicogenomics Approaches. Chem Res Toxicol 24, 1251-1262.

Luan, F., Kleandrova, V.V., Gonzalez-Diaz, H., Ruso, J.M., Melo, A., Speck-Planche, A., Cordeiro, M.N., 2014. Computer-aided nanotoxicology: assessing cytotoxicity of nanoparticles under diverse experimental conditions by using a novel QSTR-perturbation approach. Nanoscale.

Lubinski, L., Urbaszek, P., Gajewicz, A., Cronin, M.T.D., Enoch, S.J., Madden, J.C., Leszczynska, D., Leszczynski, J., Puzyn, T., 2013. Evaluation criteria for the quality of published experimental data on nanomaterials and their usefulness for QSAR modelling. Sar Qsar Environ Res 24, 995-1008.

Lusted, L.B., 1971. Signal Detectability and Medical Decision-Making. Science 171, 1217-&.

Lv, M., Huang, W., Chen, Z., Jiang, H., Chen, J., Tian, Y., Zhang, Z., Xu, F., 2015. Metabolomics techniques for nanotoxicity investigations. Bioanalysis 7, 1527-1544.

Manke, A., Wang, L.Y., Rojanasakul, Y., 2013. Mechanisms of Nanoparticle-Induced Oxidative Stress and Toxicity. Biomed Res Int.

Marchese Robinson, R.L., Cronin, M.T., Richarz, A.N., Rallo, R., 2015. An ISA-TAB-Nano based data collection framework to support data-driven modelling of nanotoxicology. Beilstein J Nanotechnol 6, 1978-1999.

Marchese Robinson, R.L., Lynch, I., Peijnenburg, W., Rumble, J., Klaessig, F., Marquardt, C., Rauscher, H., Puzyn, T., Purian, R., Aberg, C., Karcher, S., Vriens, H., Hoet, P., Hoover, M.D., Hendren, C.O., Harper, S.L., 2016. How should the completeness and quality of curated nanomaterial data be evaluated? Nanoscale 8, 9919-9943.

Marquardt, C., Kuhnel, D., Richter, V., Krug, H.F., Mathes, B., Steinbach, C., Nau, K., 2013. Latest research results on the effects of nanomaterials on humans and the environment: DaNa - Knowledge Base Nanomaterials. J Phys Conf Ser 429.

Massart, D.L., Vandeginste, B.G., Buydens, L.M.C., Lewi, P.J., Smeyers-Verbeke, J., De Jong, S., 1997. Handbook of Chemometrics and Qualimetrics: Part A. Elsevier Science Inc., New York, NY, USA.

Mikolajczyk, A., Gajewicz, A., Rasulev, B., Schaeublin, N., Maurer-Gardner, E., Hussain, S., Leszczynski, J., Puzyn, T., 2015. Zeta Potential for Metal Oxide Nanoparticles: A Predictive Model Developed by a Nano-Quantitative Structure-Property Relationship Approach. Chem Mater 27, 2400-2407.

Miller, J.B., Hobbie, E.K., 2013. Nanoparticles as macromolecules. J Polym Sci Pol Phys 51, 1195-1208.

Mitchell, J.B.O., 2014. Machine learning methods in chemoinformatics. Wires Comput Mol Sci 4, 468-481.

Murdock, R.C., Braydich-Stolle, L., Schrand, A.M., Schlager, J.J., Hussain, S.M., 2008. Characterization of nanomaterial dispersion in solution prior to In vitro exposure using dynamic light scattering technique. Toxicol Sci 101, 239-253.

Nel, A.E., 2013. Implementation of alternative test strategies for the safety assessment of engineered nanomaterials. J Intern Med 274, 561-577.

Netzeva, T.I., Worth, A.P., Aldenberg, T., Benigni, R., Cronin, M.T.D., Gramatica, P., Jaworska, J.S., Kahn, S., Klopman, G., Marchant, C.A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G.Y., Perkins, R., Roberts, D.W., Schultz, T.W., Stanton, D.T., van de Sandt, J.J.M., Tong, W.D., Veith, G., Yang, C.H., 2005. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships - The report and recommendations of ECVAM Workshop 52. Atla-Altern Lab Anim 33, 155-173.

Newcombe, R.G., 1998. Two-sided confidence intervals for the single proportion: Comparison of seven methods. Stat Med 17, 857-872.

Norinder, U., Carlsson, L., Boyer, S., Eklund, M., 2014. Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination. J Chem Inf Model 54, 1596-1603.

OECD, 2004. THE REPORT FROM THE EXPERT GROUP ON (QUANTITATIVE) STRUCTURE-ACTIVITY RELATIONSHIPS [(Q)SARs] ON THE PRINCIPLES FOR THE VALIDATION OF (Q)SARs. Organisation for Economic Co-operation and Development.

OECD, 2007. GUIDANCE DOCUMENT ON THE VALIDATION OF (QUANTITATIVE)STRUCTURE-ACTIVITY RELATIONSHIPS [(Q)SAR] MODELS; OECD SERIES ON TESTING AND ASSESSMENT. Organisation for Economic Co-operation and Development.

OECD, 2009. Preliminary Review of OECD Test Guidelines for their Applicability to Manufactured Nanomaterials, Publications in the Series on the Safety of Manufactured Nanomaterials. Organisation for Economic Co-operation and Development.

OECD, 2010. GUIDANCE MANUAL FOR THE TESTING OF MANUFACTURED NANOMATERIALS: OECD's SPONSORSHIP PROGRAMME; FIRST REVISION, Series of Safety of Manufactured Nanomaterials. Organisation for Economic Co-operation and Development.

OECD, 2012. GUIDANCE ON SAMPLE PREPARATION AND DOSIMETRY FOR THE SAFETY TESTING OF MANUFACTURED NANOMATERIALS, Series on the Safety of Manufactured Nanomaterials. Organisation for Economic Co-operation and Development.

OECD, 2013a. Co-Operation on Risk Assessment: Prioritisation of Important Issues on Risk Assessment of Manufactured Nanomaterials - Final Report, Publications in the Series on the Safety of Manufactured Nanomaterials. Organisation for Economic Co-operation and Development.

OECD, 2013b. GUIDANCE DOCUMENT ON DEVELOPING AND ASSESSING ADVERSE OUTCOME PATHWAYS, Series on Testing and Assessment. Organisation for Economic Co-operation and Development.

OECD, 2014a. Ecotoxicology and Environmental Fate of Manufactured Nanomaterials: Test Guidelines Publications in the Series on the Safety of Manufactured Nanomaterials. Organisation for Economic Co-operation and Development.

OECD, 2014b. Report of the OECD expert meeting on the physical chemical properties of manufactured nanomaterials and test guidelines, Publications in the Series on the Safety of Manufactured Nanomaterials. Organisation for Economic Co-operation and Development.

Oksel, C., Ma, C.Y., Liu, J.J., Wilkins, T., Wang, X.Z., 2015a. (Q)SAR modelling of nanomaterial toxicity: A critical review. Particuology 21, 1-19.

Oksel, C., Ma, C.Y., Liu, J.J., Wilkins, T., Wang, X.Z., 2017. Literature Review of (Q)SAR Modelling of Nanomaterial Toxicity Modelling the Toxicity of Nanoparticles, Part of the Advances in Experimental Medicine and Biology book series (AEMB, volume 947), pp. 103-142.

Oksel, C., Ma, C.Y., Wang, X.Z., 2015b. Current situation on the availability of nanostructure-biological activity data. Sar Qsar Environ Res 26, 79-94.

Oksel, C., Ma, C.Y., Wang, X.Z., 2015c. Structure-activity relationship models for hazard assessment and risk management of engineered nanomaterials. Procedia Engineering 102, 1500-1510.

Palmer, D.S., Mitchell, J.B.O., 2014. Is Experimental Data Quality the Limiting Factor in Predicting the Aqueous Solubility of Druglike Molecules? Mol Pharmaceut 11, 2962-+.

Pathakoti, K., Huang, M.J., Watts, J.D., He, X.J., Hwang, H.M., 2014. Using experimental data of Escherichia coli to develop a QSAR model for predicting the photo-induced cytotoxicity of metal oxide nanoparticles. J Photoch Photobio B 130, 234-240.

Pavan, M., Worth, A.P., 2008. Publicly-accessible QSAR software tools developed by the Joint Research Centre. Sar Qsar Environ Res 19, 785-799.

Petersen, E.J., Henry, T.B., Zhao, J., MacCuspie, R.I., Kirschling, T.L., Dobrovolskaia, M.A., Hackley, V., Xing, B.S., White, J.C., 2014. Identification and Avoidance of Potential Artifacts and Misinterpretations in Nanomaterial Ecotoxicity Measurements. Environ Sci Technol 48, 4226-4246.

Piir, G., 2014. QDB archive #119. QsarDB repository, http://dx.doi.org/10.15152/QDB.119.

Potter, T., Matter, H., 1998. Random or rational design? Evaluation of diverse compound subsets from chemical structure databases. J Med Chem 41, 478-488.

Powers, K.W., Brown, S.C., Krishna, V.B., Wasdo, S.C., Moudgil, B.M., Roberts, S.M., 2006. Research strategies for safety evaluation of nanomaterials. Part VI. Characterization of nanoscale particles for toxicological evaluation. Toxicol Sci 90, 296-303.

Powers, K.W., Palazuelos, M., Moudgil, B.M., Roberts, S.M., 2007. Characterization of the size, shape, and state of dispersion of nanoparticles for toxicological studies. Nanotoxicology 1, 42-51.

Przybylak, K.R., Madden, J.C., Cronin, M.T.D., Hewitt, M., 2012. Assessing toxicological data quality: basic principles, existing schemes and current limitations. Sar Qsar Environ Res 23, 435-459.

Puzyn, T., Leszczynska, D., Leszczynski, J., 2009. Toward the Development of "Nano-QSARs": Advances and Challenges. Small 5, 2494-2509.

Puzyn, T., Rasulev, B., Gajewicz, A., Hu, X.K., Dasari, T.P., Michalkova, A., Hwang, H.M., Toropov, A., Leszczynska, D., Leszczynski, J., 2011. Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. Nat Nanotechnol 6, 175-178.

Rauscher, H., Sokull-Kluttgen, B., Stamm, H., 2013. The European Commission's recommendation on the definition of nanomaterial makes an impact. Nanotoxicology 7, 1195-1197.

Richarz, A.N., Avramopoulos, A., Benfenati, E., Gajewicz, A., Golbamaki Bakhtyari, N., Leonis, G., Marchese Robinson, R.L., Papadopoulos, M.G., Cronin, M.T., Puzyn, T., 2017. Compilation of Data and Modelling of Nanoparticle Interactions and Toxicity in the NanoPUZZLES Project. Adv Exp Med Biol 947, 303-324.

Ross, T.D., 2003. Accurate confidence intervals for binomial proportion and Poisson rate estimation. Comput Biol Med 33, 509-531.

Roy, K., Ambure, P., Aher, R.B., 2017. How important is to detect systematic error in predictions and understand statistical applicability domain of QSAR models? Chemometr Intell Lab 162, 44-54.

Roy, K., Das, R.N., Ambure, P., Aher, R.B., 2016. Be aware of error measures. Further studies on validation of predictive QSAR models. Chemometr Intell Lab 152, 18-33.

Rucker, C., Rucker, G., Meringer, M., 2007. y-Randomization and its variants in QSPR/QSAR. J Chem Inf Model 47, 2345-2357.

Ruusmann, V., Sild, S., Maran, U., 2014. QSAR DataBank - an approach for the digital organization and archiving of QSAR model information. J Cheminform 6, 25.

Ruusmann, V., Sild, S., Maran, U., 2015. QSAR DataBank repository: open and linked qualitative and quantitative structure-activity relationship models. J Cheminformatics 7.

Sahlin, U., 2014. Assessment of uncertainty in chemical models by Bayesian probabilities: Why, when, how? Journal of computer-aided molecular design.

Sahlin, U., Golsteijn, L., Iqbal, M.S., Peijnenburg, W., 2013. Arguments for Considering Uncertainty in QSAR Predictions in Hazard and Risk Assessments. Atla-Altern Lab Anim 41, 91-110.

Sahlin, U., Jeliazkova, N., Oberg, T., 2014. Applicability Domain Dependent Predictive Uncertainty in QSAR Regressions. Mol Inform 33, 26-35.

Serrano-Andres, L., Avramopoulos, A., Li, J.B., Labeguerie, P., Begue, D., Kello, V., Papadopoulos, M.G., 2009. Linear and nonlinear optical properties of a series of Ni-dithiolene derivatives. J Chem Phys 131.

Sheridan, R.P., 2012. Three Useful Dimensions for Domain Applicability in QSAR Models Using Random Forest. J Chem Inf Model 52, 814-823.

Sheridan, R.P., 2013. Using Random Forest To Model the Domain Applicability of Another Random Forest Model. J Chem Inf Model 53, 2837-2850.

Sheridan, R.P., Feuston, B.P., Maiorov, V.N., Kearsley, S.K., 2004. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. J Chem Inf Comp Sci 44, 1912-1928.

Silva, F., Arezes, P., Swuste, P., 2015. Systematic design analysis and risk management on engineered nanoparticles occupational exposure. Sho2015: International Symposium on Occupational Safety and Hygiene, 350-352.

Singh, K.P., Gupta, S., 2014. Nano-QSAR modeling for predicting biological activity of diverse nanomaterials. Rsc Adv 4, 13215-13230.

Sizochenko, N., Rasulev, B., Gajewicz, A., Kuz'min, V., Puzyn, T., Leszczynski, J., 2014. From basic physics to mechanisms of toxicity: the "liquid drop" approach applied to develop predictive classification models for toxicity of metal oxide nanoparticles. Nanoscale 6, 13986-13993.

Stefaniak, A.B., Hackley, V.A., Roebben, G., Ehara, K., Hankin, S., Postek, M.T., Lynch, I., Fu, W.E., Linsinger, T.P.J., Thunemann, A.F., 2013. Nanoscale reference materials for environmental, health and safety measurements: needs, gaps and opportunities. Nanotoxicology 7, 1325-1337.

Sushko, I., Novotarskyi, S., Korner, R., Pandey, A.K., Cherkasov, A., Lo, J.Z., Gramatica, P., Hansen, K., Schroeter, T., Muller, K.R., Xi, L.L., Liu, H.X., Yao, X.J., Oberg, T., Hormozdiari, F., Dao, P.H., Sahinalp, C., Todeschini, R., Polishchuk, P., Artemenko, A., Kuz'min, V., Martin, T.M., Young, D.M., Fourches, D., Muratov, E., Tropsha, A., Baskin, I., Horvath, D., Marcou, G., Muller, C., Varnek, A., Prokopenko, V.V., Tetko, I.V., 2010. Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. J Chem Inf Model 50, 2094-2111.

Tantra, R., Oksel, C., Puzyn, T., Wang, J., Robinson, K.N., Wang, X.Z., Ma, C.Y., Wilkins, T., 2015. Nano(Q)SAR: Challenges, pitfalls and perspectives. Nanotoxicology 9, 636-642.

Taylor, R., 1995. Simulation Analysis of Experimental-Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. J Chem Inf Comp Sci 35, 59-67.

Tetko, I.V., Maran, U., Tropsha, A., 2017. Public (Q) SAR Services, Integrated Modeling Environments, and Model Repositories on the Web: State of the Art and Perspectives for Future Development. Mol Inform 36.

Tetko, I.V., Sushko, I., Pandey, A.K., Zhu, H., Tropsha, A., Papa, E., Oberg, T., Todeschini, R., Fourches, D., Varnek, A., 2008. Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: Focusing on applicability domain and overfitting by variable selection. J Chem Inf Model 48, 1733-1746.

Thomas, D.G., Gaheen, S., Harper, S.L., Fritts, M., Klaessig, F., Hahn-Dantona, E., Paik, D., Pan, S., Stafford, G.A., Freund, E.T., Klemm, J.D., Baker, N.A., 2013. ISA-TAB-Nano: A Specification for Sharing Nanomaterial Research Data in Spreadsheet-based Format. Bmc Biotechnol 13.

Thomas, D.G., Klaessig, F., Harper, S.L., Fritts, M., Hoover, M.D., Gaheen, S., Stokes, T.H., Reznik-Zellen, R., Freund, E.T., Klemm, J.D., Paik, D.S., Baker, N.A., 2011. Informatics and standards for nanomedicine technology. Wires Nanomed Nanobi 3, 511-532.

Todeschini, R., Ballabio, D., Grisoni, F., 2016. Beware of Unreliable Q2! A Comparative Study of Regression Metrics for Predictivity Assessment of QSAR Models. J Chem Inf Model 56, 1905-1913.

Topliss, J.G., Costello, R.J., 1972. Chance Correlations in Structure-Activity Studies Using Multiple Regression-Analysis. J Med Chem 15, 1066-&.

Topliss, J.G., Edwards, R.P., 1979. Chance Factors in Studies of Quantitative Structure-Activity-Relationships. J Med Chem 22, 1238-1244.

Toropov, A.A., Rallo, R., Toropova, A.P., 2015. Use of Quasi-SMILES and Monte Carlo Optimization to Develop Quantitative Feature Property/Activity Relationships (QFPR/QFAR) for Nanomaterials. Curr Top Med Chem 15, 1837-1844.

Toropov, A.A., Toropova, A.P., 2015. Quasi-QSAR for mutagenic potential of multi-walled carbon-nanotubes. Chemosphere 124, 40-46.

Toropov, A.A., Toropova, A.P., Benfenati, E., Gini, G., Puzyn, T., Leszczynska, D., Leszczynski, J., 2012. Novel application of the CORAL software to model cytotoxicity of metal oxide nanoparticles to bacteria Escherichia coli. Chemosphere 89, 1098-1102.

Toropova, A.P., Toropov, A.A., Benfenati, E., Gini, G., Leszczynska, D., Leszczynski, J., 2011. CORAL: QSPR models for solubility of [C-60] and [C-70] fullerene derivatives. Mol Divers 15, 249-256.

Toropova, A.P., Toropov, A.A., Benfenati, E., Korenstein, R., 2014. QSAR model for cytotoxicity of SiO2 nanoparticles on human lung fibroblasts. J Nanopart Res 16.

Toropova, A.P., Toropov, A.A., Rallo, R., Leszczynska, D., Leszczynski, J., 2015. Optimal descriptor as a translator of eclectic data into prediction of cytotoxicity for metal oxide nanoparticles under different conditions. Ecotox Environ Safe 112, 39-45.

Tropsha, A., 2010. Best Practices for QSAR Model Development, Validation, and Exploitation. Mol Inform 29, 476-488.

Tropsha, A., Gramatica, P., Gombar, V.K., 2003. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. Qsar Comb Sci 22, 69-77.

Visser, U., Abeyruwan, S., Vempati, U., Smith, R.P., Lemmon, V., Schurer, S.C., 2011. BioAssay Ontology (BAO): a semantic description of bioassays and high-throughput screening results. Bmc Bioinformatics 12.

Vriens, H., Mertens, D., Regret, R., Lin, P., Locquet, J.P., Hoet, P., 2017. Case Study III: The Construction of a Nanotoxicity Database - The MOD-ENP-TOX Experience. Adv Exp Med Biol 947, 325-344.

Walkey, C.D., Olsen, J.B., Song, F.Y., Liu, R., Guo, H.B., Olsen, D.W.H., Cohen, Y., Emili, A., Chan, W.C.W., 2014. Protein Corona Fingerprinting Predicts the Cellular Interaction of Gold and Silver Nanoparticles. Acs Nano 8, 2439-2455.

Wehrens, R., Putter, H., Buydens, L.M.C., 2000. The bootstrap: a tutorial. Chemometr Intell Lab 54, 35-52.

Weininger, D., 1988. Smiles, a Chemical Language and Information-System .1. Introduction to Methodology and Encoding Rules. J Chem Inf Comp Sci 28, 31-36.

Winkler, D.A., 2016. Recent advances, and unresolved issues, in the application of computational modelling to the prediction of the biological effects of nanomaterials. Toxicol Appl Pharm 299, 96-100.

Wirnitzer, U., Herbold, B., Voetz, M., Ragot, J., 2009. Studies on the in vitro genotoxicity of baytubes®, agglomerates of engineered multi-walled carbon-nanotubes (MWCNT). Toxicology Letters 186, 160-165.

Wold, S., Ruhe, A., Wold, H., Dunn, W.J., 1984. The Collinearity Problem in Linear-Regression - the Partial Least-Squares (Pls) Approach to Generalized Inverses. Siam J Sci Stat Comp 5, 735-743.

Wood, D.J., Carlsson, L., Eklund, M., Norinder, U., Stalring, J., 2013. QSAR with experimental and predictive distributions: an information theoretic approach for assessing model quality. J Comput Aid Mol Des 27, 203-219.

Worth, A., Aschberger, K., Asturiol Bofill, D., Bessems, J., Gerloff, K., Graepel, R., Joossens, E., Lamon, L., Palosaari, T., Richarz, A., 2017. Evaluation of the Availability and Applicability of Computational Approaches in the Safety Assessment of Nanomaterials, EUR 28617 EN, Publications Office of the European Union, Luxembourg.

Worth, A.P., Bassan, A., Gallegos, A., Netzeva, T.I., Patlewicz, G., M., P., Tsakovska, I., M., V., 2005. The Characterisation of (Quantitative) Structure-Activity Relationships: Preliminary Guidance. European Commission, Joint Research Centre, Ispra.

Worth, A.P., Cronin, M.T.D., 2001. The use of bootstrap resampling to assess the uncertainty of Cooper statistics. Atla-Altern Lab Anim 29, 447-459.

Xia, T., Kovochich, M., Liong, M., Madler, L., Gilbert, B., Shi, H.B., Yeh, J.I., Zink, J.I., Nel, A.E., 2008. Comparison of the Mechanism of Toxicity of Zinc Oxide and Cerium Oxide Nanoparticles Based on Dissolution and Oxidative Stress Properties. Acs Nano 2, 2121-2134.

Yang, H., Liu, C., Yang, D.F., Zhang, H.S., Xi, Z.G., 2009. Comparative study of cytotoxicity, oxidative stress and genotoxicity induced by four typical nanomaterials: the role of particle size, shape and composition. J Appl Toxicol 29, 69-78.

Yasri, A., Hartsough, D., 2001. Toward an optimal procedure for variable selection and QSAR model building. J Chem Inf Comp Sci 41, 1218-1227.

Zhang, H., Ji, Z., Xia, T., Meng, H., Low-Kam, C., Liu, R., Pokhrel, S., Lin, S., Wang, X., Liao, Y.-P., Wang, M., Li, L., Rallo, R., Damoiseaux, R., Telesca, D., Mädler, L., Cohen, Y., Zink, J.I., Nel, A.E., 2012. Use of Metal Oxide Nanoparticle Band Gap To Develop a Predictive Paradigm for Oxidative Stress and Acute Pulmonary Inflammation. Acs Nano 6, 4349-4368.