Representing and describing nanomaterials in predictive nanoinformatics

Wyrzykowska, Ewelina

2022-09

Wyrzykowska , E , Mikolajczyk , A , Lynch , I , Jeliazkova , N , Kochev , N , Sarimveis , H , Doganis , P , Karatzas , P , Afantitis , A , Melagraki , G , Serra , A , Greco , D , Subbotina , J , Lobaskin , V , Banares , M A , Valsami-Jones , E , Jagiello , K & Puzyn , T 2022 , ' Representing and describing nanomaterials in predictive nanoinformatics ' , Nature Nanotechnology , vol. 17 , no. 9 , pp. 924-932 . https://doi.org/10.1038/s41565-022-01173-6

http://hdl.handle.net/10138/356290
https://doi.org/10.1038/s41565-022-01173-6

acceptedVersion

Representing and describing nanomaterials in predictive nanoinformatics

Article *in* Nature Nanotechnology · August 2022

DOI: 10.1038/s41565-022-01173-6

18 authors, including:

Ewelina Wyrzykowska
QSAR Lab
9 PUBLICATIONS **75** CITATIONS
SEE PROFILE

Alicja Mikolajczyk
University of Gdansk
42 PUBLICATIONS **894** CITATIONS
SEE PROFILE

Iseult Lynch
University of Birmingham
392 PUBLICATIONS **28,142** CITATIONS
SEE PROFILE

Nina Jeliazkova
Ideaconsult Ltd
119 PUBLICATIONS **4,121** CITATIONS
SEE PROFILE

nature nanotechnology

REVIEW ARTICLE

https://doi.org/10.1038/s41565-022-01173-6

Representing and describing nanomaterials in predictive nanoinformatics

Ewelina Wyrzykowska[1,14], Alicja Mikolajczyk[1,2,14], Iseult Lynch[3], Nina Jeliazkova[4], Nikolay Kochev[4,5], Haralambos Sarimveis[6], Philip Doganis[6], Pantelis Karatzas[6], Antreas Afantitis[7], Georgia Melagraki[8], Angela Serra[9,10], Dario Greco[9,10,11], Julia Subbotina[12], Vladimir Lobaskin[12], Miguel A. Bañares[13], Eugenia Valsami-Jones[3], Karolina Jagiello[1,2] and Tomasz Puzyn[1,2] ✉

Engineered nanomaterials (ENMs) enable new and enhanced products and devices in which matter can be controlled at a near-atomic scale (in the range of 1 to 100 nm). However, the unique nanoscale properties that make ENMs attractive may result in as yet poorly known risks to human health and the environment. Thus, new ENMs should be designed in line with the idea of safe-and-sustainable-by-design (SSbD). The biological activity of ENMs is closely related to their physicochemical characteristics, changes in these characteristics may therefore cause changes in the ENMs activity. In this sense, a set of physicochemical characteristics (for example, chemical composition, crystal structure, size, shape, surface structure) creates a unique 'representation' of a given ENM. The usability of these characteristics or nanomaterial descriptors (nanodescriptors) in nanoinformatics methods such as quantitative structure–activity/property relationship (QSAR/QSPR) models, provides exciting opportunities to optimize ENMs at the design stage by improving their functionality and minimizing unforeseen health/environmental hazards. A computational screening of possible versions of novel ENMs would return optimal nanostructures and manage ('design out') hazardous features at the earliest possible manufacturing step. Safe adoption of ENMs on a vast scale will depend on the successful integration of the entire bulk of nanodescriptors extracted experimentally with data from theoretical and computational models. This Review discusses directions for developing appropriate nanomaterial representations and related nanodescriptors to enhance the reliability of computational modelling utilized in designing safer and more sustainable ENMs.

The application of computational methods for the prediction of nanomaterial properties, including physics-based multiscale materials modelling and data-based modelling with artificial intelligence/machine learning, has great potential to accelerate the introduction of engineering nanomaterials (ENMs) into a range of advanced applications and to enhance their safe and more sustainable use at all stages of their life cycles[1–4]. Among the variety of nanoinformatics methods, data-driven quantitative structure–activity relationship (QSAR) and quantitative structure–property relationship (QSPR) models help materials scientists to gain a better grasp of the physicochemical characteristics that determine the desired functionality or are responsible for the adverse health effects of ENMs[5,6]. These relationships allow a multidimensional optimization of the material such that the useful properties are enhanced while the properties of concern are designed out at the earliest possible step (virtual design before synthesis).

In nanoinformatics QSAR modelling (nano-QSAR) that utilizes machine learning and artificial intelligence, the predicted response (for example, the physicochemical property, biological activity or toxicity of interest) is modelled by a set of predictors representing various ENM characteristics termed nanomaterials descriptors or nanodescriptors. It is also possible, especially in the case of sequential modelling (for example, by structure–activity prediction networks, SAPNet)[7], that the same nanodescriptor (for example, the zeta potential) can serve as a predictor (for example, a predictor of cellular uptake) in one model and the property of interest (the predicted response) in another (Fig. 1).

Given that even subtle changes in the characteristics of ENMs can cause changes in the property of interest, it is crucial to appropriately define the nanodescriptors. They should reflect not only the ENMs' chemical composition and the chemical structures of its components, but also other important physicochemical characteristics related to the nanomaterial as a whole (for example, size or shape). Moreover, the nanodescriptors should reflect the influence of the system (that is, the environmental or experimental conditions) on the properties of the ENMs. One can thus make a distinction between system-independent (intrinsic) and system-dependent (extrinsic) nanodescriptors.

As at present nanodescriptors can come from either the experimental characterization or from theoretical calculations, appropriate ENM characterization is a challenge faced both by experimentalists and nanoinformaticians. In experimental characterizations, there

[1]QSAR Lab Ltd, Gdańsk, Poland. [2]Laboratory of Environmental Chemoinformatics, Faculty of Chemistry, University of Gdańsk, Gdańsk, Poland. [3]School of Geography, Earth and Environmental Sciences, University of Birmingham, Edgbaston, Birmingham, United Kingdom. [4]Ideaconsult Ltd, Sofia, Bulgaria. [5]Department of Analytical Chemistry and Computer Chemistry, University of Plovdiv, Plovdiv, Bulgaria. [6]School of Chemical Engineering, National Technical University of Athens, Zografou, Athens, Greece. [7]NovaMechanics Ltd, Nicosia, Cyprus. [8]Division of Physical Sciences and Applications, Hellenic Military Academy, Vari, Greece. [9]FHAIVE, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. [10]BioMediTech Institute, Tampere University, Tampere, Finland. [11]Institute of Biotechnology, University of Helsinki, Helsinki, Finland. [12]School of Physics, University College Dublin, Belfield, Dublin, Ireland. [13]Instituto de Catálisis y Petroleoquimica, ICP CSIC, Madrid, Spain. [14]These authors contributed equally: Ewelina Wyrzykowska, Alicja Mikolajczyk. ✉e-mail: tomasz.puzyn@ug.edu.pl

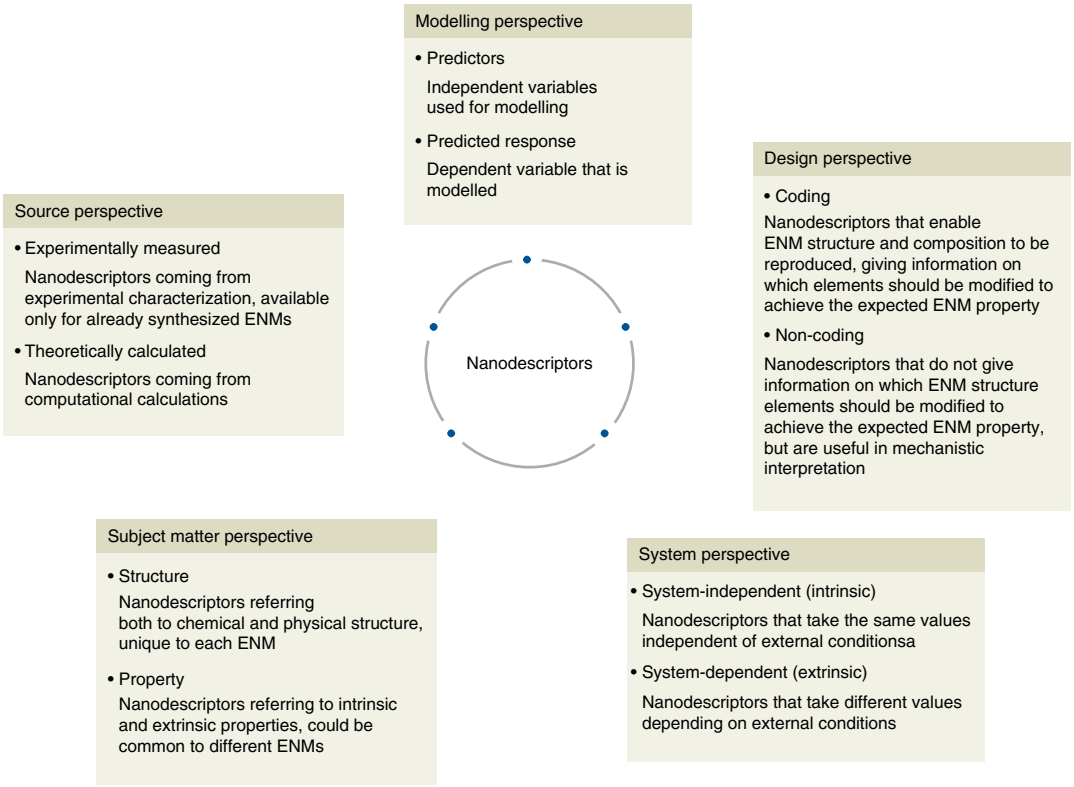

Fig. 1 | The spectrum of perspectives on nanomaterials descriptors. The figure illustrates viewpoints from the contexts of modelling, design, the behaviour of ENMs in a system, the subject matter of encoded ENM features and the sources from which nanodescriptors are obtained.

are constantly implemented new parameters and measurement modes with transparent procedures enabling reproducibility of the work, analytical accuracy, and resolution; as well as requirements for reporting formats that include metadata, data gathering, and storage tools. The availability of metadata is also an important issue in assessing the quality and completeness of experimental nanodescriptors[8,9]. Nanodescriptors from experiments conducted with different quality requirements and/or according to different protocols can introduce an additional, substantial source of uncertainty. It is possible that this variation in a group of nanomaterials could be even higher than the variation related to the real differences in the characterized properties. This, of course, could make the nanodescriptors useless for nano-QSAR modelling. It is also worth noting that a range of results could be obtained depending on the applied descriptors expressing volume, mass, surface area or the quantity of nanoparticles in the sample.

In turn, nanoinformaticians are faced with building simplified (supra)molecular structure models (that is, atomic clusters, surface fragments) that could be used in physics-based materials modelling to obtain knowledge of the electron structure and related properties of investigated ENMs. The complexity of aspects that must be considered—which are mostly related to cluster size (a reflection of the actual size of the ENM or part thereof, such as a unit cell), the character of multicomponent ENMs (doping, coating, shell composition and thickness) and surface properties (the presence of defects, charge, porosity, roughness)—is virtually endless. The challenge of building appropriate molecular models of ENMs limits the possibilities of developing newly calculated (purely theoretical) descriptors to be used in nano-QSAR for predicting the properties of not-yet-synthesized ('virtually created') ENMs.

However, independently of the source (either experimental measurements or theoretical calculations), these nanodescriptors should be collectively unique for a particular ENM and act as a 'fingerprint' enabling ENMs with the same chemical composition but different physical characteristics to be differentiated. The achievement of such a distinction between compositionally similar ENMs is essential because, owing to physical differences, these ENMs can exhibit notable variation in the properties and toxicological responses they induce[10]. Considering that in the case of ENMs size and other morphological aspects may be even more important than their chemical composition, it is relevant to include these physical characteristics in nano-QSAR models.

This seemingly inevitable need to transition towards nanodescriptors as representing unique fingerprints also highlights the distinction between 'coding' versus 'non-coding' ones. Coding nanodescriptors are those for which the ENM composition, component structures and other characteristics can be precisely reproduced (that is, size, chemical identity). In contrast, non-coding descriptors are those that do not offer such a possibility; more than one nanoparticle may be characterized by the same value of a given property, and it is not possible to deduct the identity from the property. Examples of non-coding nanodescriptors include solubility, lipophilicity and bandgap energy calculated with quantum mechanical methods[11]. Non-coding descriptors, however, also provide important links to the extrinsic properties of ENMs[12]. Analysing descriptors and categorizing them as coding or non-coding might provide insights into whether the predicted response (for example, toxicity) driver is dependent or independent of the material. This will provide hints regarding how to directly or indirectly design out the specific ENM feature linked to toxicity, thereby feeding into safer design strategies for ENMs.

One important feature of nanodescriptors, especially in investigating the toxicity of more complex structures, is interpretability. It is defined as the ability to elucidate and rationalize the underlying nanodescriptors responsible for biological behaviour[13], which needs to be translated to ease of use for the rational design of improved

ENMs and understanding of the mechanisms of the biological properties[14]. Slow progress in developing interpretable descriptors is one of the limiting factors in progress towards earlier-identified ENM milestones[14], and more recent milestones highlighted in the EU-US nanoinformatics roadmap 2030[15]. Progress on these fronts is necessary to advance ENM knowledge and applications. Extensive consultations have led to the understanding that nanodescriptors and the ontology used to depict them are interpreted differently not only by modellers and experimentalists, but even within the modelling community itself. As the challenge of appropriate representation of ENMs can only be met by close cooperation between experimentalists and theoreticians, there should be a common understanding of the basic terms and needs to facilitate more reliable nanoinformatics modelling for safe-and-sustainable-by-design (SSbD) ENMs.

An adequate computer representation of the objects (entities) is required to handle biological, chemical or ENM information and enable information systems to be built. A representation that does not reflect a particular material aspect cannot convey that aspect to the machine learning model. As an analogy with cheminformatics, if a computer representation does not include stereochemical information, any machine learning model built on this data will also ignore the stereochemistry. The development of descriptors and material representations is a dynamic field, especially in the deep learning era, and deserves its own state-of-the-art review.

In this Review, we aim to show the need for developing a comprehensive system of nanomaterials descriptors to be used in nanoinformatics. The discussion is based on a newly proposed conceptual model of ENM representation in line with the terminology used in recent regulations. By doing this, we also intend to contribute to the ongoing regulatory discussion on defining nanomaterials.

How the regulatory context affects the representation of ENMs

Chemical safety assessment in Europe is regulated by a set of legal acts[16–19], the most important of which is the REACH regulation (EC 1906/2006) that refers to the term 'chemical substance'. However, the definition of 'substance' is ambiguous when considering ENMs. According to REACH, a substance is identified by the main (≧80%) constituent. Overall, the concept of substance is sophisticated: the 'substance data model' is defined by a combination of many components (for example, molecules) playing different roles, namely: major chemical constituent(s), impurities (unintentional constituents originating from the manufacturing process or from the starting materials) and additives (intentionally added to improve the properties of the substance). When the component is known, from the nanoinformatics point of view it can be represented by a triad adopted from the classical chemoinformatics (data-based modelling of regular chemicals): (1) molecular structure (the way the atoms are connected and/or configured in space); (2) molecular descriptors, which constitute "the result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment"[12] (for example, descriptors indicating the number of particular functional groups, topological indices, energies of frontier orbitals and so on); and (3) molecular properties (such as the *n*-octanol/water partition coefficient[20] or aqueous solubility of the compound). Moreover, the substance (ENM in this case) itself can be characterized by its own properties (differing from these of its individual components), including its density, refractive index and the electric conductivity of the substance. These have to be complemented by the related metadata describing the experimental conditions and protocols[21]. Another very important piece of information concerning ENMs is the potential links between individual components of the ENM (for example, if a second coating layer is connected with covalent bonds to the first coating layer). In this context, the term composition (as also used in REACH) describes the ENM components and their relations, and should not be mistaken with the term chemical composition, which is typically associated with elemental analysis[12]. In the context of ENMs, the component identifier used by REACH (that is, the Chemical Abstract Service (CAS number) does not include a size consideration. Thus, all particulate forms with a composition of ≧80% TiO_2, regardless of their particle size or morphological characteristics, would have the same substance ID. Identifying the distinct crystalline phases of TiO_2 (that is, anatase, rutile or brookite) does not solve the challenge; these are infinite crystalline patterns and ratios. Material is an entity limited in space, and with multiple features, combinations of crystalline domains, exposed phases, surface and bulk defects, different surface terminations and so on. According to the original REACH definition, different ENMs with a single main composition would be classified as the same substance.

To address this definition gap, the recent amendment to the REACH regulation (EC No 1907/2006)[22], which entered into force on 1 January 2020, introduces the 'nanoform' of the substance to distinguish nanoscale particles of a specific substance that may also exist as larger particles or as a molecular chemical, such as an ion. Nanoform is defined as: "a form of a natural or manufactured substance containing particles, in an unbound state or as an aggregate or as an agglomerate and where, for 50% or more of the particles in the number size distribution, one or more external dimensions is in the size range 1 nm–100 nm, including also by derogation fullerenes, graphene flakes and single wall carbon nanotubes with one or more external dimensions below 1 nm"(ref. [23]). The same substance may thus have one or more different nanoforms, differing in one or several characteristics such as size distribution, number fraction of constituents, surface treatment, shape, specific surface area and other morphological, structural or chemical characteristics. Owing to the differences in the physicochemical properties of nanoforms (for example, surface chemistry), various nanoforms of the same substance are likely to have different activity (toxicity) profiles[24]. Moreover, these physicochemical properties of nanoforms, which are potentially important for their biological activity (for example, agglomeration ability), can be influenced by the conditions of the external system (for example, the dispersion medium, pH, the presence of proteins and other biomolecules)[25–27]. In a nanoinformatics context, these 'system-dependent' properties must always be accompanied by appropriate metadata describing the system and the protocols followed in the characterization of the ENM. By this legal definition, two ENMs are considered either as two nanoforms of the same substance (when they differ in the aforementioned characteristics, for example, size, shape, number fraction of constituents) or as two separate substances (when they differ by molecular descriptors of the main constituent, for example, core composition). Therefore, the ENM representation should go beyond the component or constituent's characteristic and also include the detailed characteristics of the nanoform, which are important in the regulatory context (Fig. 2).

From a regulatory perspective, each nanoform must be characterized experimentally, following Annex VI section 2.4 of REACH to prepare the registration dossier. However, from the perspective of scientists and those in industry who design and manufacture ENMs, new approaches are needed to enable built-by-design and/or safer-by-design strategies for ENMs. Thus, it would be beneficial to adapt QSAR/QSPR methodology to enable virtual screening of ENMs to identify the most promising nanoforms (those with high application potential and low toxicity), expressed in terms of appropriate descriptors. Such models would enable the user to manipulate the compositional and structural parameters very precisely to generate the required effect in the properties, activity and toxicity of the ENM and establish the boundary conditions for such behaviour[28–30].

In turn, the US Environmental Protection Agency's Toxic Substances Control Act (TSCA)[31,32] defines a 'chemical substance' as "any organic or inorganic substance of a particular molecular identity,

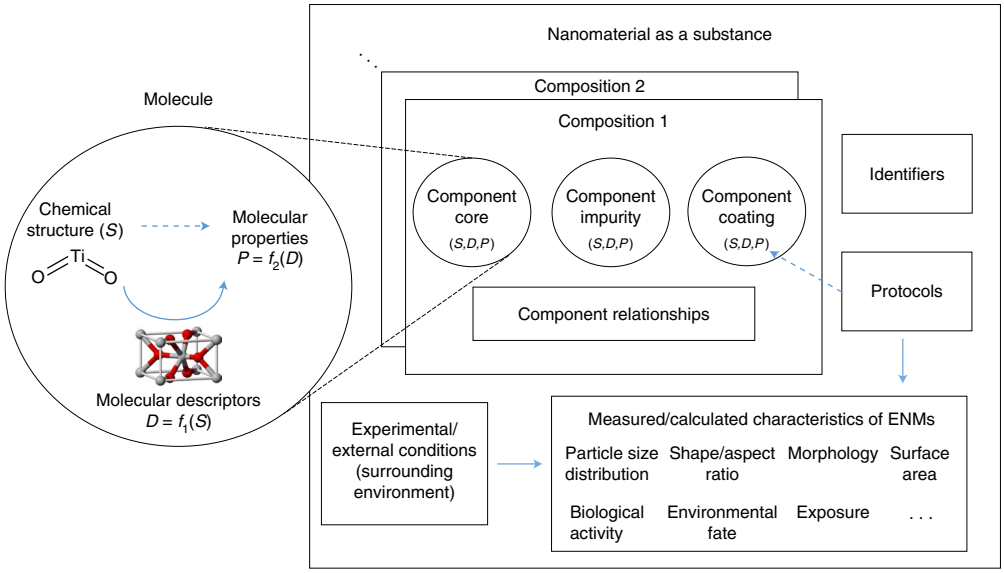

Fig. 2 | The concept of ENM representation. This concept combines the substance identity concept with the definition of a nanoform, adapted from the classic substance paradigm[21], and includes unique identifiers for the ENMs' composition. These consist of the core components, impurities and coatings and their relationships, as well as their measured or calculated characteristics supported with metadata related to protocols and experimental conditions. The symbols '*S*', '*D*' and '*P*' code chemical structure, molecular descriptors and molecular properties, respectively. The symbol '*f*' codes the mathematical function of relationships between the above mentioned elements of the chemical triad.

including any combination of these substances occurring in whole or in part as a result of a chemical reaction or occurring in nature, and any element or uncombined radical". The TSCA approach is analogous to the REACH regulation, addressing multiple components within a substance. Similarly to REACH, the TSCA also regards ENMs as a particular type of 'chemical substance' requiring additional information to be reported, such as the specific chemical identity, methods of manufacture/processing, use scenarios, exposure, safety data and so on. In this context, the data model described for ENM representation is flexible enough and could benefit both regulatory frameworks and others that are now evolving. Consequently, the descriptors calculated on top of the data model from Fig. 2 may fit the purposes of REACH and TSCA regulations alike.

All regulatory frameworks suffer from a lack of unique nano-identifiers as traditional identifiers such as CAS numbers or the most popular linear notations are inadequate. Developing ENM identifiers is a challenging problem that is not yet solved; however, a promising effort in this direction is performed by the International Union of Pure and Applied Chemistry nano-InChI working group[33,34]. Nevertheless, nano-InChI will also not solve everything but aims to integrate concepts of ENM provenance and intrinsic versus extrinsic properties and to support a domain-specific language for nanoinformatics. It is not likely that all of the nanomaterial information can be stored within nano-InChI notation; hence the descriptors calculated solely from nano-InChI should not be expected to be efficient enough for predictive nanoinformatics tasks compared with the descriptors obtained by the full data model depicted in Fig. 2. Similarly, a promising direction could be quasi-SMILES for nanomaterials that carry information not only about the chemical composition but also about experimental conditions and changes in physicochemical properties after exposure[35,36], but further development would be required.

Generally, the regulatory frameworks (REACH, TSCA and others) are slowly and cautiously progressing into the nano arena—for example, the required nano-identity information from TSCA is quite generally described, and no strict requirements are specified. The latter is due to many factors, including the ongoing discussions and lack of full consensus on those properties of ENMs that are predictive of toxicity. In this regard, having a more universal, flexible and dynamic data model might increase the potential to handle future transitions in the nanoinformatics field. This is the main objective of the data model in Fig. 2.

The proposal of the data model for ENM representation is pragmatic to enable nanoinformatics support for materials beyond single chemical structures (ENMs, advanced materials). It is based on experience and collaboration with industry and many experts who reviewed the relevant regulations in various contexts. Finally, it should serve as a map for developing a comprehensive system of nanodescriptors.

Five challenges in developing nanodescriptors

Challenge 1: describing the component relationships in multicomponent ENMs. The first relevant aspect of enhancing the description of nanomaterials is related to the complexity of their compositions. So far, substantial progress in nano-QSAR modelling has been made mainly with one-component nanostructures or attributes of nanoforms of the substance expressed through its individual constituents (for example, core, shell layers, doping, structure modification, impurities, additives and so on). However, most ENMs have a diversified functionality resulting from several components, including different shells or coatings, attached surface modifiers, bimetallic or multi-metallic cores and so on, rendering them heterogeneous materials in the sense of the chemical composition (more about ENM heterogeneity in the sense of physicochemical properties is described under Challenge 3 below). The individual components in the nanoform of the substance may show joint toxicity effects that may be closely related to the ideal additive behaviour of response/effect and/or increased (synergistic) or decreased (antagonistic) effects. As a result, a range of outcomes in relation to the mechanism of action may be observed. Specifically, the same mode of action as for individual chemicals (concentration addition); a distinct mode of action as for individual chemicals (independent action); or multifaceted and considerable deviations in the apparent properties of its individual components (a synergistic or antagonistic mode of action).

Table 1 | Overview of challenges to be faced and potential directions for resolving them

Challenge number	Description	Starting point	Future directions	Reference(s)
1	Describing the component relationships in multicomponent ENMs	Descriptors reflecting additive relationships between ENM cores and modifiers	Descriptors reflecting synergistic and antagonistic relationships between ENM components	37–39
2	Representing the influence of external conditions (surrounding environment) and different protocols	Descriptors act as the linkage between ENM properties in different conditions	Further extension and development of nano-InChI and quasi-SMILES	35,36,44,76,77
3	Considering the distributions of nanodescriptors, rather than the average values	Vectors of d_{10}, d_{50} and d_{90} values reflecting ENM size distributions	Descriptors reflecting the entire distribution instead of single points that assume a normal distribution	Not yet applied in nano-QSAR/QSPR models
4	Developing nanodescriptors for virtual screening	Theoretical descriptors based on the Pauling atom electronegativity and the Delaunay tessellation approach	Theoretical descriptors considering the influence of the external conditions	60
5	Increasing the interpretability of nanodescriptors by coupling nano-QSARs with adverse outcome pathways (AOPs)	Classical descriptors used for modelling key events	Algorithm for selecting descriptors that are better interpretable in the context of AOPs	65,67,71

d_{10} is the size for which 10% of the whole set of ENMs have a size less than this value, d_{50} is the median size of ENMs and d_{90} is the size for which 90% of the ENMs have a size below this value.

In an early attempt to predict the properties of multicomponent ENMs[37] the authors introduced 'additive descriptors', which are a linear combination of descriptors for pure components weighed by their concentrations. In this example, the catalytic efficiency of TiO_2-based surface photocatalysts modified by (poly-) metallic clusters of four metals (Au, Pd, Pt, Ag) was predicted by combining descriptors that characterize the metals weighted by their concentration. This concept was based on a strategy originally developed for the additive mechanism for mixtures of conventional chemicals. The newly developed methodology was then combined with periodic table-based descriptors and applied to improve the cytotoxicity models of the TiO_2-based ENMs modified with Ag/Au/Pd/Pt[38,39]. Nevertheless, there is still a need for a better understanding of the influence of individual ENM components and surface modifiers on the finally observed ENM features because the joint effects are not always additive; they can also be synergistic or antagonistic (Table 1).

Here it might be useful to think about the roles and functions of the different ENM constituents and whether they are 'relevant' to the predicted toxic response. For example, let us consider a polymer coating on an ENM: it may be there to provide steric stabilization against agglomeration and potentially to reduce protein binding and subsequent cellular attachment (such as polyethylene glycol, although reduced binding may depend on the degree of surface coverage of the polymer coating; that is, there may still be a strong influence from the underlying core on the protein binding). Thus a strategy that includes molecular descriptors for the polymer coating and nanodescriptors for the underlying core, and allows their relative weighting to be adjusted, might enhance the predictive power in terms of corona formation, cellular attachment uptake and eventual toxicity. For many functional multicomponent nanomaterials, their properties are not linear combinations of the properties of the components: for example, in catalysis, a combination of oxides may generate reactive properties that are not present in the individual components, and that new reactivity may, in turn, cause new modes of action that cannot be linearly inferred. Further development in this area is therefore still required, especially for cases where particular components act independently (according to different mode of actions) or their mode of action is synergistic or antagonistic. Moreover, further studies should also consider the influence of impurities on the predicted response.

Challenge 2: representing the influence of the surrounding environment and different protocols. An appropriate representation of the surrounding environment influence is especially important when one considers the toxicity of ENMs as the predicted response. Toxicity is thought to originate from unique characteristics of the pristine, individual components. In reality, however, the composition and structural characteristics that determine an ENM's properties may change during its lifetime depending on external conditions[25,40–42]. As a result, the same ENM may be safe or hazardous for the human body and the environment, depending on the external conditions (the surrounding environment)[40]. The nanodescriptors used must therefore be system-dependent.

The ENM composition (that is, core, coating, layers composed of the protein corona[43], lipids and so on) may be considered as its fingerprint in a given environment that strongly determines the translated unique properties into application potential and, simultaneously, toxicological potential. However, at the same time, the composition is not stable and may change as the ENM is transported through different environments. ENM properties may consequently vary widely before, during and after exposure to biological serum or environmental media. The REACH approach for treating substance data supports multiple compositions (see Fig. 2), which fits well with the idea that ENM composition and properties may progressively change (that is, ENM dynamics) and hence could be utilized in nanoinformatics software tools for handling ENMs.

The majority of already developed nano-QSAR models use nanodescriptors that refer to the composition, component structures and properties measured or calculated under a well-defined, unchanging set of conditions. This approach allows the relationships between a nanoform with a specific composition/structure/morphology and a given toxic response to be recognized. To address this challenge, the scientific community[44–46] proposed an extension of classic nano-QSAR models to consider both 'classic' intrinsic nanodescriptors and extrinsic nanodescriptors that characterize composition/structure concerning the surrounding conditions (that is, extrinsic or system-dependent descriptors).

Examples of ENM behaviour varying according to the surrounding environment are agglomeration or aggregation; both refer to the process by which primary nanoparticles form a group of secondary particles, but agglomerates rely on weak interaction forces that can disintegrate in bioenvironments, whereas aggregates result

from strong forces such as covalent or metallic bonds[47]. Both processes obviously alter the average size and the size distribution but also reduce the available surface area and result in different overall shapes, which is easily visible for spherical ENMs or for bundling/unbundling of nanotubes, for example[26]. It should be noted that post-synthesis treatments such as centrifugation can affect size distributions due to particle aggregation[28], so any evaluation of ENMs should take into consideration the form in which they will actually be used, and the handling steps (the so-called provenance of the ENM)[48]. Overall, the impact of the agglomeration or aggregation state is diverse[47]. Although it is believed to substantially reduce human exposure in the environment, once within biosystems (after entry or internal formation), agglomerates/aggregates typically have high toxic potentials. At the same time, whereas in most cases the reduction in a specific surface area leads to reduced nano–cell interactions, some studies report increased adverse effects[47]. New specific descriptors for the aggregation state have recently been proposed: the aggregation free energy and two numerical parameters used to correct for the observed deviation from the aggregation kinetics described by the Smoluchowski theory[49]. These descriptors are generated as a result of a coarsening strategy combining ab initio density functional theory and molecular and Brownian dynamics.

Another example of a relevant process that ENMs undergo is protein corona formation in the biological environment. Recently we have shown that protein adsorption energies can be predicted directly from the protein and nanomaterial descriptors by using an artificial neural networks model[50]. The artificial neural network model invoked for Au and TiO_2 surfaces removed the need for the expensive parameterization protocol of the united-atom multiscale model, which is used for obtaining *in silico* adsorption energies for the protein as a whole. Moreover, in another recent work[51], it was shown that the bio–nano interaction descriptors themselves could be derived from the nanoparticle structure and composition using multiscale computational models without any experimental input. The parameterization of the united-atom model would usually include a set of biased all-atom molecular dynamics simulations of the adsorption of amino acid fragments onto nano surfaces[20]. Indeed, the obtained binding energies for individual proteins can be further utilized to predict the properties of the protein corona[52]. This already enables us to construct predictive models of the biological activity of ENMs starting from the ENM composition and connecting it to the biological endpoint via an ENM-specific corona. This approach has been tested on a limited set of ENMs, and further work and testing are needed to expand the scope of this method (in progress).

We also note the importance of ENM provenance and the metadata describing the protocols used. These data include the origin or source of a batch of nano-objects along with information related to handling and any changes that may have taken place since they were originated[48], and are critical to the quality of nanoinformatics models, especially those considering dynamic evolution. Previous work to capture ENM provenance, such as the CODATA Universal description system proposed by Committee on Data for Science and Technology, which has become ASTM standard created within American Society for Testing and Materials, is now being integrated with nano-InChI and its auxiliary information as a step in this direction, although this will not fully adress the need for complete metadata reporting.

Challenge 3: considering distributions of nanodescriptors, rather than the average values. A key feature of real ENMs is their heterogeneity in the context of physicochemical characteristics[53]. Variations in the structural features of ENMs are thought to be responsible for the nonlinear effects and toxic responses after exposure to subpopulations of ENMs[54]. Although the main components of ENMs could be identified and declared harmless, parts of ENMs with different structural characteristics could induce observed effects; for example, nanoparticles much smaller than the average size can more efficiently undergo cellular uptake and cause cell dysfunction[55].

We are used to thinking only about size distributions, but in reality all properties are likely to be heterogeneous across the population of ENMs, including the distributions of surface charges, defects in the crystal structure, coating thickness and coverage of the surface with functional groups or polymeric stabilizers and so on. However, when reporting the characteristics, we usually provide the mean plus standard deviation or the mean value. Similarly, toxicity data are reported as the mean (per cell/organism and so on) averaged over the whole exposed population of cells. Is information on the homogeneity of the sample lost in this averaging? And could the true distinction between nanoforms and the true predictors for toxicity lie in the details that get lost by averaging? An important next step will be to consider whether we can move towards ENM descriptors as distributions or vectors that allow for variability of the specific property within a 'dose' of ENMs and interrogate the effect of small changes in the descriptors to be used in the establishment of nano-QSARs. For example, an ENM dispersion with a mean size of 50 nm and size distribution of ±15 nm may correspond to 50% of the particles having a size of 40 nm and 50% of the particles having a size of 60 nm, behaving as two completely different populations, which would be lost by averaging. Similarly, suppose there are on average 10 functional groups per ENM over 100 particles. In that case, there could be some particles with no functional groups, some with 20, and these could all be trafficked into cells differently or acquire different biomolecular coronas and be distributed to entirely different organs—that is, have different toxicokinetics and toxicodynamics.

In the context of ENM size distributions, the European Chemical Agency recently introduced the requirement for reporting of size distribution through the values of d_{10}, d_{50} and d_{90} (ref. [56]). In turn, a collaboration between experts from European Nanomedicine Characterization Laboratory and the US National Cancer Institute Nanotechnology Characterization Laboratory resulted in a multi-step strategy of incremental complexity for the characterization of the size distribution and size stability of nanoparticle-enabled medicinal products[57]. This strategy assumed the use of orthogonal techniques where in the first attempt, low-resolution techniques of light-scattering and tracking analysis are used for preliminary screening of sample stability and integrity, then a combination of high-resolution microscopy techniques are used for size measurement in simple buffers and in much more complex biological media. This developed approach raises the issues of the reproducibility of the manufacturing (batch-to-batch variability), long-term dispersion stability during storage, nanoparticle size changes after administration and interactions with biomolecules in physiological media[57]. The implementation of this approach in experimental ENM characterization on a broader scale would be extremely beneficial for nano-QSAR/QSPR modelling.

In a further step, it would be beneficial to introduce descriptors that reflect ENM heterogeneity via standard statistical quantities (moments of distributions) and geometrical definitions. Possible quantitative higher-order descriptors include: ENM sphericity, aspect ratios, principal moments of inertia, components of the dipole moment, the percentage of hydrophobic surface or ligand/functionalization coverage, spherical harmonics (or Fourier modes) of non-uniform ligand grafting density or charge distribution, node amplitudes of the Schulz distribution of the size of the polydisperse materials and so forth.

On the other hand, the dynamic nature of ENMs and their transformations could make the use of a moving-average approach for the nanodescriptors appropriate to allow for changes in (for example) the size as a result of dissolution, to be factored in over-time. One parameter that is well known to evolve with time, and as

ENMs move into cells and between cellular compartments, is the biomolecular corona[58,59], and thus time-resolved corona data (proteins, metabolites, etc.) could be incorporated into moving-average ENM fingerprint descriptors to capture this evolution on the basis of binding affinities, local abundances and physiological conditions. These represent exciting areas of research where advances in nanoinformatics can open up new vistas in our understanding of nanosafety and nanomedicine.

Challenge 4: developing nanodescriptors for virtual screening. The concept of nanoforms moves us beyond the descriptors of ENMs' constituents and begins to demand more complex approaches such as full-particle descriptors or multicomponent descriptors (as discussed above), albeit with extensions for polymers, small molecule ligands and mixed or non-integer oxides. The influence of the external conditions (the environment) has also been highlighted. Thus, there is a need to develop a comprehensive system of descriptors following the concept of ENM representation, as presented in Fig. 2.

QSAR methods originate from pharma, where they are used at the early stage of pre-clinical studies in virtual screening for novel active substances that are then studied empirically. The strength of virtual screening lies in the possibility of investigating thousands of chemical structures created virtually (in a computer) without needing to synthesize them. In such a case, the modellers utilize only calculated and preferably coding descriptors as predictors of the property of interest. Implementation of virtual screening into SSbD strategies for ENMs would substantially reduce time, cost and use of laboratory animals because a number of considered solutions would be eliminated even without synthesis and experimental characterization. However, there is a challenge related to the fact that the majority of nanodescriptors in use in nano-QSAR at present need to be experimentally measured. On the other hand, the use of physics-based multiscale modelling methods (for example, quantum chemical methods, density functional theory, molecular dynamics) for calculations is very demanding given the necessity of using supercomputer power even for small (from a nanotechnology point of view) nanoparticles (that is, 1–2 nm). Thus, in practice, there is still a limited number of descriptors that can be calculated when characterizing the whole nanostructure (that is, without the use of simplified molecular models of the nanomaterials).

To this end, universal theoretical (calculated) nanodescriptors were proposed to characterize diverse nanostructures and be applied in predictive modelling and virtual screening of ENMs[60]. These descriptors are calculated by considering the Pauling atom electronegativity as an empirical variable to define descriptor characteristics and using the Delaunay tessellation approach to represent the surface chemistry by defining tetrahedra in the 3D representation of the ENM structure. This idea was further explored[61] and the frequency of occurrence of each tetrahedron was used to consider large differences between large-size and small-size ENMs, while 17 additional atomic properties were considered in the calculation of these nanodescriptors. However, these descriptors do not consider all characteristics that need to be accounted for when following the new concept of ENM representation. For example, they do not include the influence of external conditions (the surrounding environment; system-dependent nanodescriptors).

In this context, the descriptors that might be used to express system-dependent agglomeration and aggregation are those based on the liquid drop model[62,63]. Originally, the liquid drop model was designed as an assembly of molecules that form the nanoparticle. However, it also can be treated as an assembly of pristine nanoparticles that have formed an aggregate or agglomerate. In this case further work is needed: how the formation of agglomerate/aggregate differs depending on the external (system-dependent) conditions needs to be determined.

In the context of virtual screening, the applicability domain of the model is one more point that needs to be considered. The applicability domain is constrained to the data used in the process of the model calibration (training data). Thus, the nanodescriptors used in the model (that is, their type, number and the range of values) determine how extensive the applicability domain is for an appropriate model. The highest reliability of computational model predictions is obtained through interpolation, where the characteristics of new ENMs are similar in terms of nanodescriptors to the ENMs used in the training data. Usually, however, the virtual screening concerns the extrapolations, where the predictions go far beyond the applicability domain borders. The credibility of such predictions is lower, but not none. The challenge is to quantify the degree of uncertainty in the predicted values. The application of Bayesian machine learning methods has great potential in this respect, but so far these methods have not been applied in virtual screening of ENMs. Nevertheless, facing this challenge would substantially increase the role of nanoinformatics in SSbD.

Challenge 5: increasing interpretability of nanodescriptors by coupling nano-QSARs with adverse outcome pathways. An approach for predicting the progressive changes of ENMs properties is needed; it could be related to both the basic physicochemical properties (for example, dissolution) and to complex human or environmental pathways (for example, AOPs). Dissolution processes can be an important factor in understanding the bio-distribution and the cellular responses to a range of different ENMs[64]. They have the potential to become key information to be used in a screening process to group ENMs with a common hazard potential based on their potential to release ionic species[65]. Several approaches to this problem can be envisioned: (1) comparisons of bond energies with solvation energies for a given ion/atom/molecule, (2) kinetic models to assess the timescale of any dissolution, (3) biased molecular dynamics simulations of free-energy barriers to the dissolution of ENMs, including surface reconstruction and changes on contact with water and, where appropriate, (4) direct molecular dynamics studies of spontaneous dissolution and the influence of surface ligands and coronas.

The ability of an ENM to dissociate, catalyse a chemical reaction and produce reactive species, thus affecting the conformation of 'vital' biomolecules or interfering in metabolic or reproductive processes, also determines its potential to cause hazardous effects. From a biological point of view, this can be explained as inducing molecular initiating events leading to the initiation of an AOP. ENM properties profoundly affect molecular processes at the bio–nano interface. Known candidate molecular initiating events for ENMs include the production of radical oxygen species, cellular uptake and lysosomal damage[66]. At the same time, nanodescriptors for interactions of ENMs with lipids, lung or cell membranes, water or oil[20] or receptor proteins are scarce, and their evaluation requires substantial developmental work. Recent computational studies have demonstrated the possibility of evaluating bio–nano interaction descriptors such as protein binding energies from first principles; that is, based only on ENM structure and composition[51] using a multiscale approach: a combination of coupled quantum chemistry, molecular dynamics, and mesoscale simulations. With this technique, the bio–nano interaction descriptors can be evaluated even for artificial materials and shapes, which makes it promising for screening ENMs at the stage of the materials and product design. Furthermore, the first-principles multiscale simulation methodology allows one to construct advanced integral descriptors such as the abundance of proteins of certain types or specific amino acids in the corona of an ENM of given structure and composition, and thus to build a model of ENM-specific protein corona[50] and through this predict the biological activity of the ENM. Moreover, detailed characterization of the ENMs after initial contact with organisms at different stages of

systemic transport can provide molecular-level nanodescriptors for 'mechanism-aware' toxicity prediction schemes[67].

Materials modelling and experimental ENM characterization after exposure to an organism could be used to develop the relevant ENM descriptors. At the first level, such descriptors would include characterization of the interfacial ENM contact with biomolecules in terms of the binding energies of biomolecule elements (amino acids, lipid headgroups and so on). Such descriptors should be organized into a bio–nano interactions database linked with initiatives such as The Human Protein Atlas[68] and equivalents for ecotoxicity species and plants[69,70]. The Human Protein Atlas contains detailed quantitative information on the relative abundances of proteins in different cellular and tissue environments that can be integrated and used to predict ENM corona formation, including characterization of the outer corona surface and prediction of the likelihood of a particular hazardous effect. To develop mechanism-aware nano-QSARs, one should systematically analyse the ENM-induced toxicity pathways and map the nanodescriptors to the molecular initiating events and the specific adverse outcome for any ENMs. As the majority of molecular initiating events responsible for the induction of adverse outcome by ENMs are non-specific, key events should be considered at this stage. However, this is only possible if the sequence of key events and the dose–response relationships that lead to eventual adverse outcome are well described and understood. A nano-QSAR model linking the properties of ENMs with the upstream key events that are essential for the initiation and manifestation of an adverse outcome could then be developed.

The only approach[71] reported so far in this area employs a nano-QSAR model to predict the transcriptomic-based pathway-level response associated with the lung tissue inflammation induced by multiwalled carbon nanotubes. This AOP-anchored model allows the prediction of the inflammatory response induced by nanotubes on the basis of their aspect ratio values. In this study, the pathway-relevant genes were also analysed, and the transcriptomic biomarkers that can be applied to assess the inflammogenic properties of nanotubes were identified. The overall assessment scheme would thus combine materials modelling, systems biology, in vivo and in vitro studies. A systematic mapping of ENMs with their induced molecular alterations and interactions in the AOP context can provide mechanistic interpretability of ENM molecular dynamics[1,24,72–74]. Furthermore, such models can be exploited when designing ENMs that exert specific biological effects[75].

Outlook. Nanoinformatics is a vibrant and apidly developing area, with enormous potential to provide new mechanistic insights into ENM interactions with living systems and to facilitate a transition to in silico risk assessment.

This Review provides a snapshot of conceptual directions in progress within the field of nanoinformatics that are necessary to achieve a complete representation of ENMs and description in terms of nanodescriptors. Under the newly proposed scheme of ENM representation, nanodescriptors should reflect not only pristine characteristics of the main constituents, but also the information on the possible relationships between the constituents. Information on the attached biomolecules in response to the ENMs surroundings and, consequently, the initiation of key events leading to adverse outcomes should also be incorporated. In addition, representing the distributions of ENMs in terms of their size, structural (for example, shape irregularities, defects in the crystal structure) and surface (charges, coating thickness and coverage of the surface and so on) features to enable particle-by-particle predictive analysis may prove to be the key to understanding and minimizing potential risks in the context of SSbD. Special attention should be paid to developing calculated coding nanodescriptors that may be utilized for virtual screening at the earliest possible stage of nanoparticle design. Finally, comprehensive documentation of the metadata for experimental nanodescriptors is essential, including the data processing and clean-up steps, for example, to ensure data provenance.

The five challenges defined above should be tackled with the highest priority to develop comprehensive models for predictive nanoinformatics.

Received: 19 October 2021; Accepted: 8 June 2022;
Published online: 18 August 2022

References

1. Fadeel, B. et al. Advanced tools for the safety assessment of nanomaterials. *Nat. Nanotechnol.* **13**, 537–543 (2018).
2. Winkler, D. A. Role of artificial intelligence and machine learning in nanosafety. *Small* **16**, 2001883 (2020).
3. Cherkasov, A. et al. QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.* **57**, 4977–5010 (2014).
4. Fourches, D. et al. Quantitative nanostructure-activity relationship modeling. *ACS Nano* **4**, 5703–5712 (2010).
5. Puzyn, T. et al. Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. *Nat. Nanotechnol.* **6**, 175–178 (2011).
6. Jeliazkova, N. et al. Towards FAIR nanosafety data. *Nat. Nanotechnol.* **16**, 644–654 (2021).
7. Rybińska-Fryca, A., Mikolajczyk, A. & Puzyn, T. Structure–activity prediction networks (SAPNets): a step beyond Nano-QSAR for effective implementation of the safe-by-design concept. *Nanoscale* **12**, 20669–20676 (2020).
8. Marchese Robinson, R. L. et al. How should the completeness and quality of curated nanomaterial data be evaluated? *Nanoscale* **8**, 9919–9943 (2016).
9. Muratov, E. N. et al. QSAR without borders. *Chem. Soc. Rev.* https://doi.org/10.1039/d0cs00098a (2020).
10. Stone, V. et al. A framework for grouping and read-across of nanomaterials-supporting innovation and risk assessment. *Nano Today* https://doi.org/10.3390/nano10102017 (2020).
11. Papadiamantis, A. G. et al. Predicting cytotoxicity of metal oxide nanoparticles using Isalos Analytics Platform. *Nanomaterials* **10**, 2493 (2020).
12. Puzyn, T. et al. in *Recent Advances in Qsar Studies: Methods and Applications* Vol. 8 (eds. Puzyn, T. et al.) 127–176 (Springer, 2010).
13. Shoombuatong, W. et al. in *Advances in QSAR Modeling* (Ed. Roy, K.) 3–55 (Springer, 2017).
14. Karakus, C. O. & Winkler, D. A. Overcoming roadblocks in computational roadmaps to the future for safe nanotechnology. *Nano Futures* **5**, 22002 (2021).
15. Haase, A. & Klaessig, F. (eds) EU US roadmap nanoinformatics 2030. *Zenodo* https://doi.org/10.5281/zenodo.1486012 (2018).
16. Mech, A. et al. Insights into possibilities for grouping and read-across for nanomaterials in EU chemicals legislation. *Nanotoxicology* **13**, 119–141 (2019).
17. Miernicki, M., Hofmann, T., Eisenberger, I., von der Kammer, F. & Praetorius, A. Legal and practical challenges in classifying nanomaterials according to regulatory definitions. *Nat. Nanotechnol.* **14**, 208–216 (2019).
18. Regulation (EC) No 1907/2006 of the European Parliament and of the Council (*EUR-Lex*, 18 December 2006); https://eur-lex.europa.eu/eli/reg/2006/1907/2014-04-10
19. Commission Regulation (EU) 2018/1881 (*EUR-Lex*, 3 December 2018); https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018R1881
20. Subbotina, J. & Lobaskin, V. Multiscale modeling of bio-nano interactions of zero-valent silver nanoparticles. *J. Phys. Chem. B* **126**, 1301–1314 (2022).
21. Kochev, N., Jeliazkova, N. & Tsakovska, I. in *Issues in Toxicology* (eds. Neagu, D., Richarz, A.-N.) 69–107 (The Royal Society of Chemistry, 2020).
22. *Commission Regulation (EU) 2018/1881 of 3 December 2018 amending Regulation (EC) No 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) as regards Annexes I, III,VI, V* (European Commission, 2018).
23. *Guidance on Information Requirements and Chemical Safety Assessment: Appendix R.6-1 for Nanomaterials Applicable to the Guidance on QSARs and Grouping of Chemicals* Version 2.0, 3 (ECHA, 2019); https://doi.org/10.2823/273911
24. Burello, E. Review of (Q)SAR models for regulatory assessment of nanomaterials risks. *NanoImpact* **8**, 48–58 (2017).
25. Lynch, I., Weiss, C. & Valsami-Jones, E. A strategy for grouping of nanomaterials based on key physico-chemical descriptors as a basis for safer-by-design NMs. *Nano Today* **9**, 266–270 (2014).
26. Lynch, I., Afantitis, A., Leonis, G., Melagraki, G. & Valsami-Jones, E. in *Advances in QSAR modeling. Challenges and Advances in Computational Chemistry and Physics* Vol. 24 (Ed. Roy, K.) 385–424 (Springer, 2017).
27. Lynch, I. & Lee, R. G. in *Innovation, Technology, and Knowledge Management* (eds. Murphy, F. et al.) 145–169 (Springer, 2016).
28. Mülhopt, S. et al. Characterization of nanoparticle batch-to-batch variability. *Nanomaterials* **8**, 311 (2018).

29. Yao, Y. et al. High-throughput, combinatorial synthesis of multimetallic nanoclusters. *Proc. Natl Acad. Sci. USA* **117**, 6316–6322 (2020).
30. Kluender, E. J. et al. Catalyst discovery through megalibraries of nanomaterials. *Proc. Natl Acad. Sci. USA* **116**, 40–45 (2019).
31. *Toxic Substances Control Act* (US EPA,1979): https://www.epa.gov/laws-regulations/summary-toxic-substances-control-act
32. *TSCA Inventory Status of Nanoscale Substances – General Approach* (US EPA, 2008); https://www.epa.gov/reviewing-new-chemicals-under-toxic-substances-control-act-tsca/control-nanoscale-materials-under
33. Nano-InChI working group; https://www.inchi-trust.org/nanomaterials/
34. Lynch, I. et al. Can an InChI for nano address the need for a simplified representation of complex nanomaterials across experimental and nanoinformatics studies? *Nanomaterials* **10**, (2020).
35. Toropova, A. P. & Toropov, A. A. Nanomaterials: Quasi-SMILES as a flexible basis for regulation and environmental risk assessment. *Sci. Total Environ.* **823**, 153747 (2022).
36. Toropov, A. A., Sizochenko, N., Toropova, A. P. & Leszczynski, J. Towards the development of global nano-quantitative structure-property relationship models: zeta potentials of metal oxide nanoparticles. *Nanomaterials* **8**, 243 (2018).
37. Mikolajczyk, A. et al. Nano-QSAR modeling for ecosafe design of heterogeneous TiO_2-based nano-photocatalysts. *Environ. Sci. Nano* **5**, 1150–1160 (2018).
38. Mikolajczyk, A. et al. A chemoinformatics approach for the characterization of hybrid nanomaterials: safer and efficient design perspective. *Nanoscale* **11**, 11808–11818 (2019).
39. Roy, J., Ojha, P. K. & Roy, K. Risk assessment of heterogeneous TiO_2-based engineered nanoparticles (NPs): a QSTR approach using simple periodic table based descriptors. *Nanotoxicology* **13**, 701–716 (2019).
40. Svendsen, C. et al. Key principles and operational practices for improved nanotechnology environmental exposure assessment. *Nat. Nanotechnol.* **15**, 731–742 (2020).
41. Amos, J. D. et al. The NanoInformatics Knowledge Commons: capturing spatial and temporal nanomaterial transformations in diverse systems. *NanoImpact* **23**, 100331 (2021).
42. Di Cristo, L. et al. Grouping hypotheses and an integrated approach to testing and assessment of nanomaterials following oral ingestion. *Nanomaterials* **11**, 2623 (2021).
43. Afantitis, A., Melagraki, G., Tsoumanis, A., Valsami-Jones, E. & Lynch, I. A nanoinformatics decision support tool for the virtual screening of gold nanoparticle cellular association using protein corona fingerprints. *Nanotoxicology* **12**, 1148–1165 (2018).
44. Wyrzykowska, E., Mikolajczyk, A., Sikorska, C. & Puzyn, T. Development of a novel in silico model of zeta potential for metal oxide nanoparticles: a nano-QSPR approach. *Nanotechnology* **27**, 1–8 (2016).
45. Mikolajczyk, A. et al. Zeta potential for metal oxide nanoparticles: a predictive model developed by a nano-quantitative structure-property relationship approach. *Chem. Mater.* **27**, 2400–2407 (2015).
46. Grzelczak, M., Liz-Marzan, L. M. & Klajn, R. Stimuli-responsive self-assembly of nanoparticles. *Chem. Soc. Rev.* **48**, 1342–1361 (2019).
47. Liu, Y., Zhu, S., Gu, Z., Chen, C. & Zhao, Y. Toxicity of manufactured nanomaterials. *Particuology* **69**, 31–48 (2022).
48. Baer, D. R., Munusamy, P. & Thrall, B. D. Provenance information as a tool for addressing engineered nanoparticle reproducibility challenges. *Biointerphases* **11**, 04B401 (2016).
49. Mancardi, G. et al. Multi-scale modelling of aggregation of TiO_2 nanoparticle suspensions in water. *Nanomaterials* **12**, 217 (2022).
50. Alsharif, S. A., Power, D., Rouse, I. & Lobaskin, V. In silico prediction of protein adsorption energy on titanium dioxide and gold nanoparticles. *Nanomaterials* **10**, 1967 (2020).
51. Rouse, I. et al. First principles characterisation of bio–nano interface. *Phys. Chem. Chem. Phys.* **23**, 13473–13482 (2021).
52. Rouse, I. & Lobaskin, V. A hard-sphere model of protein corona formation on spherical and cylindrical nanoparticles. *Biophys. J.* **120**, 4457–4471 (2021).
53. Buzea, C., Pacheco, I. I. & Robbie, K. Nanomaterials and nanoparticles: sources and toxicity. *Biointerphases* **2**, MR17–MR71 (2007).
54. Rabanel, J.-M. et al. Nanoparticle heterogeneity: an emerging structural parameter influencing particle fate in biological media? *Nanoscale* **11**, 383–406 (2019).
55. Adjei, I. M., Peetla, C. & Labhasetwar, V. Heterogeneity in nanoparticles influences biodistribution and targeting. *Nanomedicine* **9**, 267–278 (2014).
56. *Appendix for Nanoforms Applicable to the Guidance on Registration and Substance Identification* (ECHA, 2019); https://doi.org/10.2823/832485
57. Caputo, F., Clogston, J., Calzolai, L., Rösslein, M. & Prina-Mello, A. Measuring particle size distribution of nanoparticle enabled medicinal products, the joint view of EUNCL and NCI-NCL. A step by step approach combining orthogonal measurements with increasing complexity. *J. Control. Release* **299**, 31–43 (2019).
58. Lundqvist, M. et al. The evolution of the protein corona around nanoparticles: A test study. *ACS Nano* **5**, 7503–7509 (2011).
59. Chetwynd, A. J., Zhang, W., Thorn, J. A., Lynch, I. & Ramautar, R. The nanomaterial metabolite corona determined using a quantitative metabolomics approach: a pilot study. *Small* **16**, 2000295 (2020).
60. Yan, X. et al. In silico profiling nanoparticles: predictive nanomodeling using universal nanodescriptors and various machine learning approaches. *Nanoscale* **11**, 8352–8362 (2019).
61. Yan, X., Sedykh, A., Wang, W., Yan, B. & Zhu, H. Construction of a web-based nanomaterial database by big data curation and modeling friendly nanostructure annotations. *Nat. Commun.* **11**, 1–10 (2020).
62. Sizochenko, N. et al. 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 (2014).
63. Sizochenko, N., Jagiello, K., Leszczynski, J. & Puzyn, T. How the 'liquid drop' approach could be efficiently applied for quantitative structure-property relationship modeling of nanofluids. *J. Phys. Chem. C* **119**, 25542–25547 (2015).
64. Utembe, W., Potgieter, K., Stefaniak, A. B. & Gulumian, M. Dissolution and biodurability: Important parameters needed for risk assessment of nanomaterials. *Part. Fibre Toxicol.* **12**, 11 (2015).
65. Lin, S. et al. Zebrafish high-throughput screening to study the impact of dissolvable metal oxide nanoparticles on the hatching enzyme, ZHE1. *Small* **9**, 1776–1785 (2013).
66. Kokot, H. et al. Prediction of chronic inflammation for inhaled particles: the impact of material cycling and quarantining in the lung epithelium. *Adv. Mater.* **32**, 2003913 (2020).
67. Ellis, L.-J. A. & Lynch, I. Mechanistic insights into toxicity pathways induced by nanomaterials in *Daphnia magna* from analysis of the composition of the acquired protein corona. *Environ. Sci. Nano* **7**, 3343–3359 (2020).
68. Uhlen, M. et al. Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.* **28**, 1248–1250 (2010).
69. Wheeler, K. E. et al. Environmental dimensions of the protein corona. *Nat. Nanotechnol.* **16**, 617–629 (2021).
70. Smythers, A. L. & Hicks, L. M. Mapping the plant proteome: tools for surveying coordinating pathways. *Emerg. Top. Life Sci.* **5**, 203–220 (2021).
71. Jagiello, K. et al. Transcriptomics-based and AOP-informed structure-activity relationships to predict pulmonary pathology induced by multiwalled carbon nanotubes. *Small* **17**, 2003465 (2020).
72. Myden, A., Hill, E. & Fowkes, A. Using adverse outcome pathways to contextualise (Q)SAR predictions for reproductive toxicity – a case study with aromatase inhibition. *Reprod. Toxicol.* **108**, 43–55 (2022).
73. Ellison, C. M., Piechota, P., Madden, J. C., Enoch, S. J. & Cronin, M. T. D. Adverse outcome pathway (AOP) informed modeling of aquatic toxicology: QSARs, read-across, and interspecies verification of modes of action. *Environ. Sci. Technol.* **50**, 3995–4007 (2016).
74. Seo, M., Chae, C. H., Lee, Y., Kim, H. R. & Kim, J. Novel QSAR models for molecular initiating event modeling in two intersecting adverse outcome pathways based pulmonary fibrosis prediction for biocidal mixtures. *Toxics* **9**, 59 (2021).
75. Halappanavar, S. et al. Adverse outcome pathways as a tool for the design of testing strategies to support the safety assessment of emerging advanced materials at the nanoscale. *Part. Fibre Toxicol.* **17**, 16 (2020).
76. Toropova, A. P., Toropov, A. A. & Benfenati, E. QSPR as a random event: solubility of fullerenes C[60] and C[70]. *Fuller. Nanotub. Carbon Nanostruct.* **27**, 816–821 (2019).
77. Toropov, A. A. & Toropova, A. P. Quasi-SMILES and nano-QFAR: united model for mutagenicity of fullerene and MWCNT under different conditions. *Chemosphere* **139**, 18–22 (2015).

Acknowledgements

This work was funded via the European Union's H2020 project NanoSolveIT (grant agreement number 814572), with additional support from the EU H2020 project NanoInformaTIX (grant agreement number 814426) and EU H2020 project PATROLS (grant agreement number 760813). V.L. and J.S. were supported by Science Foundation Ireland grant number 16/IA/4506.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Tomasz Puzyn.

Peer review information *Nature Nanotechnology* thanks Robert Rallo and David Winkler for their contribution to the peer review of this work.