ELSEVIER

A computational view on nanomaterial intrinsic and extrinsic features for nanosafety and sustainability

Giulia Mancardi [a,*], Alicja Mikolajczyk [b,m,*], Vigneshwari K. Annapoorani [c], Aileen Bahl [d], Kostas Blekos [e], Jaanus Burk [f], Yarkın A. Çetin [g], Konstantinos Chairetakis [e], Sutapa Dutta [c], Laura Escorihuela [g], Karolina Jagiello [b,m], Ankush Singhal [h], Rianne van der Pol [h], Miguel A. Bañares [i], Nicolae-Viorel Buchete [c], Monica Calatayud [j], Verónica I. Dumit [d], Davide Gardini [k], Nina Jeliazkova [l], Andrea Haase [d], Effie Marcoulaki [e], Benjamí Martorell [g], Tomasz Puzyn [b,m], G.J. Agur Sevink [h], Felice C. Simeone [k], Kaido Tämm [f], Eliodoro Chiavazzo [a,*]

[a] Politecnico di Torino, Corso Duca degli Abruzzi, 24, Torino 10129, Italy
[b] Laboratory of Environmental Chemometrics, Institute for Environmental and Human Health Protection, Faculty of Chemistry, University of Gdansk, 80-308 Gdansk, Poland
[c] School of Physics & Institute for Discovery, University College Dublin, Belfield, Dublin 4, Ireland
[d] German Federal Institute for Risk Assessment (BfR), Department of Chemical and Product Safety, Berlin, Germany
[e] Institute of Nuclear & Radiological Sciences & Technology, Energy & Safety, National Centre for Scientific Research ńDemokritosż, Athens, Greece
[f] Institute of Chemistry, University of Tartu, Ravila 14a, 50411 Tartu, Estonia
[g] Departament d'Enginyeria Química, Universitat Rovira i Virgili, Avinguda dels Països Catalans, 26, 43007 Tarragona, Spain
[h] Leiden Institute of Chemistry, Leiden University, P.O. Box 9502, 2300 RA Leiden, The Netherlands
[i] Institute for Catalysis, ICP-CSIC, Marie Curie 2, E-28049 Madrid, Spain
[j] Laboratoire de Chimie Theorique, CNRS, Sorbonne Université, 4 Place Jussieu, 75005 Paris, France
[k] ISTEC-CNR, Institute of Science and Technology for Ceramics - National Research Council of Italy, Faenza, Italy
[l] Ideaconsult Ltd, Sofia, Bulgaria
[m] QSAR Lab Ltd., Trzy Lipy 3, 80-172 Gdansk, Poland

In recent years, an increasing number of diverse Engineered Nano-Materials (ENMs), such as nanoparticles and nanotubes, have been included in many technological applications and consumer products. The desirable and unique properties of ENMs are accompanied by potential hazards whose impacts are difficult to predict either qualitatively or in a quantitative and predictive manner. Alongside established methods for experimental and computational characterisation, physics-based modelling tools like molecular dynamics are increasingly considered in Safe and Sustainability-by-design (SSbD) strategies that put user health and environmental impact at the centre of the design and development of new products. Hence, the further development of such tools can support safe and sustainable innovation and its regulation.

This paper stems from a community effort and presents the outcome of a four-year-long discussion on the benefits, capabilities and limitations of adopting physics-based modelling for computing suitable features of nanomaterials that can be used for toxicity assessment of nanomaterials in combination with data-based models and experimental assessment of toxicity endpoints. We review

* Corresponding authors.
E-mail addresses: Mancardi, G. (mancardigiulia@gmail.com), Mikolajczyk, A. (alicja.mikolajczyk@ug.edu.pl), Chiavazzo, E. (eliodoro.chiavazzo@polito.it).

Nomenclature

AIMD	ab initioMolecular Dynamics
AO(P)	Adverse Outcome (Pathway)
API	Application Programming Interface
BD	Brownian Dynamics
CG(MD)	Coarse-Grained (Molecular Dynamics)
CSS	Chemicals Strategy for Sustainability
DFT	Density Functional Theory
DFTB	Density Functional Tight-Binding
DLVO	Derjaguin-Landau-Verwey-Overbeek
ENM	Engineered NanoMaterials
FAIR	Findable, Accessible, Interoperable and Reusable
GA	Genetic Algorithm
GAN	Generative Adversarial Network
HOMO	Highest Occupied Molecular Orbital
hPF	hybrid Particle-Field
KE	Key Events
LCA	Life Cycle Assessment
LDM	Liquid Drop Model
LUMO	Lowest Unoccupied Molecular Orbital
MD	Molecular Dynamics
MDReaxFF	Reactive Force Field Molecular Dynamics
MIE	Molecular Initiating Events
ML	Machine Learning
MoA	Mode of Action
MODA	MOdelling DAta fiche
NAM	New Approach Methodologies
PMF	Potential of Mean Force
QM	Quantum Mechanics
QNAR	Quantitative Nanostructure–Activity Relationship
QSAR	Quantitative Structure–Activity Relationship
QSPR	Quantitative Structure–Property Relationship
RNN	Recurrent Neural Network
S(S) bD	Safe (and Sustainable) by Design
SASA	Solvent Accessible Surface Area
SCFT	Self-Consistent Field Theory
VAE	Variational Autoencoder

modern multiscale physics-based models that generate advanced system-dependent (intrinsic) or time- and environment-dependent (extrinsic) descriptors/features of ENMs (primarily, but not limited to nanoparticles, NPs), with the former being related to the bare NPs and the latter to their dynamic fingerprinting upon entering biological media. The focus is on (i) effectively representing all nanoparticle attributes for multicomponent nanomaterials, (ii) generation and inclusion of intrinsic nanoform properties, (iii) inclusion of selected extrinsic properties, (iv) the necessity of considering distributions of structural advanced features rather than only averages. This review enables us to identify and highlight a number of key challenges associated with ENMs' data generation, curation, representation and use within machine learning or other advanced data-driven models to ultimately enhance toxicity assessment. Finally, the set up of dedicated databases as well as the development of grouping and read-across strategies based on the mode of action of ENMs using omics methods are identified as emerging methodologies for safety assessment and reduction of animal testing.

Keywords: Nanoinformatics; Nanosafety; Engineered nanomaterials; Physicochemical descriptors; Materials modeling; Machine learning; Grouping approaches; Multiscale modeling; Safe and Sustainability-by-design(SSbD)

Introduction

ENMs perform key dedicated tasks in catalysis [1], medicine [2,3], agriculture [4], food [5] and energy [6,7] among many others, because of their exceptional and often unique characteristics. ENMs underpin new products and devices requiring control of matter at the nanometer scale, and the possibility to fine-tune their synthesis protocols results in countless variants with different physicochemical properties. The exceptional properties of ENMs - stemming from the manipulation of matter at the atomic scale - can also come with yet little-known risks to human health and the environment. In fact, the biological activity of ENMs is thought to be closely related to their physicochemical characteristics, which may be altered by the biological medium itself during the lifetime of ENMs (e.g., protein corona formation and structural modifications). Both the intrinsic and extrinsic physicochemical features of ENMs are key for recognizing and predicting hazards as well as assessing safety characteristics. The former are related to chemical composition, crystal structure, size, shape and surface structure, whereas the latter pertain to the non-trivial interaction with the environment. As such, a complete set of those features (here referred to as *advanced descriptors*) are critical in nanoinformatics for data-based model development, such as the popular quantitative nanostructure–activity/property models (QSAR/QSPR). In this respect, predictive models offer unprecedented opportunities for knowledge-based optimization and development of new ENMs improving their functionality and - at the same time - minimizing unexpected health and/or environmental risks. Preliminary *in silico* screening of possible versions of new ENMs can thus lead to optimal nanostructures with reduced hazardous characteristics even before the production stage. However, the pace of further progress in nanotechnology will critically depend on a synergistic knowledge integration of experimental evidence with data from reliable theoretical and computational models.

The combined study of materials modeling (nanostructure characterization) and predictive models for the design of safe

nanostructures with desired properties (Safe and Sustainability-by-design perspective, SSbD) brings along new opportunities both in the academic and industrial context. Nonetheless, significant challenges arising from both the extremely demanding computational models (physics-based and data-based) and the practically unlimited number of possible combinations of different substances and nanoforms that lie ahead. Even for a single material, there is a plethora of possible bulk and surface structures, defects and terminations, each delivering unique properties. This general challenge can only be tackled by an integrated approach. Here, we critically analyze the process associated with the *in silico* evaluation of the ENMs descriptors, with the aim of highlighting challenges and opportunities when using physics-based modelling for generating them.

Our discussion takes place under the broad perspective of the European and US legislation development and their emerging strategies related to risk assessment of nanomaterials (e.g., the EU Materials Modelling Council, the EU REACH regulation, US-EU nanoEHS initiative, the EU-US Nanoinformatics Roadmap 2030[8]). Despite the well-accepted relationship between physicochemical properties (i.e. descriptors) and the (eco)-toxicological endpoints, a comprehensive computational description of ENMs and an understanding of the basic mechanisms behind their interaction with biological media are still very challenging.

First, we focus on some of the (physics-based) computational approaches for estimating advanced descriptors at time and space scales relevant to nanosafety. In this respect, we provide an overview of the approaches for investigating ENMs electronic and atomistic structure (thus focusing on intrinsic features) up to the mesoscopic description of interactions with biological matter, such as proteins or cell membranes, hence moving towards more extrinsic features.

Second, more recent data-based models and approaches for nanosafety assessment are analyzed. As illustrated schematically in the top part of Fig. 1, computations can be performed at different space/time scales by solving appropriate physical model equations with advanced descriptors collected from each simulator. Advanced descriptors can be passed across different time and space scales, realizing a chain of multiscale simulations [9]. This vision is certainly fascinating, however, as discussed below, it comes with formidable challenges: as opposed to the high-accuracy data extracted by the physics-based models, the inherent computational cost is often prohibitively high. This leads to a low data variance that renders the subsequent translation

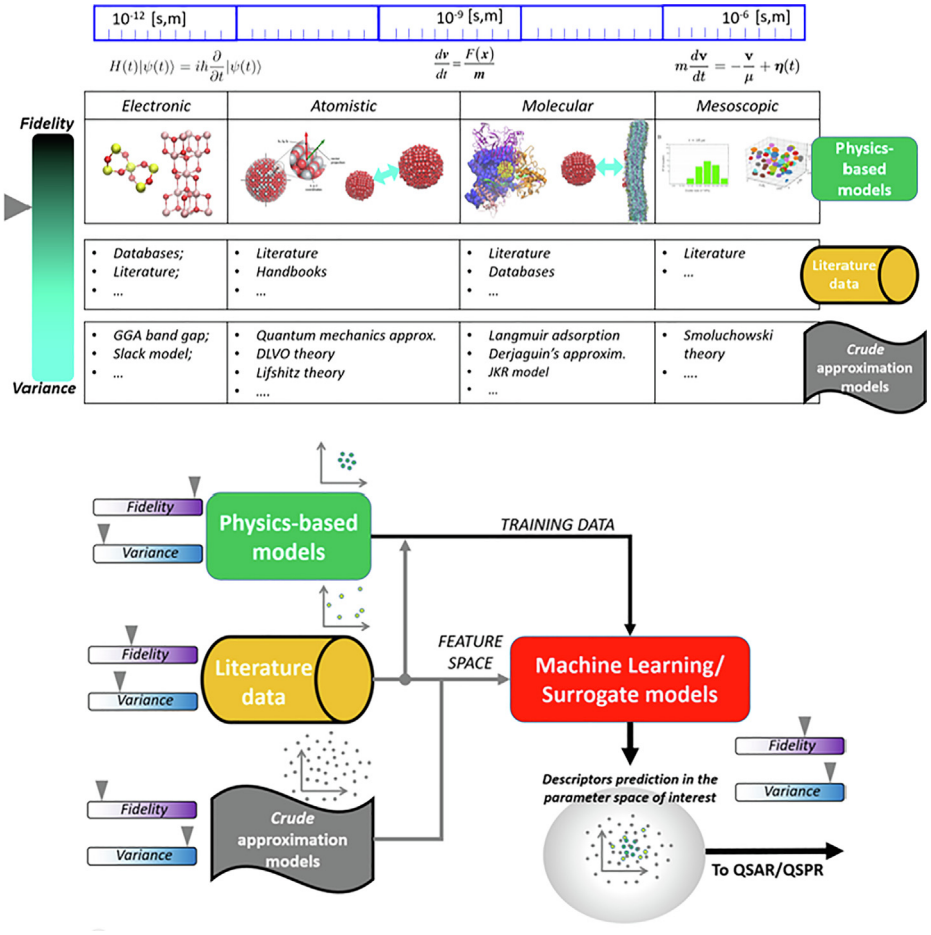

FIG. 1

TOP: Physics-based models ensure high fidelity at a high computational cost, making it possible to investigate only a restricted region of the ENM design space. Hence, the necessary variance is needed to describe the large variety of parameters of interest (e.g., particle size, coating, etc.). Thus, predicting advanced descriptors for nanosafety has to stem from different sources, such as literature data and crude approximation models characterized by a lower fidelity level. BOTTOM: Discrepancy in fidelity and variance is to be properly orchestrated, possibly using ML models [11].

of such data into input for QSAR/QSPR (or other data-based) models extremely difficult if not impossible. Details on obstacles at every relevant scale are discussed, and the current status and challenges with the European regulatory context on materials modeling and *in silico* estimate of descriptors for nanosafety purposes are reported below [10].

Safe and Sustainability-by-design (SSbD) strategy: From nanoform description to safety modeling

Within the European Green Deal, the Chemicals Strategy for Sustainability (CSS)[12] identified several actions to reduce negative impacts on human health and the environment associated with chemicals, materials, products, and services commercialized or introduced onto the EU market. In particular, the ambition of the CSS is to phase out the most harmful substances and substitute, as far as possible, all other substances of concern and otherwise minimize their use and track them. This objective requires novel approaches to analyze and compare all life cycle stages, effects, releases, and emissions for specific chemicals, materials, products, and services, and moving towards zero pollution for air, water, soil, and biota. The SSbD framework aims to support the design and development of safe and sustainable chemicals and materials with research and innovation (R&I) activities.

Although the safety of nanomaterials has been of concern to the scientific community for more than two decades, there is still a limited number of validated and regulatory-accepted alternative nano-specific approaches for assessing their human safety. In fact, the current knowledge about various adverse effects induced by nanomaterials after exposure does not yet enable a broad development of the SSbD strategy.

The safety and sustainability assessment

Reliable and efficient methods allowing, in a timely manner, to assess exposure risks should be first developed. In this context, the development of new alternative methods combining in vitro, chemical analysis as well as *in silico* models to predict the potential adverse impact of chemicals, including nanomaterials, on human health is highly needed [13]. These tools are expected to be useful for regulators and policymakers; therefore, they need to consider biologically plausible and regulatory-relevant events essential for the occurrence of possible adverse outcomes [14]. In line with this point of view is the strategy that integrates nanoinformatics models with the Adverse Outcome Pathways (AOPs) [15,16]. The AOP is a framework that describes a sequence of biological events following stressor exposure and leading to various Adverse Outcomes (AO). This concept links Molecular Initiating Events (MIE) to the series of key cellular- or tissue-level changes (so-called Key Events, KE) that culminate in the manifestation of AO. By supporting the identification of AOP-relevant events and information that might be applied to the weight of evidence-based safety decisions, AOP can serve as a framework for developing nanoinformatics models useful for regulatory actions. This would mean that events recognized as crucial to manifest nano-specific adverse effects should be considered for modeling. As an example, recently, Jagiello et al. [15] proposed a model that linked the structural properties of multiwalled carbon nanotubes with the recruitment of pro-inflammatory mediators into the lungs, that finally leads to lung fibrosis according to AOP173. Applying the AOP framework in this model enables to understand events triggers following nanotube exposure and occurring in different biological organizations (molecular, cellular, tissue, organ, individual). To sum up, with the development of the AOP-anchored nanoinoformatics models (models for key biologically plausible and regulatory-relevant events), the usage and acceptance of those approaches in making regulatory decisions about the safety of nanomaterials exposure can increase.

Additionally, the JRC (2022)[17] reviewed previous decision frameworks for materials safety to investigate how sustainability (broadly including social and economic aspects) was considered and to define a set of criteria to integrate safety and sustainability. Sustainability assessment can use conventional sustainability metrics adapted for nanotechnology products and processes. In this context, Stieberova et al. [18] used techno-economic and life cycle environmental criteria for sustainability assessment, to compare alternative nanoparticle production technologies; García-Quintero and Palencia [19] analyzed conventional quantitative sustainability metrics and proposed Life Cycle Assessment (LCA) as a suitable metric to compare, optimize and quantify bio-based nanobiotechnology and nanosynthesis protocols.

Regulatory actions to drive SSbD chemicals and materials

Predictive *in silico* modeling, accessible and searchable databases and quantitative tools for risk assessment and prevention are critical for the successful implementation of SbD/SSbD strategies in the nanosafety context and to develop relevant protocols, reference materials, realistic *in vitro* and computational models, as well as grouping and read-across methods. [20] Additional challenges stem from the need to assess the fate and reactivity of next generations of nanomaterials [21] including smart nanomaterials and nano-enabled products [22]. New Approach Methodologies (NAMs), comprise a wide range of such models and tools [23] to conduct robust and reliable chemical safety assessments and reduce animal testing.

With substantial funding, the European Commission has supported NAMs for nanomaterials, Refs. [24,25], acknowledging their potential in regulatory decision-making and innovation. NAMs could increase regulatory preparedness through fit-for-purpose data and standardized tests specific to nanomaterials [26], and their potential would be better exploited within tiered regulatory schemes. [27] Consideration of available standards (as the OECD standard for QSARs [28] and recommendations of the European Materials Modelling Council for validating *in silico* tools is essential to promote their adoption for regulatory use. On the technical side, this also enables model integration into complete computational IATA workflows [29].

As presented in Section 'Data-based models for linking ENM features to safety', there is rich literature on a variety of models for the prediction of ENMs properties and toxicity. The reverse problem, i.e. the development of NM structures featuring desired or optimal properties has also been considered in few recent works. As a representative example, preliminary works considered a conditional deep convolutional Generative Adversarial Network (GAN) using competitive learning to suggest nanopho-

tonic structures with desired optical properties [30], and a GAN to generate crystalline porous materials, in particular pure silica zeolite structures [31]. However, since there is still much to do for understanding the behavior of ENMs, several projects have been funded by the EU in the last years [32–35] to address safety and risk governance issues.

With this in mind, researchers in NanoInformaTIX [36] project have developed an optimization methodology to guide the search for safer ENMs and to support the application of safe(r)-by-design (SbD) approaches [37]. The assessment and screening processes use efficient representations coupled with quantitative tools for similarity assessment to investigate morphology-based behavior [38]. The now implemented tool can be trained using available datasets and can enable the assessment of various design options generated during the optimal search.

Hazards of ENMs

As mentioned above, the physicochemical properties of nanomaterials may significantly differ from their bulk counterparts. Despite the beneficial technological consequences associated with this, ENMs might exhibit hazardous effects on human health or the environment. The experience gathered from the employment of asbestos fibers decades ago serves as a cautionary tale for the potential hazard of nanomaterials. Asbestos were extensively used in various products with a wide range of applications because of the advantageous properties, like heat resistance/isolation and durability. However, after many years of ubiquitous quotidian presence, the insidious harmful effects on health became evident. In this regard, it was discovered that after a long latency period asbestos were able to cause lung cancer and mesothelioma [39]. This effect was linked with its high-aspect ratio fiber structure and frustrated phagocytosis. Gathering materials by common features leading to similar adverse effects is a first step to group and read-across. Therefore, investigating the potential hazards of nanomaterials is crucial to guarantee their safety enabling their responsible usage. Understanding their potential hazard is important to engineer relevant modifications and undertake the necessary measurements to minimize exposure, while still profiting from their various beneficial properties. Understanding their potential hazard requires describing surface structure, properties and reactivity, which define how these engineered nanomaterials interact with each other and with the environment. The latter determines the trend to aggregate, which affects fate and exposure; moreover it shapes their surface properties, thus altering interaction among particles (fate, exposure) and modifying reactivity. Although there are several exposure scenario for nanomaterial toxicity, inhalation is considered the most critical uptake route since it allows the particles to reach the sensitive tissues deep within the lungs [40]. There, the particles could be taken up by lung cells or interact with the immune system, leading to harmful effects. Asbestos is not the only 1-dimensional material potentially harmful. Other materials, with very different chemistry and composition, share common features. Studies on specific types of carbon nanotubes suggested that they may show similar toxicity as asbestos fibers at lung level [41,42]. Similarly, research on silver nanoparticles showed genotoxicity and pointed out that smaller-sized particles exhibited higher toxicity in in vitro settings [43]. It is important to emphasize that not all nanomaterials are associated with risks for human health. Their toxicological potential depends on different factors such as size, shape, functionalization, besides chemical composition. Ultimately, toxic effect is a phenomenon triggered at the surface of nanomaterials, and how it interacts with its environment. Bulk properties do not typically correlate with surface properties or structure; therefore, no direct translation of bulk characteristics can be made to surface reactivity. Experimental characterization may be hampered by this, which must be taken in consideration when defining characterization strategies. From a more methodological perspective, material characterization should include detailed information on the surface and its defects, how the latter and the former depend on the underlying bulk structure and defects and how those are affected and, ultimately, shaped by the interaction with the environment or biological media. The combination of these properties results in countless possible nanomaterials that can enter the market, which can rapidly overwhelm the current risk assessment procedure.

For this reason, SSbD strategies are particularly important, as they have the potential to facilitate and expedite risk assessment, necessary to minimize the potential risks associated with the use of nanomaterials throughout their lifecycle. Moreover, SSbD strategies are crucial to ensure the responsible development of nanomaterials and their applications.

The role of materials modeling in ENMs' hazard assessment

It is worth stressing that accurate physics-based modelling of materials is typically not meant to directly simulate the basic mechanisms underpinning ENMs toxicity. On the contrary, such tools and methods can be used to compute properties that can better capture the complexity of ENMs composition and the influence of external conditions on toxicity, as opposed to difficult, time-consuming and expensive experiments (when experiments are possible). Despite the benefits of using materials modelling to calculate nano-descriptors, even until a few years ago, the scale and complexity of system simulations were challenging, and there was a shortage of models to predict important properties, such as the NM dissolution rate [8].

One may thus conclude that it is highly desirable to base nanosafety assessment upon accurate intrinsic and extrinsic properties, i.e. *features* that represent a particular structure and chemical nature of the ENMs of interest and how these properties affect or are affected by their (biological) environment. In the following, such features or *advanced descriptors* and their evaluation by means of physics-based models is discussed in more detail. Interpolation- or extrapolation-based methods like QSAR and ML will be served by a deterministic calculation of such advanced descriptors in a part of the parameter space that is expensive or hard to assess experimentally, or when experimental data are unclear or incomplete. On the other hand, physics-based models may also provide *direct* insight into the relation between observed ensemble properties (phenomena) and system parameters (simple and advanced descriptors) that stems from a systematic computational screening for one or more well-defined design parameters. Examples of phenomena considered in this review are protein absorption, NP aggregation and membrane binding. An instance of direct insight gained by computational

means is the finding that small nanoparticles (≪d, with *d* the membrane thickness, usually 4–5 nm) can simply permeate through the membrane, similar to small molecules, while large nanoparticles (≫d) will be fully engulfed or wrapped by the membrane upon binding. The binding characteristics for nanoparticle sizes comparable to the membrane thickness is still unclear [44].

A separation of descriptors into intrinsic and extrinsic provides a straightforward basis for the selection of the most appropriate physics-based model. Section 'Data-based models for linking ENM features to safety' discusses the current knowledge on ENMs and the (additional, advanced) features that need to be considered, as well as the progress on nanomaterial representations and data-based models, like nano-QSARs, that can receive and process the calculated nano-descriptors.

Data-based models for linking ENM features to safety

The critical role of a general representation of ENMs

In the attempt of gaining a deep rationale from data linking ENMs and their observed safety properties, an essential prerequisite is the ability to represent complex materials in well structured and machine-readable format. Unfortunately, to date, the lack of a standardized semantic characterization of the structural ENMs features and environmental variables makes it difficult to aggregate, curate and evaluate data from different sources and to use them for simulations or for training new (data-driven) models; thus making the meaningful integration of datasets a very demanding task. Web databases and repositories for chemical substances (e.g., ChEMBL, PDBe, ZINC15, Pubmed etc.) use standardized linear notations for substance identification (SMILES, SYBYL Line Notation, or InChI) [45,46]. These notations are also employed in deep learning tools to guide the generation of feasible molecular structures with desirable properties (e.g., Generative Adversarial Networks (GANs), Variational Auto-Encoders (VAEs), Recurrent Neural Networks (RNNs), etc.). The extension of these notations to polymers, mixtures, reactions, etc. has also been proposed, but ENMs entail additional challenges compared to conventional chemicals. The key properties of ENMs have a strong dependence on the physical and structural features. ENMs characteristics such as the spatial relationship between components, their relative sizes, etc., all play an important role in their inherent and emergent properties. What is more, many of these properties are environment-dependent, as they are affected by various external factors such as temperature and concentration through non-linear dependencies, making ENMs descriptions even more subtle and complicated.

A complete ENM representation, therefore, would require the inclusion of information on many more aspects than what current linear notations provide. Extending the current representations to include such advanced information content is not a straightforward task. Lynch et al. [47] have initiated an extensive discussion among various stakeholders and proposed a framework for an InChI standard applied to ENMs as well as a roadmap for its development. They aimed to address the variety of complex nanostructures, using a hierarchical approach that introduces new layers on the InChI notation for the size, shape, crystal structure, and ligand binding of the ENM, and, possibly, extrinsic and surface properties. Recently, Blekos et al. proposed principles for a more accurate, complete, flexible and incremental representation approach. The principles were demonstrated through the development of extensions to nano-InChI to encode morphology/mixture properties and statistical distributions of properties and to store metadata and enable their reuse.

The proposed representation framework could also provide the theoretical background to gradually capture the real particle dynamics under specific environmental conditions. [38] These are currently under consideration in the Nanomaterials InChI Working Group (https://www.inchi-trust.org/nanomaterials/) and their proposal for a new InChI standard. A hope for the near future is that an increasing number of advanced descriptors - as those discussed below in Section 'Physics-based models for nanostructure characterization' - will be gradually incorporated into such standard notations.

Data-based models for safety assessment

In nanoinformatics, popular data-based models include Quantitative Structure Activity Relationship modeling (so-called nano-QSAR/-QSPR) that utilizes Machine Learning (ML) and artificial intelligence to predict the desired response (e.g., biological activity, toxicity endpoints or any physicochemical property of interest). Those approaches are based on a set of computational and/or experimentally developed descriptors representing nanoparticle structure. As a representative example, Wyrzykowska, Mikolajczyk, et al. [48] proposed a concept in which a nanostructure is characterized by a triad that describes: (i) molecular structure; (ii) molecular descriptors; (iii) molecular properties that correspond to chemical composition and chemical structures of its components. The proposed triad covers the characterization of the intrinsic properties of nanoparticle structure (so-called system-independent or intrinsic (nano-) descriptors (S-descriptors in Fig. 2, left panel). [49].

Recent advancements in nano-QSAR modeling have introduced hybrid models that enhance predictive capabilities by integrating multiple modeling approaches. These hybrid models often combine molecular dynamics simulations data with machine learning techniques to better understand and predict the complex behavior of nanomaterials in biological systems and their interactions with the environment.[50,51] Although QSAR/QSPR can be certainly regarded as valuable tools to complement experimental studies on chemicals and nanomaterials, they come with multiple challenges that should be carefully analyzed. In particular, there is an absence of comprehensive methods to characterize nanoparticles, which are essential for a standard QSAR procedure or a dependable computational approach that accurately represents the unique features of nanostructures – in short, a lack of trustworthy nano-descriptors.

Furthermore, depending on the surrounding environment, ENMs features may change; thus, the data-based models should be developed based not only on the characterization of ENMs chemical composition/chemical structures of its components but also on the influence of the environment (including experimental conditions, or biological medium). In other words, the so-called system-dependent (extrinsic) nano-descriptors or the environment (E-descriptors) should also be considered to describe nanostructure as a whole (Fig. 2, right panel) [48,49].

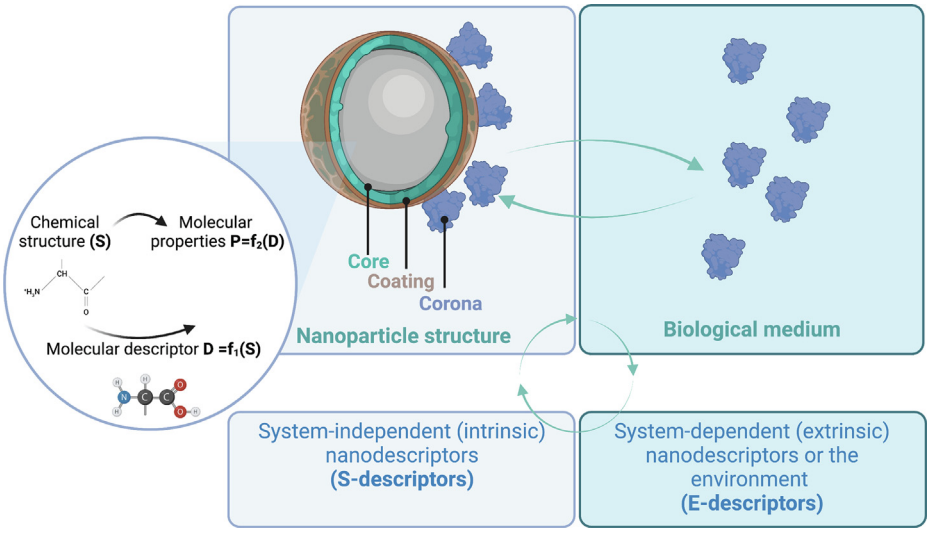

FIG. 2

Intrinsic and extrinsic descriptors: Intrinsic descriptors (i.e. system-independent, sometimes also referred to as *S-descriptors)* refer to the composition, components' structures, and properties that may be measured or calculated under a well-defined, unchanging set of conditions. On the other hand, extrinsic descriptors (i.e. system-dependent, sometimes also referred to as *E-descriptors*) represent the influence of the surrounding environment, and may be varying in time. See text and Section 'Physics-based models for nanostructure characterization' for more details..

In the following two subsections, we provide a brief critical review of intrinsic and extrinsic descriptors in preparation to the most recent and advanced computational approaches to estimate them, as discussed in detail in Section 'Physics-based models for nanostructure characterization'.

System-independent intrinsic features

System-independent (intrinsic) nano-descriptors refer to the composition, components' structures and properties that may be measured or calculated under a well-defined, unchanging set of conditions. A first example of advanced descriptors that were successfully applied for data-based modeling of nanoparticles was developed based on quantum mechanical calculations (so-called QM Descriptors) [52]. The QM Descriptors that describe the core chemistry of the structure were proposed in 2011 by Puzyn et al. [52] QM descriptors reflect the electronic form of a chemical compound. They are obtained by applying quantum mechanics to the appropriate molecular model of an ENMs structure. During the last 10 years, different research groups have been working to develop more sophisticated types of descriptors that are not related to detailed atomistic simulations. For example, in 2012, Toropov et al. [53] proposed SMILES-based optimal descriptors based on the encoded one-, two- and three-element SMILES attributes of a compound and can be calculated with the CORAL software [54]. This idea was then extended to the simplex representation of molecular structure (SiRMS). Here, another type of descriptor based on the Liquid Drop Model (LDM descriptors) was proposed by Sizochenko et al. in 2014 [55]. The methodology of LDM descriptors calculation assumes that an ENM can be represented as a spherical drop in which elementary molecules are tightly packed. At the same time, the density of clusters is equal to the particle mass density [56]. The proposed methodology is based on the thermodynamically most stable unit cell of the considered crystal structure, that is replicated in three dimensions. Afterwards, a spherical ENM is created by removing all atoms outside the indicated diameter. This is a clear simplification: In fact, bulk-cut surfaces will restructure, and even a spherical particle may exhibit regular regions as well as a variety of defects. [57]. Moreover, surfactants may reorder the surface differently [58]. Clearly, during the last decade, computational scientists and nanoinformaticians active within the European safety community have made a considerable effort to integrate knowledge from existing EU completed and ongoing projects within EU FP7 and HORIZON 2020 to develop a more comprehensive approach for nanostructure characterization [59]. However, we should also recognize that several key challenges related to the appropriate representation and description of ENMs structure are still ahead. Among others, those aspects have been highlighted by results provided by Mikolajczyk et al., Wyrzykowska, Mikolajczyk, et al. [60,49]. As far as intrinsic properties of ENMs are concerned, one of the main challenges is related to the description of the complexity of nanomaterials composition, thus clearly requiring the development and use of more advanced computational tools capable to evaluate/estimate descriptors that are difficult or even impossible to access by current experiments.

While such aspects are discussed in detail below within Section 'Physics-based models for nanostructure characterization', an additional critical point, partly related to the latter computational assessment of intrinsic ENMs features, is related to the lack of detailed material characterization in the published literature. Too often, indeed, only nominal composition, shape and size of nanomaterial are reported, with their possible coating only vaguely (if at all) defined. Clearly, those uncertainties add even more complexity and make benchmarking of computational material modeling particularly difficult.

System-dependent extrinsic features

While using QSAR/QSPR or other data-based models, a second challenge concerns representing the influence of external conditions (surrounding environment) [48]. A recently published

study [49] indicates that the system-dependent (extrinsic) nano-descriptors, also referred to as *environment descriptors* (E-descriptors), are much more critical for controlling and managing out the properties of ENMs than nanostructure characteristics themselves. Thus, in addition to standard characterization, the experimentalist should provide information about changes in the structure of the nanoparticles depending on the environment (the surrounding conditions). As a result, next to the core and coating, surface properties such as protein corona formation (so-called "biomolecular corona") play an essential role in characterizing ENMs' behavior and may be considered its fingerprint in a biological medium, see also Section 'Extrinsic advanced descriptors: Mesoscopic level' below for more details.

The system-dependent (extrinsic) nano-descriptors are crucial in describing physicochemical properties such as electrophoretic mobility or zeta potential value under specified conditions, reflecting the hydrophobicity, biomolecular corona, dissolution rate, sorption, surface reactivity, degree of aggregation/agglomeration, or ENM persistence.

Developing the system-dependent (extrinsic) nano-descriptors is challenging because the nanostructure may change during its lifetime due to its transport through different environments. In fact, nano-bio interactions are in principle driven by the nanomaterials fate in biological and environmental compartments, meaning how materials translocate, act as carriers of further toxicants and finally come in touch with biological targets. Under this perspective, the surface charge and wettability are considered the key determinants of the fate and behavior of nanomaterials dispersed in the exposure media.

Furthermore, the electrostatic interactions that keep particles dispersed, preventing or promoting contact with cell membranes, depend on the surface potential shown at the slipping plane (Zeta potential), as well as other important properties that drive nano-bio reactivity such as hydrophilicity or the surface interaction with biomolecules solubilized in the media. In this respect, the identification of the pH at which the Zeta potential is equal to zero (isoelectric point) allows making hypothesis on the type of acid/base behavior of surfaces and on the presence of charged molecules specifically adsorbed, as well as on the colloidal stability of nano-dispersed phases and of the occurring of hetero-aggregation phenomena [60–63]. In addition, the Zeta potential is very useful for the design optimization of surface functionalization strategies applied to control nanoparticles reactivity both for nanosafety and nanomedicine purposes, because it is predictive of the amount and type of coating that masks surface sites, providing a new biological identity to the dispersed phases [60,64]. Few studies report *in silico* models for the prediction of Zeta potential based on physicochemical intrinsic properties [60,65].

One of the most promising approaches that could cope with the challenge of extrinsic descriptors is an application of atomistic simulations complemented with coarse-grained models of ENM and biomolecules (see the dedicated sections below). In such multiscale approach, the coarse-grained models (nano and microscale) are parameterized and possibly validated by the data obtained from more detailed models at smaller scales (atomistic and quantum chemical). In this context, novel descriptors can also be derived from the statistics of adsorbed molecules after an analysis of the biomolecular corona [66,67].

Modern machine-learning methods

Machine learning (ML) is a branch of artificial intelligence (AI) that involves training computer programs to make predictions or take actions based on data. In ML, algorithms are designed to learn from data, instead of being explicitly programmed to perform specific tasks. The idea is to provide the computer with a large amount of data, and then use this data to teach the computer how to identify patterns, make predictions, or classify new data. At its core, ML includes four basic notions: algorithms, models, data, and training. ML algorithms are designed to learn from data and make predictions or take actions based on that data. The algorithms generate models, which are representations of the patterns and relationships in the data. The quality of the model depends on the quality and quantity of the data used to train it, as well as the algorithm's ability to learn from that data. Therefore, the process of training an ML model involves feeding it with labeled data, measuring its performance, and refining the model until it achieves satisfactory accuracy on new, unseen data.

The rationale for using ML approaches for nanosafety assessment is based on the need to efficiently process, analyze, and extract meaningful information from possibly vast amounts of data generated from both computational and experiments. ML algorithms can learn complex relationships and patterns from those data sets, enabling researchers to make predictions, optimize material properties, and identify novel materials with desired (e.g. less toxic) characteristics.

The advantages of using ML - as opposed to more traditional approaches - include:

- Accelerated materials discovery and optimization: ML algorithms can quickly process large amounts of data and identify potential new material candidates or modifications for further investigation.
- Reduced experimental and computational costs: By predicting material properties and hidden pattern identification, ML can help reduce the number of experiments and simulations required in the development process.
- Enhanced understanding of complex material systems: ML can capture non-linear relationships and intricate patterns in data, leading to better insights into the underlying physics and chemistry of materials.

However, there are also challenges and limitations associated with such ML approaches:

- Data quality and availability: ML algorithms rely on high-quality and abundant data, which can be a limiting factor in materials science, where data may be scarce, noisy or heterogeneous.
- Interpretability and explainability: ML models can be complex and difficult to interpret, making it challenging to understand the underlying reasons for their predictions and build trust in their outcomes.
- Overfitting and generalization: ML models may be prone to overfitting, namely they perform well on the training data but poorly on unseen data, thus reducing predicting accuracy.

ML techniques have been applied to a wide range of applications and fields including nanoinformatics [68]. In such a con-

text, ML methods have the potential to significantly impact the design, characterization, and safety assessment of ENMs. However, the availability of high-quality and abundant data is still an important aspect to address.

Among other challenges, a particularly important aspect associated to data-based modelling in nanoinformatics is the proper handling of highly imbalanced datasets. Imbalanced datasets are characterized by a skewed class distribution (e.g., over 1:100 observations in the minority class compared to the majority class), where usually the minority (underrepresented) class is the most interesting one to predict. For example, in an unbalanced dataset, there could be a majority class of nanomaterials that are composed of a single element (e.g. gold nanoparticles) and minority classes of nanomaterials that are composed of multiple other elements. As a result, when training a ML to classify the nanomaterials the model is likely to be biased towards the majority class and may not perform well on the minority class of high interest.

To address the issue of class imbalance, various data-level and algorithm-level approaches have been proposed over the last decades [69–71]. Data level approaches are addressing class imbalance via resampling (undersampling and oversampling), as well as via evolutionary algorithms for sampling, active learning for selecting the most appropriate data points or more recently adversarial learning algorithms for new points generation and meta-learning [72]. Random undersampling (i.e. removal of observations from the majority class), Near Miss, Tomek links (i.e. removal of boundary observations) are examples of undersampling (or downsampling) algorithms. The disadvantage of the latter methods is the loss of useful information as well as possible increase of the data bias.

Oversampling approaches include random oversampling (multiplication of observations from the minority class), SMOTE (finding nearest neighbors of the minority class observations and adding points on the line joining the point and the nearest neighbor) [73]. A review of SMOTE variants can be found in Ref. [74] SMOTE extensions to handle multiclass and multilabel (MLSMOTE) classification and regression (SMOTER) have been proposed [75], and, more recently, DeepSMOTE [76] and GraphSMOTE [77]. Oversampling has the advantage of retaining all the information and usually performs better than undersampling. However, oversampling may increase the probability of overfitting.

Another important aspect to be considered is that accuracy is not an appropriate metric for assessing model performance on imbalanced datasets. Algorithm-level approaches aim at modifying the loss metric, assigning different costs to penalize errors in each class (cost-sensitive training) or introducing new algorithms which can inherently deal with imbalanced data. Instead of accuracy, the recommended metrics are confusion matrix, precision and recall, F1 score, Kappa, Area under curve (AUC) (see also Ref. [74]). Ensemble algorithms using bagging and boosting (e.g., tree ensembles as Random Forest) are known to be more robust in imbalanced settings. Recent literature addresses handling imbalanced data (also known as *long tail learning*) by deep learning methods [78,79]. Imbalanced datasets are typical in high throughput screening [80–82] and chemogenomics [83].

ML algorithms can be broadly categorized into supervised and unsupervised learning. We begin with a discussion of supervised learning in the following subsection.

Supervised learning

Supervised learning works on labelled data, with the goal of approximating a function that maps input to labelled data. A prototypical example would be the ability to link a set of relevant intrinsic and extrinsic advanced descriptors for a number of ENMs to their observed (eco-) toxicological endpoints, e.g. the survival and/or reproduction rate of a chosen organisms. Supervised learning techniques include a large variety of algorithms and methods like Decision Trees, Random Forest, Support Vector Machines, various classifiers like k-Nearest Neighbors, Neural Networks and Instance-Based Learning methods (see Fig. 3, left panel) [84,85].

In one of the earliest applications of ML methods in manufactured nanoparticles, Fourches et al. used Support Vector Machine-based classification and kNN-based regression to generate Quantitative Nanostructure–Activity Relationship (QNAR) models to predict biological activity profiles of novel nanomaterials [86]. Later, Puzyn et al. [52] presented a method to quickly test the potential toxicity of engineered nanoparticles. They applied a multiple regression method combined with a Genetic Algorithm (GA-MLR) to create a model that described the cytotoxicity of 17 different types of metal oxide nanoparticles to bacteria *Escherichia coli* [52]. Gernand and Casman performed a regression-tree-based meta-analysis on rodent pulmonary toxicity exposed to uncoated, non-functionalized carbon nanotubes. They reported the application of Regression Tree models, Random Forest models, and a random-forest-based dose–response model [87]. Winkler et al. used novel sparse ML methods to model the biological effects of nanoparticles with various compositions, including iron oxide nanoparticles and gold nanoparticles. They employed Bayesian neural networks using both linear and nonlinear ML methods [88].

Evolutionary approaches have also been explored. Le and Winkler [89], for example, reviewed the use of artificial evolutionary methods for the identification and optimization of novel materials. They report uses of genetic algorithms to investigate the properties of bimetallic core–shell and titanium dioxide nanoparticles [90,91]. Martinez et al. presented decision tree models based on evolutionary algorithms that classified nanoparticle aggregates into morphological classes [92]. kNN algorithms have also been used to model the toxicological properties of nanomaterials. Wang et al. used a kNN algorithm to develop QNAR models for biological activity profiles like cellular uptake in various human cells and the ability to induce oxidative stress [93]. Kovalishyn et al. used kNN, random forest, and neural network methods to generate models for the analysis of eco/toxicological and physicochemical properties for metal and metal oxide nanoparticles [94].

Various Neural Network architectures have also been investigated in relation to predictive nanoinformatics. Gomez-Bombarelli et al. trained a Deep Neural Network to automatically generate novel chemical structures and demonstrated their method on small drug-like molecules [95]. Hataminia et al. used

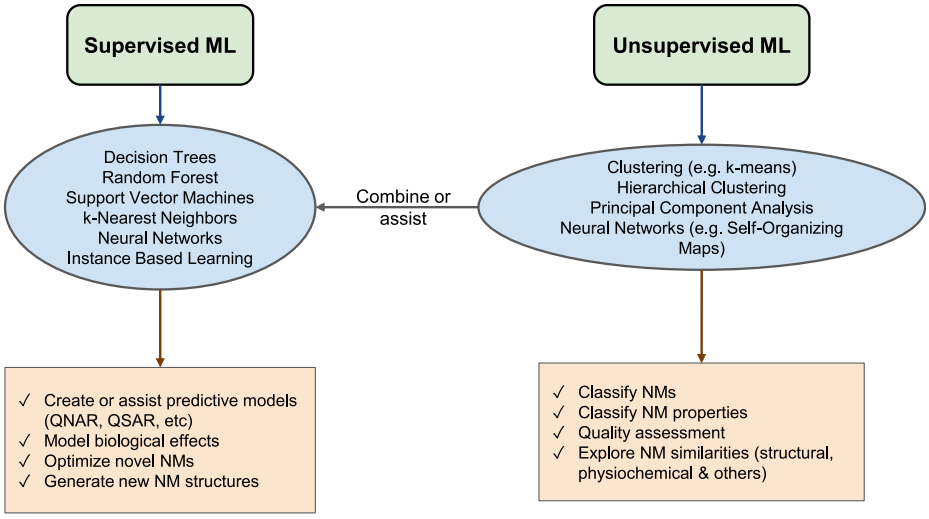

FIG. 3

The use of Supervised and Unsupervised Machine ML in predictive nanoinformatics: Supervised ML methods (Decision Trees, Random Forest, Neural Networks etc.) are often used to create or improve predictive models, while Unsupervised ML methods (Clustering, Self-Organizing Maps, etc.) are more often used to group and explore nanomaterials properties or in combination with Supervised ML methods.

a neural network to model the cytotoxicity of iron oxide nanoparticles to kidney cells [96]. Lazerovits et al. trained a Deep Neural Network to predict nanoparticle biological fate immediately after intravenous injection and tested it by predicting nanoparticle spleen and liver accumulation [97]. Very recently, Balraadjsing et al. investigated the performance of various supervised ML algorithms in the context of acute Daphnia Magna nanotoxicity prediction. Here, the Authors created classification models based on Random Forest, Neural Networks and kNN algorithms [98].

Li et al. used multi-target Random Forest Regression to predict the performance of sunscreen based on the type of titanium dioxide nanoparticle additives. Based on those models, they demonstrated the use of inverse design models that identify nanoparticle configurations based on desired sunscreen properties [99].

While the literature provides numerous examples of supervised learning techniques applied to nanoinformatics, the choice of an appropriate method is not always straightforward and often depends on several factors. For example, the choice between classification and regression usually depends on the type of problem being addressed: classification is used for categorical output variables, while regression is used for continuous output variables. Various factors, such as data size, dimensionality, complexity, and desired model interpretability, influence the choice of a specific algorithm.

Some general guidelines that can be considered when selecting an algorithm include the following: Decision Trees and Random Forest are well-suited for problems with mixed data types (numerical and categorical), and they provide easily interpretable results. Support Vector Machines (SVM) are appropriate for high-dimensional data and complex decision boundaries; however, they might be computationally intensive for large datasets. k-Nearest Neighbors (kNN) is a simple, instance-based method that performs well with small datasets but can be computationally intensive and sensitive to noise for larger datasets. Lastly, Neural Networks are effective for modeling complex patterns and relationships in high-dimensional data, but they may require more extensive computational resources and may not provide easily interpretable results.

Unsupervised learning

Unsupervised learning mostly deals with unlabelled data. This is particularly advantageous when data labelling is a resource-intensive task. The goal of unsupervised learning is to learn or discover the structure and patterns of the input data, and it is often based on exploring similarity among the input variables. In the case of nanomaterials, unsupervised learning is thus useful to explore similarities-based advanced descriptors. Clustering algorithms are the most representative algorithms of unsupervised learning. Popular examples include k-means, Principal Component Analysis, various Neural Networks like Self-Organizing Maps and Hierarchical Clustering (see Fig. 3, right panel) [100,85,85].

Unsupervised learning techniques have been used for the classification of nanomaterials or nanomaterial properties and quality assessment. They have also been used in combination with supervised or other statistical methods to assist, for example, the development of QSAR and neural network models. Wang et al. used Principal Component Analysis to analyze the structure toxicity relationship for various nanoparticles and identified the physicochemical properties of the nanoparticles that are risk factors for cytotoxicity [101]. Jha et al. modeled the toxicity of nanomaterials using a multivariate statistical analysis approach: through a multivariate Principal Component Analysis, they selected descriptors that optimally separated toxic from non-toxic nanomaterials [102]. Sizochenko et al. used a Self-Organizing Map in their hybrid approach to identify hidden patterns of toxicity among nanoparticles and to determine the underlying factors responsible for the toxicity [103]. Sizochenko et al., to evaluate the genotoxicity of metal oxide nanoparticles, developed another hybrid supervised and unsupervised ML

TABLE 1

Summary of the main challenges associated to data-based modelling in the ENMs safety context.

Identified challenge	Recommendations for future research
Representation and description of intrinsic properties of ENMs is much more complex and challenging as compared to chemicals	Effort should be devoted to include additional and more comprehensive information layers to the nano-InChI notation following the ongoing work in [46,38]. Additional intrinsic descriptors should be developed on the basis of experimental data but also advanced material modeling tools, such as quantum and atomistic simulations as discussed in Section 'Physics-based models for nanostructure characterization' below. Attention should be paid to the use of computational techniques as their robustness and the energetic needs associated with massive calculations might be a constrain in the next future.
Detrimental lack of well characterized ENMs data: go beyond nominal data	An effort should be made to design experimental and theoretical characterization including interfacial regions, key in the ENMs reactivity. In particular, the accurate description of the electronic structure at the surface is needed to capture specific interactions and mechanisms.
The influence of external conditions may play an even more critical role as compared to intrinsic ENMs properties	Experimental testing and characterization should focus on approaches (and report data) capable to track changes in the nanoparticle structure as a function of environmental conditions.
Unlike intrinsic features, extrinsic ones are time dependent and may change during ENMs lifetime while transported through different environments	As discussed in detail in Section 'Physics-based models for nanostructure characterization', some of the advanced material modeling approaches can be used both for estimating the value of extrinsic advanced descriptors under disparate conditions and for gaining further understanding on the basic mechanisms underpinning their change in time.
Training datasets for ML based models are often imbalanced with the most interesting material class being underrepresented	Use of undersampling and/or oversampling techniques trying to minimize loss of information and overfitting. Explore data-level and algorithm-level approaches. Develop new methods or adapt existing techniques, such as SMOTE and its variants. More focus on evaluating model performance using metrics suitable for imbalanced datasets, such as precision, recall, F1 score, Kappa, and AUC.
Addressing class imbalance issues in supervised learning for nanoinformatics.	Utilize data augmentation techniques, resampling methods (e.g., undersampling and oversampling), cost-sensitive training, and ensemble approaches to mitigate imbalance and improve prediction accuracy.
Ensuring adequate data quality and quantity for effective supervised learning.	Investigate ensemble methods such as bagging and boosting, feature selection techniques, and regularization methods to enhance model performance and stability.
Ensuring generalizability of ML models to new data.	Evaluate model performance on external datasets to ensure generalizability and robustness. Utilize cross-validation techniques to assess model performance on different subsets of the data. Pay attention to model overfitting, which occurs when the model is too complex and performs well on the training data but poorly on new, unseen data. Regularization methods such as L1 and L2 regularization can be used to reduce overfitting and improve generalizability.
Selecting appropriate ML algorithms for specific data structures and goals.	Consider various factors, such as data size, dimensionality, and complexity when selecting specific ML algorithms. For example, Decision Trees and Random Forests are good for mixed data types and provide interpretable results, while Support Vector Machines (SVM) are suitable for high-dimensional data and complex decision boundaries. k-Nearest Neighbors (kNN) is a simple, instance-based method for small datasets, while Neural Networks are effective for modeling complex patterns and relationships in high-dimensional data but may require more computational resources.
Integration of different ML models for a comprehensive understanding of ENM safety.	The literature provides numerous examples of ML models applied to nanoinformatics. However, the majority of the published studies focus on a single ML algorithm or approach. A comprehensive understanding of ENM safety requires the integration of multiple sources of data, including experimental and theoretical, and the use of different ML models to capture the complexity and variability of the system. The integration of different ML models can enable a more accurate and robust prediction of ENM safety by incorporating different types of information and reducing the uncertainty associated with each individual model. However, integrating different models poses several challenges, including data compatibility, model complexity, and the need for appropriate validation methods.

approach where they used a Self-Organizing Map to estimate relative distances between nanoparticles [85]. Kotzabasaki et al. also used a hybrid approach to develope a model for the prediction of genotoxicity of Multi-Walled Carbon Nanotubes. The latter Authors used the information derived from the experimental characterization of CNTs and a combination of Principal Compo-

nent Analysis and supervised classification techniques to improve the accuracy of the analysis in their parameters [84].

Similarly to the supervised learning, selecting an appropriate method is often contingent on the specific goals and data structure. Unsupervised learning aims to identify underlying patterns, structures, or relationships within the data without relying on labeled outcomes. Several factors, including data size, dimensionality, complexity, and the nature of the problem, influence the choice of a specific algorithm. Some general guidelines to consider when selecting an unsupervised learning algorithm are as follows:

Clustering methods, such as K-means, DBSCAN, and hierarchical clustering, are suitable for partitioning data into groups based on similarity or distance metrics. They are particularly useful when exploring the intrinsic structure of the data or identifying previously unknown subgroups.

Dimensionality reduction techniques, including Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE), are employed to project high-dimensional data into lower-dimensional spaces. These methods can aid in data visualization, noise reduction, and improving the performance of other machine learning algorithms.

Autoencoders, a type of neural network, are effective for unsupervised feature learning and representation. They can be used to reduce the dimensionality of data, denoise data, or learn more complex and abstract features.

For the sake of clarity, in the Table 1 below, we summarize what we believe are the most important open issues in data-based modeling assessment of ENMs safety. Concurrently, we provide recommendations on possible effective actions that could help addressing such challenges in the near future.

Physics-based models for nanostructure characterization

In general, physics-based modelling techniques can be grouped according to the dominating length and time scales (or resolution), and equivalent representations at different resolutions can be related to each other via a coarsening or fine-graining procedure, in which degrees of freedom (electrons, atoms) are averaged out (forward-mapping) or introduced (back-mapping). For instance, the familiar quantum methods at the *electronic* level explicitly consider electrons and atoms, enabling them to calculate several intrinsic properties of the material with great precision, including the electronic bandgap and surface reactivity. At the *atomistic* level considered by classical molecular dynamics, explicit electronic structures like bonds are only implicitly represented in a force field. The resulting enhanced sampling rates can be exploited to study larger system, for instance, the long-term conformational dynamics of one protein or the interaction between a nanoparticle and several biomolecules.

Moving to even larger scales, interactions between ENMs and their surroundings, reflected by extrinsic ENM properties, often involve processes that exhibit a considerable disparity in length and time scales. From a computational perspective, they, therefore, require yet another - *mesoscopic* - level to be tractable *in silico*: a level that is and will not be genuinely within reach of atomistic methods for some time, despite the continuous advances in computer power. For instance, even when investigating the uptake of tiny particles such as fullerenes (<1 nm) by human cells (10–100 μm), which first cluster to larger aggregates in solution before binding to a membrane due to their hydrophobicity, the spatial scales that have to be adequately represented span many orders of magnitude.

This challenge is even more significant for the (competitive) binding of one or more molecules like proteins onto a nanoparticle, since they generally experience conformational changes on various length scales upon absorption. The same holds for time, as diffusion or translocation processes in dense molecular environments like membranes and intracellular spaces take place at timescales that are orders of magnitude greater than the fastest vibrations inside a molecule - molecular bond stretching - that set the elementary time scale. At the most basic level, this challenge can be faced by experiment, via trial and error for individual setups. Yet, as the experimental resolution is also essentially limited at the bottom where many relevant molecular mechanisms of interest take place, one may formulate theoretical or computational answers for more well-posed questions.

In the following, we will concentrate on three extrinsic properties for which classical molecular dynamics significantly falls too short: nanoparticle clustering, the formation of a protein-corona, and nanoparticle uptake by a lung membrane.

A subset of advanced descriptors stemming from such calculations is reported in the Appendix, where more detailed information is reported. We note that, especially at the mesoscopic level, this overview of progress in the field of nanoparticle simulations is not exhaustive. It should also be noted that much effort has been put into ensuring *equivalence* at the highest level of resolution, e.g. by proper parameterization of tight-binding methods at the quantum level and determination of proper force fields for atomistic molecular dynamics. At the mesoscopic level, many developments are quite recent and the issue of equivalence is more complicated to satisfy, which hampers the application range of genuine multi-scale methodology based on systematic coarse graining for the aim of deriving descriptors for specific materials. For this reason, many mesoscopic investigations have focussed on evaluating mechanisms for generic setups, which cannot be directly related to specific systems that are needed for advanced descriptors. For details of such activities, we refer the reader to published reviews [104].

Intrinsic advanced descriptors: Electronic level

The quantum–mechanical computation of nanoparticle properties relevant to nanotoxicology is still a challenge, as size and time scales are shorter than typical biological phenomena. The key aim is to reveal the electronic structure nature in complex toxicity mechanisms. A valid strategy involves the accurate determination of descriptors (heat of formation, lattice energy, enthalpies of cation detachment) and band structure (bandgap, HOMO and LUMO levels) that could be related to toxicity endpoints by data-based modelling. This approach was successfully proved in Nano-QSAR models [52].

The computational costs of *ab initio* calculations, typically density functional theory (DFT), remain a strong limitation to the deployment of quantum-chemical descriptors, mainly due the size of realistic nanoparticles (from few to hundreds of

nanometers) and the complexity of the surface region (often unknown in experimental data). A detailed quantum chemical description of those systems severely increases the time needed to acquire data.

Faster computational schemes use the so-called *quasi-ab initio* methods based on DFT, applying expansions of the atomic electronic structure using the Tight Binding approximation (DFTB) [105]. Such techniques, which are 3,000 to 30,000 times faster than regular DFT, allow considering more realistic sizes and conditions such as solvation or molecule adsorption. To compute dynamic effects, *ab initio* molecular dynamics (AIMD) is a promising technique, yet at a too high computational cost.

In the last years, some of the aforementioned techniques have been successfully applied to investigate nanosized titania with sizes 2.2–4.4 nm. An example of a spherical 3 nm titanium dioxide nanoparticle is shown in Fig. 4, panels A and B. The model contains more that 1200 atoms, 6000 electrons and it requires 50000 cpu hours [59] for routine DFT characterization. In panel A, the titanium dioxide nanoparticle is shown, with an inset highlighting some surface atoms ; it can be seen that they differ from the bulk geometry. A scheme of DOS (Density of States) is shown as an example of electronic descriptor associated with chemical reactivity. In panel B, the decrease in energy during a DFTB simulation in water is shown, and the corresponding RDF (radial distribution function) for pairs of atoms (right), accounting for the surface structure in the proximity of the solvent molecules.

Other properties recently reported in the literature for titanium dioxide nanoparticles concern crystallinity [106] and hydration [107,108]. Interestingly, DFTB tools have been recently applied to understand biological processes such as the inactivation of SARS-CoV-2 virus [109], or the coating by biological molecules [109,111]. It is however advised to carefully check the applicability of DFTB methods for electronic structure characterization as their accuracy critically depends on the parameterization; for instance the oxygen vacancy energy for different titanium dioxide termination is not always well captured by DFTB [112].

Extension of quantum-based methods to the field of (nano) toxicology should be naturally observed in the next years, provided that a fundamental understanding of the physicochemical behavior behind toxicological effects is achieved. This involves several aspects.

First, the identification of new relevant descriptors based on a detailed knowledge of the electronic structures of realistic complex interfaces. The quality of the descriptors is crucial for the precision of the results and should be carefully addressed. Experiments focused on revealing the role of the surface in the toxicology mechanisms should be designed to properly interplay with electronic structure calculations.

Second, we envision interesting progress in the near future due to the development of faster and more accurate techniques to account for the complex structure and reactivity, based on machine and deep learning [113–116] or automatized structure screening tools (like Grand Canonical [117], evolutionary [118] or clustering [119] algorithms). However, they should still be tested on realistic nanotoxicity models and could possibly become routine methods accounting for composition, coating and biological media.

Intrinsic advanced descriptors: Atomistic and Molecular level

In atomistic (also called classical or all-atom) MD simulations, atoms are represented by particles interacting according to Newton's second law. In such an approach, electron dynamics is neglected, dramatically speeding up the simulations (as compared to DFT calculations), and system of larger dimensions become affordable (with the typical computational domain having up to dozens of nanometers edge length). Given the size range of nanoparticles, MD is a powerful tool to study such systems in a detailed way that is often far from the experimental capabilities.

With respect to the interaction of nanoparticles with biological moieties, MD was recently used to determine the membrane binding energies for nanoparticles made from three bare materials (silver, silica and titanium dioxide) of three different sizes (1, 3 and 5 nm diameter) via MD-based Potential of Mean Forces (PMFs) using umbrella sampling [120]. To cover the diversity of responses that are possible for nanoparticle binding of a real human lung membrane, which is facilitated by a mixture of different lipids, each with their own phase behavior, a membrane with an equivalent lipid composition was considered. Calculating binding free energies via the PMF along a normal reaction coordinate (typically the nanoparticle-membrane distance) for real materials provides a new and useful advanced descriptor for QSAR, especially when the particle size effect can be explored within the domain where the membrane response is sensitive. For non-spherical nanoparticles, however, the PMF calculation becomes less trivial and involves several reaction coordinates. Moreover, the nanoparticle sizes within reach of MD typically only cover part of the domain of interest, and are significantly below the experimental sizes.

While current MD results are exciting and make a first step towards the *in silico* assessment of advanced nano-descriptors for nanosafety, they also illustrate the essential challenge associated with more resolved computational approaches. In the first place, experimental nanoparticles are usually at least 50 nm in size, meaning that the maximum sizes considered in current studies [104], i.e. in the order of the membrane thickness, are still far from most of the real applications, especially given that many advanced descriptors are size-dependent and that it is unknown if and how one may extrapolate. The most important drawback, however, is the cost of individual simulations given the immense nanoparticle design space, which comprises size, shape, elasticity, charge, composition/hydrophobicity, and surface modification [104], if we leave dynamic surface modifications, such as the formation of a protein corona, by the surrounding medium out of the picture. The majority of existing MD studies of nanoparticle-nanoparticle and nanoparticle-membrane interactions thus focus on the molecular understanding that can be gained for specific experimental setups. A very promising application of MD is the calculation of binding energies between vitamins and specific nanoparticle interfaces, represented as an infinite slab, using metadynamics, for the purpose of validating QSAR predictions and nano-descriptors [121].

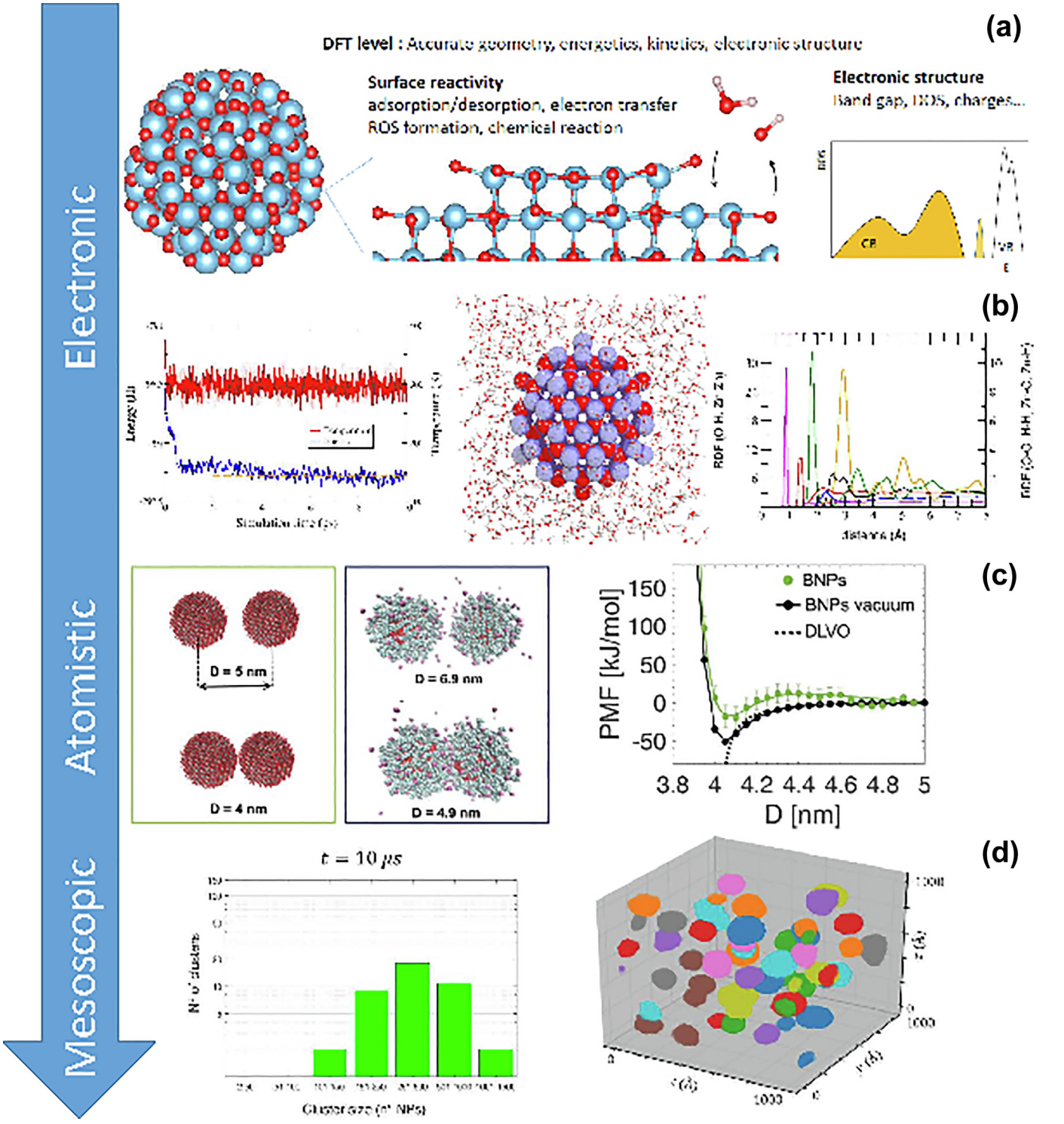

FIG. 4

Multi Scale Materials Modelling: from the electronic level to the atomistic description up to the mesoscopic level. Panel A: DFT level allows accounting explicitly for electrons, providing accurate descriptors for geometry, energetics and electronic structure; Panel B: (left) Energy and temperature stabilization of ZnO nanoparticle of 2 nm in water medium at 300 K for 10 ps simulation (centre) ZnO nanoparticle of 2 nm in diameter in water (right) Radial Distribution Function of ZnO nanoparticle of 2n in water medium at 300 K. Panel C: bare and coated nanoparticles on the left and the potential of mean force for the approaching of two identical nanoparticles in a vacuum and in water and the comparison with the original DLVO theory (Reproduced from Ref. [9] with permission from the Royal Society of Chemistry); Panel D: Brownian Dynamics simulations of nanoparticle clustering, cluster size distribution on the left and snapshot of the simulated system on the right (Reproduced with permission from Mancardi et al., MDPI Manomaterials 2022 [59]).

Theoretical descriptors for small organic molecules [122] can be readily calculated with various levels of theory to represent most molecular features. In the case of nanoparticles, the size is the obvious limiting factor for the calculation of whole particle nano-descriptors where all atoms are considered. To address those issues, full particle molecular nano-descriptors developed

by Tämm et al. [123,124] were calculated directly and solely from the structure of the considered nanoparticles. Such calculations can be performed with LAMMPS program [125] together with Buckingham potential [126] and Wolf summation [127]. The developed set of atomistic nano-descriptors is based on the chemical composition, potential energy, lattice energy, topology, size, and force vectors. Reported studies based on the latter approach can cover different properties of the nanoparticles including also differentiating such properties in core and shell regions of the nanoparticle [123,124,128]. The core region usually captures similar properties to bulk material, while the shell region is expected to account for the special nanoparticle properties. While there are different methods for the *in silico* generation of nanoparticles, there is still a need to define and study the property differences in core and shell regions of the nanoparticle. Therefore, a software tool has been recently developed to define the shell depth for all nanoparticles (https://nanogen.me/shell-depth). The tool requires the *xyz* type files as input and calculates the optimal shell depth for nanoparticles together with the average coordination numbers for different atom types in the nanoparticle.

As far as reactive phenomena are concerned, attempts to go beyond classical force field simulations could be based on reactive force fields (ReaxFF) [129–131]. Such an approach unlocks the possibility of studying larger computational systems at a fraction of the cost as compared to quantum-level simulations (i.e. DFT and DFTB). However, it requires a delicate tuning of several dozens of parameters against energetics from first principle simulations at lower scales. This fine-tuning often leads to a lack of generality of force field parameters and may need a case-by-case optimization.

Nanomaterial dimension and shape are important to determine the corresponding toxicological endpoints: the formation of aggregates from small nanoparticles modulates the amount of material entering the cells, regulating its interaction with the DNA. Aggregates are too large to study using Classical Molecular Dynamics: this kind of system can be simulated at the mesoscopic level, as discussed below.

Moving from intrinsic to extrinsic descriptors

For the physics-based determination of intrinsic, and also for extrinsic descriptors, DFT or classical MD are the first methods of choice. In contrast to mesoscopic methods, which are based on averaging, they incorporate the most direct and material-specific reaction, chemical, conformational, and interaction detail possible. Yet, for all these methods, the total computational effort invested for the determination of material properties is proportional to the number of degrees of freedom (basis functions, atoms or groups of atoms) that have to be taken into account multiplied by the number of discrete (time) steps needed for numerically stable minimization or equilibration of the system at hand. Even on exascale high-performance computing environments, where the original calculation/simulation can be divided into parts and distributed over multiple processors, nowadays even up to a million, this proportionality represents a serious computational limitation [132].

Intrinsic properties, which are determined for (solvated) isolated molecular systems, generally fall inside the reach of the most detailed DFT or MD modelling, because the system size and equilibration times usually remain manageable. Yet, although quantum methods have a clear advantage over classical MD, their applicability is restricted to small systems compared to most experimental ENMs, see Section 'Intrinsic advanced descriptors: Electronic level', meaning that one should extrapolate or concentrate on surface properties. Classical MD pushes this boundary up somewhat, and is capable of simulating, for instance, the formation of a stable ligand coating around a nanoparticle of relevant size, or the exploration of the protein folding funnel on a second scale in explicit solvent. The most detailed methods are thus particularly useful for providing implicit descriptors.

This situation changes when dealing with extrinsic descriptors. Nanoparticle aggregation and uptake, as well as biomolecule absorption, are all at play upon the release of nanoparticles into a biological environment. The length and time scales involved in these processes are many orders of magnitude greater than the elementary Angstrom and femtosecond scales of classical MD. In particular, when computationally evaluating the nanoparticle interaction with a surrounding bio-matrix of structured lipid envelopes and unstructured mixtures of shorter and longer biomolecules, care should be taken in selecting proper system sizes. After all, these systems should sufficiently represent the key elements in the larger open system and be large enough to avoid computational artifacts due to boundary conditions.

As such, multi-scale approaches, in which the system is evaluated at a coarse time and length scale at some stage, become compulsory [132]. In this context, fine-grained electronic and atomistic methodology still serves a distinct role as a reference for equivalence, via systematic mapping, or by providing structural input for realistic evaluation at a coarser level, e.g. relevant nanoparticle-nanoparticle interaction potentials or (static) protein conformations for docking to a nanoparticle.

A general issue in assessing ecotoxicity by multiscale physics-based modelling is that, historically, mesoscopic modelling primarily aims at providing fundamental insight into phenomena at a structural or kinetic level rather than determining accurate extrinsic descriptors, since it puts stringent constraints on the way atomistic or molecular detail is absorbed into a coarser description. Also after coarsening it is wise to build up complexity step-by-step, testing the averaging procedure at each increment. Therefore, in membrane binding, the current focus is on calculating advanced extrinsic descriptors like binding energies for passive rather than spontaneous (protein-induced) binding. In the foreseeable future, the addition of more players, including the many lipid types that render a cellular membrane, membrane-bound and adsorbed proteins, and the cytoskeleton, will become an option, and the particular role of these actors can be studied *in silico*.

Extrinsic advanced descriptors: Mesoscopic level

The simulation of aggregation phenomena involving thousands of nanoparticles is simply too demanding to be carried out by all-atom MD, and coarse-graining procedures are mandatory to make such simulations feasible. An effective strategy is to employ Brownian Dynamics (BD) simulations, see Fig. 4, panel D [9,59]. Here, each nanoparticle is represented by a spherical bead charac-

terized by the nanoparticle's diameter and interacting with other beads according to analytical equations describing the interaction potential (e.g. fitted to calculated expression by means of classical Molecular Dynamics for a nanoparticle pair, see Fig. 4, panel C). Applying this coarse-graining procedure makes it possible to determine new molecular descriptors ruling particle aggregation that could be fed into QSAR models [59]. Additional challenges faced by developing general computational and theoretical modelling for understanding the nanosafety of ENMs stem from the role of the environment in determining the observed toxicity. Once the ENM enters a living organism, it gets in contact with the biological molecules, in particular proteins and lipids.

Nanoparticle-proteins and nanoparticle-lipids interaction could in theory be investigated using Brownian Dynamics simulations, which can run even on a 16 core workstation, provided that all pair interactions are known; in practice, this has not yet been done because the calculation of the free energy profiles for each pair by all-atom MD is too computationally demanding. Anyway, this could be an interesting attempt to bridge the gap between the molecular simulations scale and the experimental scale.

The oldest approach for computationally determining material properties, i.e. the continuum mechanics pioneered in the 19th century by Cauchy, is in fact most suited for screening purposes, since it combines modest computational costs for realistic system sizes with a few effective screening parameters. This screening idea is at the basis of continuum Self-Consistent Field Theory (SCFT), which was developed to describe phase behavior and phase separation dynamics in block copolymers (represented as flexible chains) based on an implicit molecular representation [133].

Rigid objects like nanoparticles have also been incorporated into SCFT, primarily for the purpose of modelling polymer nanocomposites [134], and we refer to early papers for details about the different approaches [135–137]. While these field-based methods possess a clear advantage of efficiency over particle-based methods like AAMD and CGMD, which stems from the choice to deal with ensembles rather than individual chains, and SCFT interactions are of the desired many-body type by definition, this is offset by the serious disadvantage of not being able to represent specific interactions at the molecular level and having no access to conformational detail. In addition, the commonly used excluded volume interactions in SCFT do not allow for phase transitions that can play a role in membrane binding processes. Only recently, hybrid particle-field (hPF) approaches such as hPF-MD [138] and single chain in mean field (SCMF)[139] have introduced the ability to combine particle-based (atomistic or segmental) molecular detail with the efficiency and multi-body nature of Hamiltonians from continuum theory like SCFT. Phase transitions and/or coexistence have also been added recently [140]. Until these hybrid methods offer a validated solution for the need to combine efficiency with specificity and molecular detail, we conclude that SCFT is useful for the investigation of general phenomena, but not suited for the extraction of extrinsic material descriptors.

A popular and very efficient representation of lipid membranes is that of a thin elastic sheet without any molecular detail. Shape, dynamics and responses to deformation are dictated by a continuum Helfrich free energy that depends on a few collective properties like bending rigidity and lateral membrane tension, which can be directly related to specific membrane compositions via particle-based simulation. An extended Helfrich model developed later provided straightforward conditions for nanoparticle uptake, *i.e.* the balance of energy needed to stretch and bend the membrane around the nanoparticle and the energetic gain of nanoparticle binding. The latter is provided by the adhesion energy density of (coated) nanoparticles in the contact region and, as was later found, also in near-contact regions. It should be noted that extracting adhesion energy density for real materials from more detailed descriptions is, unfortunately, a far from simple task [141]. Early on, Deserno et al. used the continuum model to show that a tensionless membrane can only adopt two states: an unbound state where the membrane is flat, or a state where the nanoparticle is fully wrapped by the membrane [142,143]. Accounting for factors that were missing in this original Helfrich-based analysis, such as the membrane thickness and interaction range, Raatz et al. and Spangler et al. found that also partially wrapped cases could be stable [144,145]. While providing important energetic insight, and, therefore, being of potential use for restricted but quick screening, the lack of lipid detail already seriously hampers the use of such methods for extracting extrinsic material descriptors. Another disadvantage is that the nature of the membrane deformation upon nanoparticle binding is not an outcome but required *a priori*, introducing a risk of overlooking alternative binding mechanisms. The historical solution to this issue is to employ highly coarse-grained particle-based models with implicit solvent for studying generic membrane dynamics and nanoparticle-membrane interactions. Using such a model, the wrapping characteristics for nanoparticles up to 40 nm was considered, *i.e.* the entire range for the mechanism is expected to switch, and a discontinuous transition from partial to full wrapping was predicted around 10 – 15 nm nanoparticles [145,146]. Since these models lack solvent, which is known to modulate the (free) energy landscape, and they also lack the resolution to distinguish between different lipids, they do not represent a decisive step forward in the search for accurate descriptors.

Summarizing, one may conclude that these efficient molecule-based mesoscopic methods are useful for extracting information about general balances and mechanisms for systems of relevant size, but they were never meant to provide extrinsic nanoparticle descriptors. Just this, balancing (chemical) information and efficiency with the aim to retain the necessary detail, is the main purpose of recent developement in systematic coarse-grained methodology. Although there are several ways to perform systematic coarse graining, depending on the characteristics of the reference atomistic system that one wants to reproduce, they are all based on lumping groups of atoms into CG particles or beads. The most popular method, CG Martini, combines 2–4 heavy atoms in a single bead. Generating a description in terms of CG beads does not only reduce the computational load, but it also softens the interactions. As a result, also the system evolution is significantly accelerated. Methods based on such types of coarse graining have been applied to study lipid partitioning in general, the binding of elastic nanoshells [147,148], the adhesion of anisotropic nanoparticles [149,150] and of functionalized nanoparticles [151,152]. Also the role of bending and adhesion in the distribution of multiple nanoparticles inside the membrane has been investigated, both in terms of a generic representation in a highly

CG method [153] and for more chemically resolved CG representations [154–156]. Yet, while these CGMD studies have either been designed to properly represent a particular experimental nanoparticle or to obtain insight into more general binding mechanisms, very few have focused on the challenge of developing a transferable representation or map from the atomistic to the coarse-grained domain. Yet, determining such a map that is valid for all nanoparticles of the same material is a prerequisite for the extraction of advanced descriptors and trends that enable extrapolation.

One very recent example of such a new development is the special CG nanoparticle representation within the familiar Martini CG approach that is required for studying binding and translocation pathways of realistic silver nanoparticles across solvated lipid barriers in the lungs, see Fig. 5. Whereas the modelling community has thus far generally approached the fundamentals of such large-scale phenomena via implicit-solvent continuum [142,143] or highly coarse-grained descriptions [145,146,157], a core–shell CG representation was developed that is transferable with respect to size and enables the simulation of relevant nanoparticle sizes including solvation effects, in the size range where interesting switching in binding behavior is expected [158]. The development of this transferable map was based on matching potentials of mean force (PMFs) for silver nanoparticles obtained using all-atom molecular dynamics [120]. The systematic development of transferable CG representations for the other materials, such as silica and titanium dioxide that were also considered in atomistic studies, is a future desire. As the determination of binding free energies for CGMD NP by standard methods like umbrella sampling gets prohibitive with increasing size, string methods could be employed as an alternative [159].

A key requirement in nanosafety assessment is how to include systematically a detailed molecular description of ENM-protein interactions. In biological environments, proteins organize on ENM surfaces forming the so-called nanoparticle protein corona (NPC) structures which play a central role in biological interactions and nanotoxicity [160–162]. The NPC formation around a variety of nanoparticles was evidenced and characterized in terms of its biochemical composition by several experimental studies [163,164]. However, its effects on biological interactions and implications for nanosafety considerations remain largely unknown [163,164–169]. Similarly as above, a main research challenge is how to develop efficient yet accurate computational methods and tools that can bridge the gap between a detailed, molecular-level description of ENMs interacting with solvents and biomolecules such as proteins at an atomistic level, and the much larger scale (i.e., tens or hundreds of nanometers to micrometers) corresponding to biological structures such as NPCs or cellular membranes [170]. Coarse-grained computational methods of protein-covered nanoparticles are often limited to modeling entire proteins as single particles. Such models are successful in showing how nanomaterials type, size, and shape can lead to diverse protein composition of the NPC [171,172]. However, detailed atomistic aspects of modeling protein interactions are required to calculate other key experimentally-relevant mesoscopic descriptors such as the

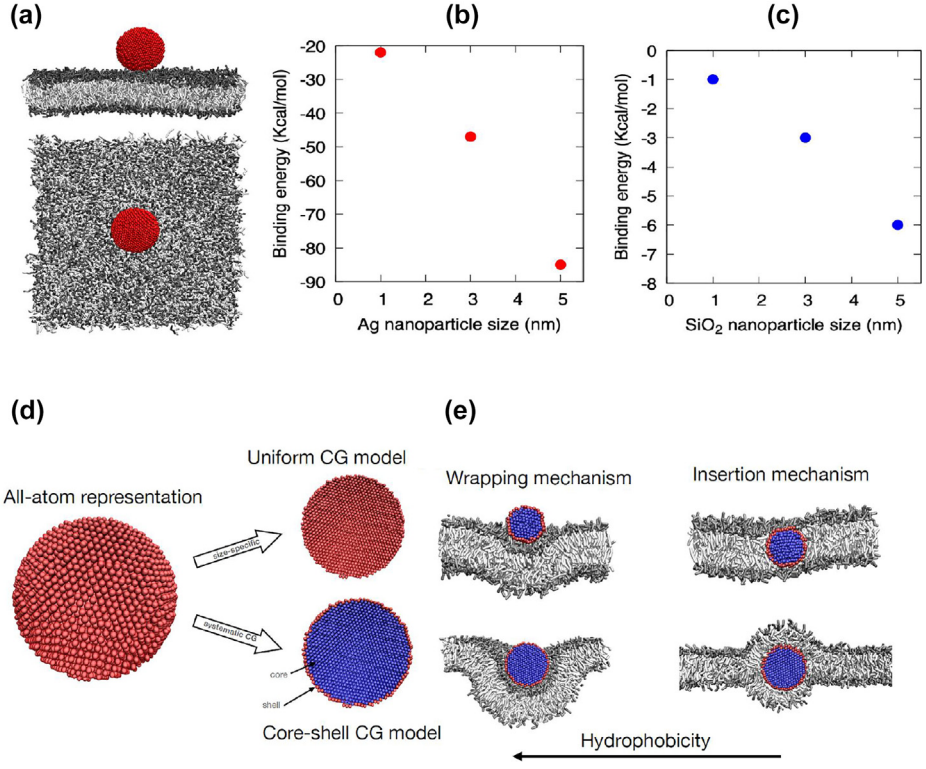

FIG. 5

(a) All-atom MD of nanoparticle/membrane/water system, (b) all-atom MD binding free energy for three different sizes of Ag nanoparticles, (c) all-atom MD binding free energy for silica for three nanoparticle sizes (graphs reproduced from Ref. [120] with permission from the Royal Society of Chemistry). (d) Standard uniform CG model and the new core–shell CG model, reproduced with permission from Singhal et al., MDPI Nanomaterials, 2022[158]. (e) With increasing hydrophobicity, the mechanism of direct insertion into the model lung membrane switches to a wrapping mechanism..

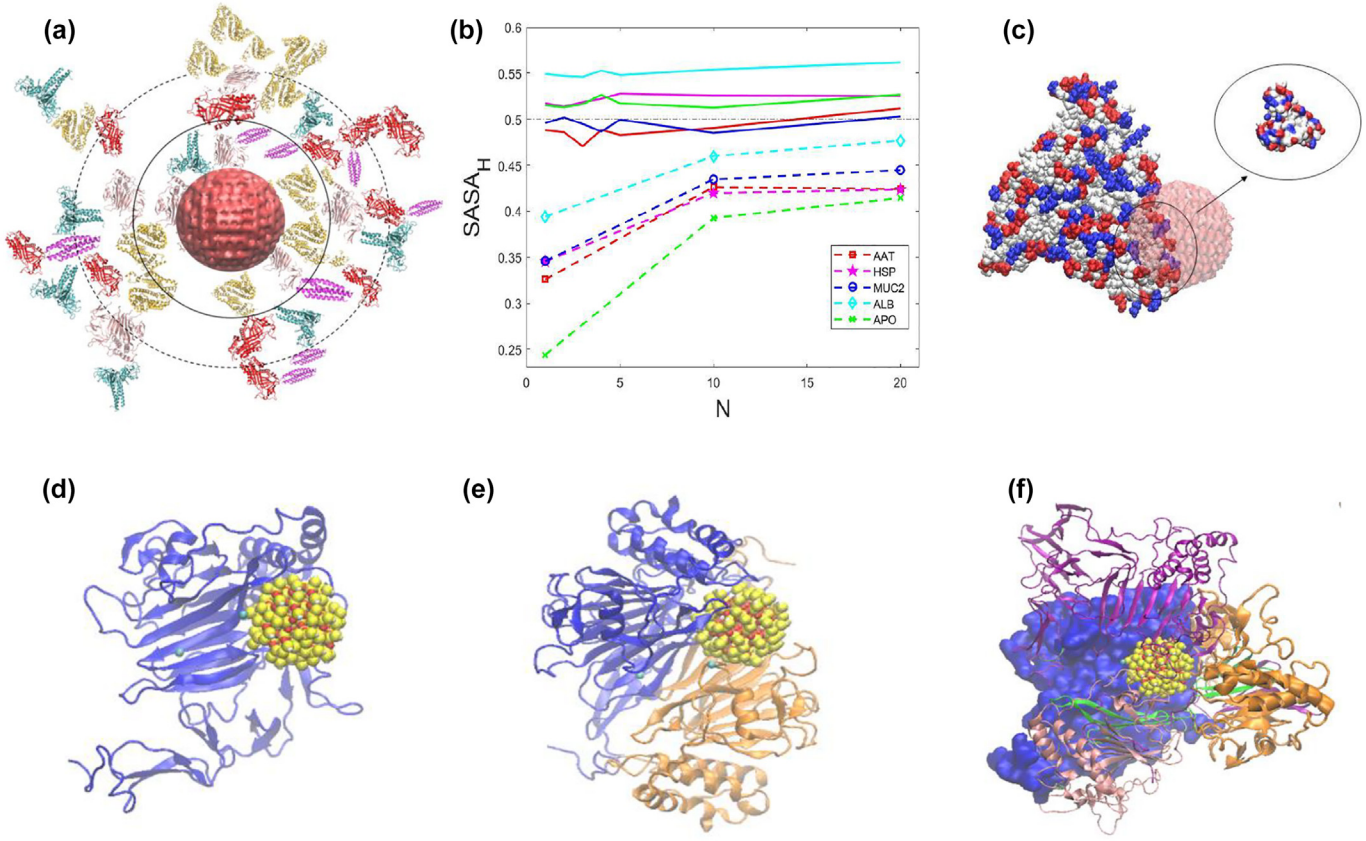

FIG. 6

(a) Schematic atomistic model of an NPC. A "soft corona" layer (dashed line) of loosely bound proteins surrounds a "hard" corona layer (continuous black line), proximal to the ENM's surface. Even for spherical nanoparticles, the overall shape and biophysical properties of the NPC surface will depend on its composition. (b) Comparing values of the $SASA_H$ for various coronas around a 4 nm spherical silica nanoparticle (dashed) with the values calculated for similar protein aggregates without including a nanoparticle (continuous). (c) Atomistic corona models allow the identification of protein residues that may play significant roles at the nanoparticle-protein interfaces. (d-e-f) Building an atomistic model of an NPC by sequential docking of protein structures (mucin, pre-equilibrated using MD) on a spherical silica nanoparticle. Mesoscopic descriptors such as $SASA_H$ can be estimated as statistical averages over results from docking multiple representative structures.

hydrophobic fraction of the solvent accessible surface area ($SASA_H$) (see Fig. 6(a-b-c)). Recent developments in atomistic and multiscale computational methods allow unique opportunities to probe the detailed molecular mechanisms that modulate interactions at bio-nano interfaces [173,174,29]. This approach can be extended to the calculation of mesoscopic biophysical descriptors for an NPC and relies on simplified models allowing further computational studies of protein interactions with ENMs (see Fig. 6(a-b-c)) [163]. This has the potential of (i) unveiling the role of specific proteins in NPC's stability and biophysical properties (e.g., hydrophobic surface area, charged patches), and (ii) quantifying the way in which nanoparticle corona properties and protein–protein interactions in the corona are modulated for different nanoparticle types. First, all-atom molecular dynamics (MD) simulations of key plasma proteins (e.g., human serum albumin, fibrinogen, immunoglobulin gamma-1 chain-C, complement C3, and apolipoprotein A1) can be used to study adsorption on typical nanoparticle surfaces (e.g., titanium dioxide or silica). For binary protein–protein interactions (e.g., only two interacting proteins) it is possible to perform exhaustive atomistic MD simulations, both in the vicinity of nanoparticle surfaces and in bulk to compare directly the results and infer the influence of the presence of specific nanoparticles on the dynamic and thermodynamics aspects of protein–protein interactions. Fig. 6(c)) illustrates the possibility to identify residues crucial to protein-nanoparticle interactions in a specific system (here, human serum albumin-titanium dioxide). In the second stage, the molecular mechanisms of protein-nanoparticle interactions are probed by looking at the dynamic and structural proteins of several proteins (and possibly lipids) in the crowded environment of nanoparticle coronas, using also molecular docking simulations and, depending on systems size, coarse-grained simulations of mixtures of multiple proteins [175,176] that can investigate the formation of the protein layer on the nanoparticle surface, as illustrated in Fig. 6(d-e-f). Preliminary studies on multi-protein docking on nanoparticles, suggest that knowledge of protein composition and conformations (e.g., refined from MD simulations) can be used to estimate the overall biophysical properties of NPCs, such as the hydrophobic fraction of their solvent-accessible surface area, and surface charge distributions [177]. Outstanding challenges in modeling the interactions of biological molecules in contact with ENMs are to extend the capability of docking programs to include more than just a few tens of proteins [178–181], and to include detailed information on the specific corona molecular composition that is seldom available [170]. Additionally, besides proteins, it is expected that

future studies will also include (i) other components such as lipids [182] and glycans [183] which play pivotal roles in ENMs uptake and could also be key for the modeled ENM systems, as well as (ii) an accurate description of the corresponding surface functionalization [184]. Finally, metallic nanoparticles deserve a special mention as they are routinely employed in cancer therapy, where they need to be selectively delivered to the tumor tissues. Among all metals, gold nanoparticles are widely used as radiosensitizing agents because of their biocompatibility and simplicity of synthesis [185]. Coated metallic nanoparticles are used in catalysis, self-assembly, imaging, drug delivery, and sensing applications. Metallic nanoparticles are very sensitive to the local environment because of a phenomenon called "localized surface plasmon resonance" deriving from the collective oscillation of surface electrons. [186] When coated with a monolayer ligand, the metallic nanoparticles' properties such as metal reduction and colloidal stability can be adjusted for the desired application [187]. Atomistic and mesoscale simulations allowed an understanding of the atypical distribution of multiple ligands on gold and silver nanoparticles observed in the experiment [188], as well as the adsorption of biomolecules on gold nanoparticles of different sizes [189]. We report in Fig. 7 an example of how molecular modelling simulations (here all-atom molecular dynamics simulations) can be used to investigate the behavior of metallic nanoparticles, in particular, to capture adsorption phenomena of polymers. Interestingly in [190], it is clearly shown that the design of shape and topology of the surface in metallic nanoparticles is a rather effective strategy to control the preferential polymer coating in some particle regions as compared to others. Furthermore, using similar simulations tools, other works [188] have investigated the precise patterning of coadsorbed surfactants on silver and gold nanoparticles. Hence, by targeting more effective control on the crystallographic features of metallic nanoparticles, we can envision an improved control of ENMs coating thus also significantly affecting their toxicological properties.

Discussion: Challenges and perspectives

On one hand, from the above overview, it clearly emerges that the accuracy in predicting possible hazards of ENMs critically relies upon the ability to use features beyond what can be accessed in typical experimental tests and characterization. Importantly, those features depend on intrinsic and extrinsic properties aiming at describing both materials and biological environments.

Nonetheless, when it comes to the direct link of ENMs to their expected toxicity endpoints, it is fair to say that the current status of the development of the hardware and algorithms does not allow a brute-force assessment of nanosafety. A more viable approach is expected to be the calculation of intrinsic and extrinsic advanced descriptors using a plethora of methodologies mostly developed and used in other fields (e.g. materials modeling and biochemistry) to extract input features for data-based or statistical models (e.g. QSAR, ML algorithms) to finally link them to the toxicity endpoints.

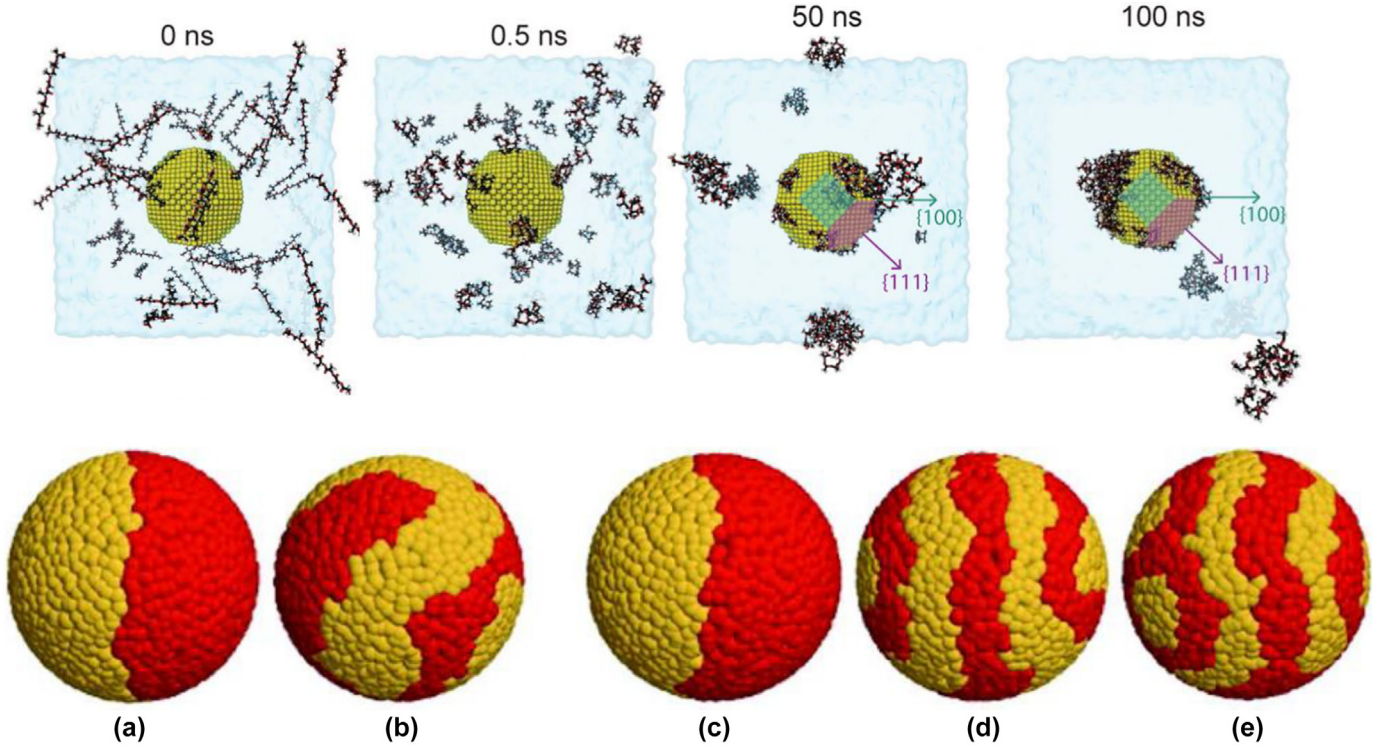

FIG. 7

Upper panel: MD snapshots of self-assembly simulations of 60 PLGAs and the Au nanoparticle in aqueous solution at 0, 0.5, 50, and 100 ns, reproduced with permission from Cappabianca et al., ACS Omega, 2022[190]; Bottom panel: Equilibrium structures obtained by mesoscale simulations of self-assembly of binary mixtures of surfactants with varying length difference or bulkiness difference on a spherical nanoparticle. Dark (red) beads and light (yellow) beads represent head groups of the two species of surfactants, from Ref. [188]. See text for more details..

However, even the latter approach comes with formidable challenges, mostly associated with the size and complexity of ENMs of practical use. On one hand, ENMs of experimental interest may have dimensions orders of magnitude larger than their computationally affordable counterparts. On the other hand, precise compositions of particles and their coating are often known with little detail level: This calls for a crucial effort of the relevant scientific community in future experimental works where, in addition to the valuable measurement of toxicological endpoints, a more comprehensive characterization beyond nominal values of ENMs is requested.

Furthermore, currently, an interesting (and perhaps necessary) approach seems to be the *hybridization* of pure physics-based models with disparate data sources. In particular, due to a practically unlimited number of different ENMs with great chemical and geometrical variety, a truly extended adoption of advanced descriptors in data-based models for nanoinformatics looks inconceivable without leveraging high-fidelity multiscale modeling data from both literature and crude or analytical (yet computationally efficient) approximations models. As a representative example, biased classical molecular dynamics simulations can certainly be used to accurately compute the Potential of Mean Force between nanoparticle pairs. At the same time, classical approaches such as the theory of Derjaguin-Landau-Verwey-Overbeek (DLVO) cannot be discarded and efforts should be devoted to finding new approaches capable of orchestrating and integrating such multi-fidelity and multi-source data. Another example of a similar synergy has been described above in the manuscript and it has to do with CGMD based simulations of cell membranes and the corresponding continuum Helfrich free energy models.

One possibility would be the adoption of descriptors from crude approximation models, characterized by a lower fidelity level and available literature data as a subset of features for ML models where physics-based model results are used as training sets. Such an approach has been proven successful in significantly improving ML model predictions in the presence of a small or incomplete training dataset [11,193]. We thus envision a more comprehensive and multi-layered approach where detailed physics-based models (first layer), literature/experiment data (second layer), and crude approximation models (third layer) are synergetically managed for the estimate of relevant descriptors by means of surrogate (statistical) and ML models (see Fig. 1, bottom panel). In this respect, further research is requested to investigate to what extent the small variance of high-fidelity data from physics-based model predictions combined with the high-variance (and low-fidelity) of other sources can deliver descriptors predictions with high/medium fidelity and variance that are suitable for QSAR/QSPR models.

It is also worth stressing that, beyond the mere usefulness of advanced descriptors for data-based models targeting nanosafety assessment, further improved physics-based models describing phenomena from the electronic up to the mesoscopic level can offer the unique opportunity of gaining insights at the nanolevel, which are hardly accessible by experimental techniques and therefore remain critical for i) unveiling basic mechanisms behind the possible hazardous character of ENMs; ii) possibly complement information from less detailed and less demanding models.

For all the above reasons, we expect and hope that the progress of high-performance computing power combined with advanced tools for acceleration of atomistic simulations [194] can soon calculate advanced descriptors more competitively as compared to experiments in terms of both the amount and quality of generated data.

Another interesting aspect to highlight is related to the possible exploitation of the large body of knowledge available in the context of modern computational approaches to investigate nanoparticles in other technological fields such as catalysis and materials science [195–196,77,197–200]. Specifically, we refer to:

1. High-throughput screening of nanoparticles in drug delivery [201]
2. Barcoded nanoparticles for high throughput *in vivo* discovery of targeted therapeutics [202].
3. Computational high-throughput screening of alloy nanoclusters for electrocatalytic hydrogen evolution [203]

It would be desirable that the Safe and Sustainability-by-design strategies in toxicology could take advantage of this knowledge and tools, combining efficiently biology and medicine with physics and chemistry.

Furthermore, an even larger amount of data on nanosafety of ENMs will likely be generated in the near future. It is therefore of increasing importance to comply with the FAIR principles, so that metadata and data can be efficiently reused in data-based models for predicting the hazard of ENMs. Specifically, in this respect, the role of databases for collecting and storing a large amount of curated and well structured data is likely to play a major role soon. As such, we discuss one prototypical case study in the following subsection.

Finally, as a long-term goal, we expect that nanoinformatics models based on advanced descriptors could be integrated with AOPs, to better assess the potential exposure throughout the entire life of the nanoparticles. In this respect, modern grouping strategies taking into account the mode of actions and developed based on ML techniques processing data from omics studies appear particularly promising and therefore it will be specifically discussed below in a dedicated subsection.

Databases: The eNanoMapper case study

Storage and organization as well as the organization of curated data from disparate sources within databases appear critical for nanosafety assessment. To this end, below we review a prototypical case study. The eNanoMapper database is an open-source chemical substance data management solution [204], adopted by more than 20 European projects and facilitating the Findable, Accessible, Interoperable and Reusable (FAIR) data collection and reuse of the nanosafety community. To provide aggregated findability, accessibility, and interoperability across project-specific databases, the Nanosafety Data Interface (https://search.data.enanomapper.net) was created, and currently represents one of the largest searchable nanosafety data collections [205]. The eNanoMapper is based on data and software originally developed to represent industrial chemicals and related experimental or calculated data. It was one of the first cheminformatics platforms to offer open REST Application Programming Interface (API) sup-

porting integrated services such as data, descriptor calculations, and ML [206,207]. A visual representation of the eNanoMapper data model is reported in Fig. 8, where substances are characterized by names and IDs, which can be multiple, the composition refers to the components of the material (core, coating, chemical structure), each of them having different properties; the same material can have different compositions. A protocol consists of measurements of a specific endpoint in given conditions, related protocols form an investigation entity; finally, different substances can be grouped into an assay entity when the same protocol applies, giving an extremely flexible structure [208]. With the explosive growth of material databases, ML frameworks, and their success in material modeling, it is critical to explore the link between the estimated material properties and experimentally measured safety or functional properties. In NanoInformaTIX [36], both experimental data from selected use cases and also calculated descriptors are stored in the eNanoMapper database and an effort is devoted towards providing open source libraries to facilitate integration with data analysis frameworks and developing exploratory data analysis methods. The validation of computational models relies on high-quality experimental data; such data may not always be complete, and it is necessary to identify data gaps and, eventually, to generate additional data based on experimental and theoretical chemistry and on biology. The goal of this computational and theoretical endeavor is the realization of safe nanoparticles and nanomaterials. To fulfil such expectations, theoretical descriptors must be translated into measurable parameters, and this challenge entails a deep knowledge of the mode of action of nanoparticles that lead to harmful outcomes. On another level, identification and estimation of theoretical drivers of toxicity can facilitate the prioritization of experimental tests, which, due to the uncontrollable inhomogeneity of any ensemble of nanoparticles, is always needed for a robust risk assessment of nano-enabled technology.

Grouping approaches

Regulatory processes, which often rely on *in vivo* testing, are outpaced by the increasing number of ENMs on the market. To cope with this situation, the lack of data and to ensure the safety of new materials, grouping approaches emerge as an interesting method. Those approaches are accepted within the overarching EU chemicals regulation REACH (EC 1907/2006) and are commonly used to consider more than one chemical at the same time [209–211]. Within an established group, data gaps can be filled by read-across. Here, existing data on a particular (eco) toxicological endpoint linked to one or several source chemicals can be employed to estimate the same property of one or more target chemical(s).

Chemicals can be grouped on the basis of well-defined physicochemical similarities, like common functional groups, precursors, and/or breakdown products. However, ENMs pose an additional challenge compared to chemicals, since there is a very limited understanding of how individual physicochemical parameters influence cellular uptake and toxicity. In addition, the properties of ENMs can change depending on the surrounding medium and over time. To establish grouping approaches for ENMs, it is essential to understand how the individual physicochemical properties are linked to toxicity. To support a grouping justification, additional information on a common Mode of Action (MoA) or toxicity mechanism is advantageous [209]. The MoA of a substance describes the functional or physiological

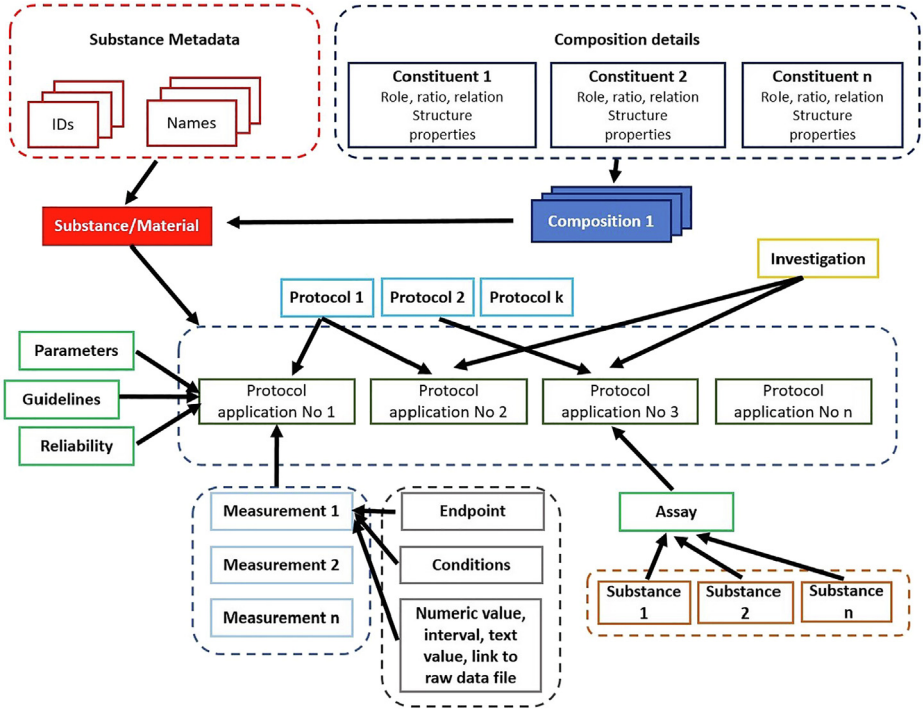

FIG. 8

eNanoMapper/Ambit data model adapted from Ref. [208]. Substances are characterized by their composition and identified by name and ID, which can be multiple. Complex relations between the substance components can be specified.

changes it causes to a living organism or cell. One of the drawbacks is that toxicity mechanisms are only partially understood, and for plenty of ENMs variants, the precise MoA remains elusive. In this regard, the potential of systems biology to contribute to the development of reliable grouping approaches for ENMs should not be obliterated. Modern omics-based approaches (transcriptomics, proteomics, metabolomics), in combination with sophisticated bioinformatics and data analysis tools, are very important to characterize toxicity pathways and unravel relationships between individual physicochemical properties and cellular responses. This knowledge can then be applied in the context of grouping and categorization. Moreover, omics approaches are of relevance for the development of AOPs and to establish reliable, comprehensive testing strategies building on the known MoA [212]. AOPs are a conceptual construct that integrates known information from various sources in a sequential chain of causally linked key events that cover different levels of biological organization (i.e. cellular, organ level) starting with a molecular initiating event leading to the final adverse outcome [213]. The knowledge gained from grouping approaches can then be directly used for SSbD of ENMs. Currently, there are several ENM grouping frameworks with different approaches [214–216]. However, there are only very few case studies for which the frameworks have been applied to ENMs.

Recently, several publications have taken advantage of different techniques, including ML approaches. [217–219] Additionally, the incorporation of omics datasets to support the development of more accurate grouping strategies of ENMs take into account the mode of action [220–223].

Bioinformatics and ML techniques are thus essential when approaching cellular effects comprehensively, particularly in combination with high-throughput techniques like omics. Fig. 9 depicts the ML random forest approach used for the grouping of ENMs, with this strategy, the biological activity of ENMs can be predicted provided that physicochemical properties are known. Omics studies are a massive source of data sets, which comprehensively describe the cellular alterations caused by any treatment, importantly in this case, by ENMs. Thus, omics methods are highly useful to identify the MoA of ENMs to be employed in the grouping approaches as additional biological descriptors. So far, most of the efforts to understand the MoA of ENMs have been undertaken in the field of transcriptomics [224,225]. However, proteomics can be even more informative since it depicts the cellular alterations much closer to the phenotype than transcriptomics. The challenge here is the standardization of the methods, particularly for data analysis and interpretation [226,227].

Meta-analysis of publicly available proteome data targeted to specific organ alterations is being carried out to investigate the MoA of ENMs within the organ, based on proteome alteration evidence. Relevant datasets from publicly accessible proteomics databases such as PRIDE are identified in the first step. These data do not necessarily involve only ENM treatments, but also other alterations like disease, cancer, and chemical treatments. A recently developed workflow by the BfR for standardized data analysis

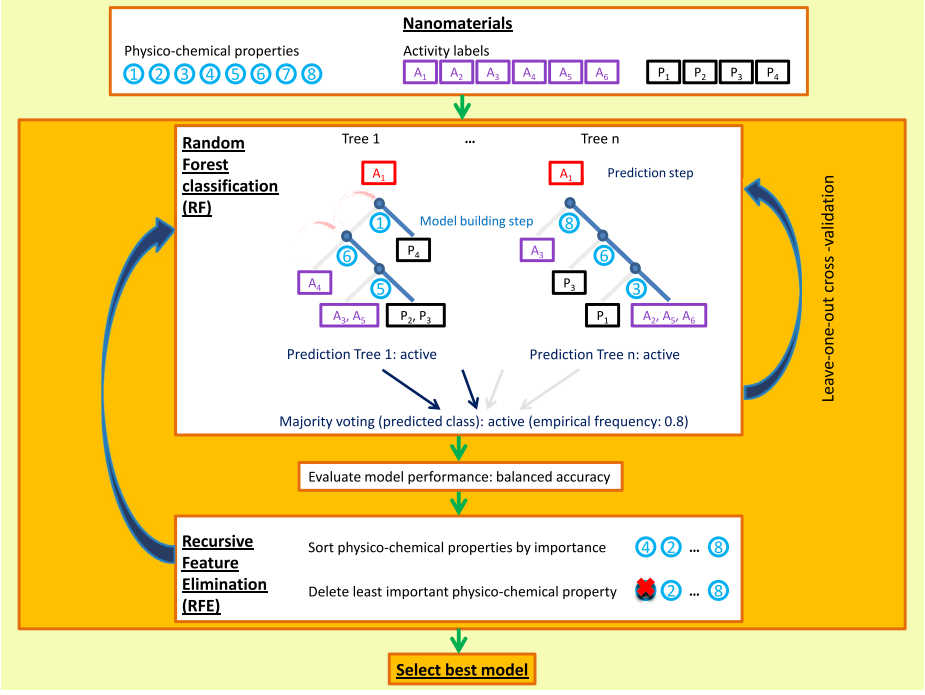

FIG. 9

Schematic representation of the random forest approach used for grouping of ENMs. ENMs are described by a set of different physicochemical properties, and an activity label is assigned to each ENM (based on the outcome of biological assays). ENMs are then subjected to a random forest model, in which ENMs are divided into active and passive materials, based on splits made on their physicochemical properties in each tree. Recursive feature elimination is used to select only the most important physicochemical properties for achieving maximum accuracy. To address and avoid a large overfitting bias, the model is validated in a leave-one-out approach. This approach can be used to predict the biological activity of new ENMs, for which the physicochemical properties are known.

was applied to the data sets. These results are then integrated into proteomics data *de novo* generated from studies evaluating the effect of ENMs in vitro. Correlations between nanomaterials effects and organ-specific alterations can thus be detected.[228]

Concluding remarks

Gaining a clear and deep rationale behind the nanosafety characteristics of ENMs is a multifaceted issue still posing formidable challenges: nanomaterials present far more complex physico-chemical properties than their macroscopic counterparts, having high surface reactivity and the ability to enter living cells, potentially causing damage to cells or entire organisms.

In this work, mostly focusing on a computational perspective, we made an effort to review and discuss state-of-the-art physics-based models for computing both intrinsic and extrinsic ENMs properties that are crucial for setting up reliable data-based models for nanosafety assessment. In this spirit, one major aim of this work was the identification of the most critical roadblocks towards computer-aided support of nanosafety assessment. Importantly, we have identified and discussed opportunities in advancing the field and in opening research directions, as conveniently summarized in Tables 1 and 2. We have extensively illustrated recently-used methods for computing advanced descriptors and the current associated challenges mainly leading to low data variance despite the expected higher fidelity. Hence, possible suggestions on hybridization strategies for moving towards models with both higher data variance and fidelity, as well as the inclusion of Adverse Outcome Pathways (AOPs) are envisioned.

We stress that the reported analysis and suggested guidelines for future reflect our current best understanding of the field after several years of discussion among experts in multi-disciplinary yet disparate related fields. As such, we hope that this work can serve as a stimulus for future multidisciplinary research, and could thus help nucleate further breakthroughs in the computational nanosafety assessment of ENMs.

Data availability

This is a review article and data availability does not apply.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 814426 (NanoInformaTIX project).

TABLE 2

Summary of recommendations and possible future research avenues in the context of using material modeling techniques for computing advanced descriptors.

Modelling technique	Recommendations for future research
Quantum Computations (DFT and DFTB)	Those are among the most time-consuming methods for extracting advanced descriptors. The most recent methodologies based on machine learning algorithms are promising to expand the set of systems, but attention should be paid to their robustness and the energetic needs associated with massive calculations. Pluridisciplinar physico-chemical approaches, such as those widely used in other technological fields like catalysis, should be used on a regular basis, to capture toxicity mechanisms on the molecular level. This may lead to new descriptors and increase the predictive power.
Reactive Atomistic Simulations	Suitable reactive force fields are still to be developed for accurately predicting the development of ENMs surface electrical charges. In this respect, reaxFF [191] or the more recent Machine Learning based potentials [192] are expected to likely play an important role in coping with sufficiently large particles beyond the capacity of standard DFT and DFTB computations.
Classical All-Atom Molecular Dynamics (AAMD)	Calculate the adsorption affinity of small metabolites on nanoparticles using enhanced sampling techniques such as metadynamics. Molecular nano-descriptors in [124] have proved very computationally effective to cope with large size particles, nonetheless they are limited to non-metallic particles: Additional effort should be spent to relax this constraint. AAMD is also needed for the molecular docking of proteins to ENM surfaces (see below). To speed up future calculations, an MD database of typical conformations and their corresponding thermodynamic weights could be pre-built for proteins that play a major role in interacting with ENMs (i.e., albumin, mucin, etc.) to increase the accuracy of docking-based calculations of nano-descriptors (e.g., $SASA_H$).
Molecular Protein Docking	Current docking software is optimised for protein-drug and protein–protein interactions (PPIs). Docking approaches rely on initial AAMD studies (see above) of the proteins used. Future research would benefit from optimization for sampling efficiently and accurately protein-ENM surfaces (i.e., inorganic materials). Future docking programs should be able to handle multiple molecules, including larger proteins with diverse conformations and in a much larger number than currently possible (i.e., from tens to hundreds or even thousands). One possible approach to bridge the gap between the limitations of docking programs and the complex and large-scale nature of PPIs and protein-ENM interactions is the development of advanced data-driven and machine learning approaches.
Coarse-Grained Molecular Dynamics (CGMD)	Estimating absorption energy densities by AAMD for pre-screening by continuum Helfrich methods. Calculating CG PMFs for all possible core–shell combinations to generate a matrix for mapping material-specific atomistic PMFs. When addressing extrinsic descriptors for ENMs-cell membrane binding, overcoming the current passive configurations, by adding many lipid types rendering the cellular membranes as well as membrane proteins and the cytoskeleton.
Brownian Dynamics	Simulating nanoparticles in contact with biological molecules to bridge the gap between the molecular modelling scale and the experimental scale. More effective approaches are needed to compute and use PMFs in case of non-spherical or anisotropic particles.

Appendix AList of physics-based descriptors that can be calculated through materials modelling techniques

Scale	Method	Advanced descriptor	Units	Typical values for a 3 nm TiO_2 nanoparticle	Notes and limitations
Quantum - Space scale: 10^{-11}-10^{-9}m Time scale: static	DFT & DFTB	Standard enthalpy of formation	eV	-9 to −8	Very accurate but ideal only for small nanoparticles up to 4 nm in diameter. Highly dependent on the basis set and the pseudopotential of choice. Absolute energy values are only useful for comparison with other structures.
		Total energy	eV	-90000 to −10000	
		Electronic energy	eV	≈ −88000	
		Energy of the Highest Occupied Molecular Orbital (HOMO)	eV	≈ −3.2	
		Energy of the Lowest Unoccupied Molecular Orbital (LUMO)	eV	≈ −3.1	
		HOMO–LUMO energy gap	eV	≈ 0–3	
		Valence band width	eV	7.3	
		Conduction band width	eV	2.6	
		Fermi level	eV	≈ −3.3	
		Hydration energy	eV	≈ −143	
		Vertical ionization potential	eV	≈ 3.4	
		Vertical electron affinity	eV	≈ −3	
		Oxygen vacancy formation energy	eV	3 to 4	
Molecular, atomistic - Space scale: 10^{-9}-10^{-8}m - Time scale: static	Quantum mechanics at DFT level	Chemical composition (x9)	-	N.A.	Fast to calculate because there is no need of expensive energy minimization. Not for metallic ENMs. Calculable also for larger nanoparticles up to 60 nm. Possible use in QSAR and ML models. Detailed description of those quantities can be found in Refs. [122,124]
		Potential energy (x9)	eV	-75 to −18	
		Topology descriptors (x9)	-	2.5 to 6	
		Size descriptors (x3)	Å, $Å^2$, $Å^3$	N.A.	
		Lattice energy descriptors (x5)	eV, eV/Å, eV/$Å^2$, eV/$Å^3$	-110 to 0	
		Force field descriptors (x27)	-	N.A.	
	ReaxFF MD	Surface charge	C/m^2	-0.06	Need a reactive force field developed for a similar system
		Particle-Membrane binding free energy	eV	≈ 0	
		Aggregation free energy	kJ/mol	54	
		Solvent Accessible Surface Area (SASA)	$Å^2$	61.27 ś 0.68	Note: See Ref. [120].
Molecular scale with atomistic resolution. Space scale: 10^{-10}-10^{-8}m	AAMD and Molecular Docking	Hydrophobic SASA fraction ($SASA_H$)	% SASA	33.5–54.5	
		Polar SASA fraction ($SASA_P$)	% SASA	40–50	Note: estimated for 4 nm TiO_2 NPs with diverse protein compositions of

(continued on next page)

(CONTINUED)

Scale	Method	Advanced descriptor	Units	Typical values for a 3 nm TiO_2 nanoparticle	Notes and limitations
					the NP corona. Variance can be smaller for specific systems.
		Negatively and positively charged SASA fractions ($SASA+$ & $SASA-$)	% SASA	6–16	
Mesoscopic - Space scale: 10^{-9}-10^{-6}m - Time scale: 10^{-9}-10^{-6}s	Brownian Dynamics	Deviation from Smouluchowski's aggregation kinetics theory	-	N.A.	Simulations are fast but require previous AAMD simulations to get the force field parameters
	Coarse Grained MD	Membrane bending rigidity and other membrane related descriptors	-	N.A.	

References

[1] D. Astruc, Chemical Reviews 120 (2020) 461–463.
[2] A. Gizzatov et al., Advanced functional materials 24 (29) (2014) 4584–4594.
[3] A. Cardellini, M. Fasano, E. Chiavazzo, et al., Phys. Lett. Sect. A Gen. At. Solid State Phys. 380 (20) (2016) 1735–1740.
[4] S.S. Mukhopadhyay, Nanotechnol. Sci. Appl. (2014) 63–71.
[5] B.S. Sekhon, Nanotechnol. Sci. Appl. (2010) 1–15.
[6] E. Chiavazzo, P. Asinari, Nanoscale research letters 6 (1) (2011) 1–13.
[7] E. Chiavazzo, P. Asinari, International journal of thermal sciences 49 (12) (2010) 2272–2281.
[8] A. Haase, F. Klaessig, EU US Roadmap Nanoinformatics 2030 (2018) 1–126.
[9] A. Cardellini et al., Nanoscale 11 (9) (2019) 3925–3932.
[10] Nat. Nanotechnol. 16 (6) (2021) 607. https://www.nature.com/articles/s41565-021-00911-6.
[11] Y. Zhang, C. Ling, npj Comput. Mater. 4 (1) (2018) 28–33.
[12] Communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions chemicals strategy for sustainability towards a toxic-free environment, in: COM/2020/667 Final, 2020.
[13] S.F. Bezerra et al., Contact Dermatitis 84 (2) (2021) 67–74.
[14] K. Jagiello, K. Ciura, Nanoscale 14 (18) (2022) 6735–6742.
[15] K. Jagiello et al., Small 17 (15) (2021).
[16] K. Jagiello et al., Environ. Sci. Nano 9 (5) (2022) 1675–1684.
[17] C. Caldeira, R. Farcal, C. Moretti, L. Mancini, H. Rauscher, J. Riego Sintes, S. Sala, K. Rasmussen, Safe and sustainable by design chemicals and materials: review of safety and sustainability dimensions, aspects, methods, indicators, and tools, Publications Office of the European Union, 2022.
[18] B. Stieberova et al., J. Clean. Prod. 241 (2019).
[19] A. García-Quintero, M. Palencia, Sci. Total Environ. 793 (2021) 148524.
[20] E. Marcoulaki et al., NanoImpact 23 (2021) 100337.
[21] L.J. Johnston et al., NanoImpact 18 (2020) 100219.
[22] S. Gottardo et al., NanoImpact 21 (2021) 100297.
[23] P. Nymark et al., Small 16 (6) (2020).
[24] S.H. Doak et al., Small 18 (17) (2022).
[25] L. Bajard et al., Environ. Res. 114650 (2022).
[26] S.I. Gomes, J.J. Scott-Fordsmand, M.J. Amorim, Nano Today 40 (2021) 101242.
[27] C. Westmoreland et al., Regul. Toxicol. Pharmacol. 135 (2022) 105261.
[28] OECD, Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models, 2014.
[29] A. Afantitis et al., Comput. Struct. Biotechnol. J. 18 (2020) 583–602.
[30] S. So, J. Rho, Nanophotonics 8 (7) (2019) 1255–1261.
[31] I. Kim et al., Microb. Pathog. 149 (2020) 104290.
[32] NanoSolveIT, NanoSolveIT Horizon 2020 project, https://cordis.europa.eu/project/id/814572.
[33] Gov4Nano, Gov4Nano Horizon 2020 project, https://cordis.europa.eu/project/id/814401.
[34] NANORIGO, NANORIGO Horizon 2020 project, https://cordis.europa.eu/project/id/814530.
[35] RiskGONE, RiskGONE Horizon 2020 project, https://cordis.europa.eu/project/id/814425.
[36] NanoInformaTIX, NanoInformaTIX Horizon 2020 project, https://cordis.europa.eu/project/id/814426.
[37] Blekos, K., Marcoulaki, E. (2023). A Bayesian-based screening framework for optimal development of safe-by-design nanomaterials. In *Computer Aided Chemical Engineering* (in press), Elsevier.
[38] K. Blekos et al., Journal of Cheminformatics 15 (1) (2023) 1–17.
[39] H. Nagai, S. Toyokuni, Arch. Biochem. Biophys. 502 (1) (2010) 1–7.
[40] F.S. Bierkandt et al., Toxicology research 7 (3) (2018) 321–346.
[41] K. Donaldson et al., Particle and fibre toxicology 7 (1) (2010) 5.
[42] A.A. Shvedova et al., American Journal of Physiology-Lung Cellular and Molecular Physiology 289 (5) (2005) L698–L708.
[43] M.V. Park et al., Biomaterials 32 (36) (2011) 9810–9817.
[44] D.J. Smith et al., Appl. Phys. 51 (29) (2018).
[45] R.W. Homer et al., J. Chem. Inf. Model. 48 (12) (2008) 2294–2307.
[46] S.J. Coles et al., Org. Biomol. Chem. 3 (10) (2005) 1832–1834.
[47] I. Lynch et al., Nanomaterials 10 (12) (2020) 1–44.
[48] E. Wyrzykowska et al., Nat. Nanotechnol. 17 (2022) 924–932.
[49] M. Swirog et al., Sci. Total Environ. 840 (2022) 1–7.
[50] M. Fronzi et al., Nanomaterials 12 (21) (2022).
[51] H. Li et al., J. Phys. Chem. B 127 (15) (2023) 3596–3605.
[52] T. Puzyn et al., Nat. Nanotechnol. 6 (3) (2011) 175–178.
[53] A.A. Toropov et al., Chemosphere 89 (9) (2012) 1098–1102.

[54] A.A. Toropov et al., Saudi J. Biol. Sci. 26 (6) (2019) 1101–1106.
[55] N. Sizochenko et al., Nanoscale 6 (22) (2014) 13986–13993.
[56] N. Sizochenko et al., J. Phys. Chem. C 119 (45) (2015) 25542–25547.
[57] R. Slapikas, I. Dabo, S.B. Sinnott, Comput. Mater. Sci. 209 (2022) 111364.
[58] D.L. Liao, B.Q. Liao, J. Photochem. Photobiol. A Chem. 187 (2–3) (2007) 363–369.
[59] G. Mancardi et al., MDPI Nanomater. 12 (217) (2022) 1–18.
[60] A. Mikolajczyk et al., Chem. Mater. 27 (7) (2015) 2400–2407.
[61] S. Ortelli et al., Environ. Sci. Nano 4 (2017) 1264–1272.
[62] S. Ortelli et al., Cellulose (2018).
[63] S. Ortelli et al., J. Colloid Interface Sci. 546 (2019) 174–183.
[64] S. Ortelli et al., Colloids Surfaces B Biointerfaces 207 (2021) 112037.
[65] N. Sizochenko et al., NanoImpact 22 (December 2020) (2021) 100317.
[66] C.D. Walkey et al., ACS Nano 8 (3) (2014) 2439–2455.
[67] I. Rouse, D. Power, E.G. Brandt, M. Schneemilch, K. Kotsis, N. Quirke, A.P. Lyubartsev, V. Lobaskin, (2020). arXiv:2007.04017.
[68] S.I. Gomes et al., Nanoscale 13 (35) (2021) 14666–14678.
[69] N. Chawla, Data Mining for Imbalanced Datasets: An Overview, in: O. Maimon, L. Rokach (Eds.), Data Min. Knowl. Discov. Handb, Springer, US, 2010, pp. 875–886.
[70] A. Fernández, S. García, M. Galar, R.C. Prati, B. Krawczyk, F. Herrera, Learning from Imbalanced Data Sets, Springer Cham, 2018.
[71] G. Haixiang et al., Expert Syst. Appl. 73 (2017) 220–239.
[72] H.B. Lee, H. Lee, D. Na, S. Kim, M. Park, E. Yang, S.J. Hwang, Learning to Balance: Bayesian Meta-Learning for Imbalanced and Out-of-distribution Tasks, ICLR 2020, 2020.
[73] N.V. Chawla et al., J. Artif. Intell. Res. 16 (2002) 321–357.
[74] P. Branco, L. Torgo, R.P. Ribeiro, (2015). arXiv:1505.01658.
[75] A. Fernández et al., J. Artif. Intell. Res. 61 (2018) 863–905.
[76] D. Dablain, B. Krawczyk, N.V. Chawla, DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data, IEEE Trans. Neural Networks Learn. Syst., 2021.
[77] T. Zhao, X. Zhang, S. Wang, WSDM 2021 - Proc. 14th ACM Int. Conf. Web Search Data Min. (2021) 833–841.
[78] M. Buda, A. Maki, M.A. Mazurowski, Neural Networks 106 (2018) 249–259.
[79] J.M. Johnson, T.M. Khoshgoftaar, J. Big Data 6 (1) (2019) 1–54.
[80] A.V. Zakharov et al., J. Chem. Inf. Model. 54 (3) (2014) 705–712.
[81] J. Liu et al., Chem. Res. Toxicol. 28 (4) (2015) 738–751.
[82] K. Klimenko et al., PLoS One 14 (3) (2019) e0213848.
[83] N. Sturm et al., J. Cheminform. 12 (1) (2020) 1–13.
[84] M. Kotzabasaki et al., Nanoscale Adv. 3 (11) (2021) 3167–3176.
[85] N. Sizochenko et al., Ecotoxicol. Environ. Saf. 185 (2019) 109733.
[86] D. Fourches et al., ACS Nano 4 (10) (2010) 5703–5712.
[87] J.M. Gernand, E.A. Casman, Risk Anal. 34 (3) (2014) 583–597.
[88] D.A. Winkler et al., SAR QSAR Environ. Res. 25 (2) (2014) 161–172.
[89] T.C. Le, D.A. Winkler, Chem. Rev. 116 (10) (2016) 6107–6132.
[90] N.S. Froemming, G. Henkelman, J. Chem. Phys. 131 (23) (2009) 234103.
[91] S. Kim, K.-S. Sohn, M. Pyo, ACS Comb. Sci. 13 (2) (2011) 101–106.
[92] R. Fernandez Martinez et al., Comput. Mater. Sci. 92 (2014) 102–113.
[93] W. Wang et al., ACS Nano 11 (12) (2017) 12641–12649.
[94] V. Kovalishyn et al., Food Chem. Toxicol. 112 (2018) 507–517.
[95] R. Gómez-Bombarelli et al., ACS Cent. Sci. 4 (2) (2018) 268–276.
[96] F. Hataminia, Z. Noroozi, H. Mobaleghol Eslam, Toxicol. Vitr. 59 (2019) 197–203.
[97] J. Lazarovits et al., ACS Nano 13 (7) (2019) 8023–8034.
[98] S. Balraadjsing, W.J.G.M. Peijnenburg, M.G. Vijver, Chemosphere 307 (2022) 135930.
[99] S. Li, A.S. Barnard, Chemosphere 303 (2022) 135033.
[100] L.E. Vivanco-Benavides et al., Comput. Mater. Sci. 201 (2022) 110939.
[101] X.Z. Wang et al., Nanotoxicology 8 (5) (2014) 465–476.
[102] S.K. Jha, T.H. Yoon, Z. Pan, Comput. Biol. Med. 99 (2018) 161–172.
[103] N. Sizochenko et al., Nanoscale 10 (2) (2018) 582–591.
[104] X. Zhang, G. Ma, W. Wei, NPG Asia Mater. 13 (1) (2021).
[105] F. Spiegelman et al., Adv. Phys. X 5 (1) (2020).
[106] O. Lamiel-Garcia et al., Nanoscale 9 (3) (2017) 1049–1058.
[107] A. Cuko et al., Nanoscale 10 (45) (2018) 21518–21532.
[108] F.A. Soria, C. Di Valentin, Nanoscale 13 (7) (2021) 4151–4166.
[109] M. Kohantorabi et al., ACS Applied Materials & Interfaces 15 (6) (2023) 8770–8782.
[110] F.A. Soria, C. Daldossi, C. Di Valentin, Materials Today Energy 28 (2022) 101085.
[111] P. Siani, G. Frigerio, E. Donadoni, et al., J. Colloid Interface Sci. 627 (2022) 126–141.
[112] Y.A. Çetin et al., J. Phys.: Condens. Matter 34 (31) (2022) 314004.
[113] R. Iftimie, P. Minary, M.E. Tuckerman, Proc. Natl. Acad. Sci. U.S.A. 102 (19) (2005) 6654–6659.
[114] P.J. Ollitrault, A. Miessen, I. Tavernelli, Acc. Chem. Res. 54 (23) (2021) 4229–4238.
[115] L. Shen, W. Yang, J. Chem. Theory Comput. 14 (3) (2018) 1442–1455.
[116] S. Chmiela et al., Nat. Commun. 9 (1) (2018).
[117] M.K. Bisbo, B. Hammer, Phys. Rev. Lett. 124 (2020) 086102.
[118] X. Liu, H. Niu, A.R. Oganov, npj Computational Materials 7 (1) (2021) 199.
[119] K.H. Sørensen, M.S. Jørgensen, A. Bruix, B. Hammer, J. Chem. Phys. 148 (24) (2018) 241734.
[120] A. Singhal, G.J. Agur Sevink, Nanoscale Adv. 3 (2021) 6635–6648.
[121] B.W. Brinkmann et al., J. Chem. Inf. Model. 62 (15) (2022) 3589–3603.
[122] M. Karelson, Molecular descriptors in QSAR/QSPR, John Wiley & Sons, New York, 2000.
[123] J. Burk et al., Nanoscale 10 (46) (2018) 21985–21993.
[124] K. Tämm et al., Nanoscale 8 (36) (2016) 16243–16250.
[125] A.P. Thompson et al., Comput. Phys. Commun. 271 (2022) 108171.
[126] R. Buckingham, Proc. R. Soc. Lond. A 168 (1938) 264–283.
[127] D. Wolf et al., J. Chem. Phys. 110 (17) (1999) 8254–8282.
[128] B.B. Manshian et al., Adv. Healthc. Mater. 6 (9) (2017) 1–11.
[129] A.C. Van Duin et al., J. Phys. Chem. A 105 (41) (2001) 9396–9409.
[130] K. Chenoweth, A.C. Van Duin, W.A. Goddard, J. Phys. Chem. A 112 (5) (2008) 1040–1053.
[131] T.P. Senftle et al., npj Comput. Mater. 2 (2016) 15011.
[132] A. Hoekstra et al., Phil. Trans. R. Soc. A 377 (2018) 0180144.
[133] M. Müller, K. Katsov, M. Schick, Phys. Rep. 434 (5–6) (2006) 113–176.
[134] K.M. Langner, G.J. Sevink, Soft Matter 8 (19) (2012) 5102–5118.
[135] G.J. Sevink et al., J. Chem. Phys. 110 (4) (1999) 2250–2256.
[136] A.C. Balazs et al., J. Phys. Chem. B 104 (15) (2000) 3411–3422.
[137] S.W. Sides et al., Phys. Rev. Lett. 96 (25) (2006) 1–4.
[138] G. Milano, T. Kawakatsu, J. Chem. Phys. 130 (21) (2009).
[139] K.C. Daoulas et al., Soft Matter 2 (7) (2006) 573–583.
[140] G.J. Sevink et al., J. Chem. Phys. 153 (24) (2020).
[141] M. Schneemilch, N. Quirke, Mol. Simul. 48 (2022) 150–167.
[142] M. Deserno, W.M. Gelbart, J. Phys. Chem. B 106 (21) (2002) 5543–5552.
[143] M. Deserno, T. Bickel, Europhys. Lett. 62 (5) (2003) 767–773.
[144] M. Raatz, R. Lipowsky, T.R. Weikl, Soft Matter 10 (20) (2014) 3570–3577.
[145] E.J. Spangler, S. Upreti, M. Laradji, J. Chem. Phys. 144 (4) (2016).
[146] E.J. Spangler, M. Laradji, J. Chem. Phys. 152 (10) (2020).
[147] X. Yi, X. Shi, H. Gao, Phys. Rev. Lett. 107 (9) (2011) 1–5.
[148] X. Yi, H. Gao, Soft Matter 11 (6) (2015) 1107–1115.
[149] S. Dasgupta, T. Auth, G. Gompper, Soft Matter 9 (22) (2013) 5473–5482.
[150] S. Dasgupta, T. Auth, G. Gompper, Nano Lett. 14 (2) (2014) 687–693.
[151] R. Vácha, F.J. Martinez-Veracoechea, D. Frenkel, Nano Lett. 11 (12) (2011) 5391–5395.
[152] J. Huang, A.D. Mackerell, J. Comput. Chem. 34 (25) (2013) 2135–2145.
[153] A. Šarić, A. Cacciuto, Phys. Rev. Lett. 108 (11) (2012) 1–5.
[154] P. Angelikopoulos et al., Nanoscale 9 (3) (2017) 1040–1048.
[155] F. Simonelli et al., J. Phys. Chem. Lett. 6 (16) (2015) 3175–3179.
[156] S. Salassi et al., J. Phys. Chem. C 121 (20) (2017) 10927–10935.
[157] T. Ruiz-Herrero et al., J. Phys. Chem. B 116 (32) (2012) 9595–9603.
[158] A. Singhal, A. Sevink, Nanomaterials 12 (2022) 3859.
[159] Y. Smirnova, H. Risselada, M. Müller, PNAS 116 (2019) 2571–2576.
[160] T. Cedervall et al., Proc. Natl. Acad. Sci. U.S.A. 104 (7) (2007) 2050–2055.
[161] M. Lundqvist et al., Proc. Natl. Acad. Sci. U.S.A. 105 (38) (2008) 14265–14270.
[162] I. Lynch, K.A. Dawson, Protein-nanoparticle interactions, nanotoday 3 (2008) 40–47.
[163] C.K. Payne, J. Chem. Phys. 151 (13) (2019).
[164] T. Casalini et al., Front. Bioeng. Biotechnol. 7 (OCT) (2019) 1–14.
[165] J.E. Vance, Traffic 16 (1) (2015) 1–18.
[166] I. Capjak et al., Arh. Hig. Rada Toksikol. 68 (4) (2017) 245–253.
[167] R. García-álvarez, M. Vallet-Regí, Nanomaterials 11 (4) (2021).
[168] M. Ovais et al., Adv. Mater. 32 (22) (2020) 1–19.
[169] V. Srivastava, D. Gusain, Y.C. Sharma, Critical Review on the Toxicity of Some Widely Used Engineered Nanoparticles 54 (2015).
[170] I. Hasenkopf et al., Nano Today 46 (2022) 101561.
[171] I. Rouse et al., Phys. Chem. Chem. Phys. 23 (24) (2021) 13473–13482.
[172] J. Subbotina, V. Lobaskin, J. Phys. Chem. B 126 (6) (2022) 1301–1314.
[173] D. Power et al., Model. Simul. Mater. Sci. Eng. 27 (8) (2019) 84003.
[174] S.A. Alsharif et al., Nanomaterials 10 (10) (2020) 1–21.
[175] N.-V. Buchete, J.E. Straub, D. Thirumalai, Curr. Opin. Struct. Biol. 14 (2004) 225–232.
[176] S.J. Marrink et al., J. Phys. Chem. B 111 (27) (2007) 7812–7824.

[177] V. Karunakaran Annapoorani, et al., In preparation (2023).
[178] D. Schneidman-Duhovny et al., Nucleic Acids Res. 33 (SUPPL. 2) (2005) 363–367.
[179] S. Chibber, I. Ahmed, Biochem. Biophys. Reports 6 (2016) 63–67.
[180] B.G. Pierce et al., Bioinformatics 30 (12) (2014) 1771–1773.
[181] M.S. Ali, M. Altaf, H.A. Al-Lohedan, J. Photochem. Photobiol. B Biol. 173 (May) (2017) 108–119.
[182] T. Lima et al., Sci. Rep. 10 (1) (2020) 1–9.
[183] A. Pancaro et al., Nanoscale 13 (24) (2021) 10837–10848.
[184] A. Kurtz-Chalot et al., Mater. Sci. Eng. C 75 (2017) 16–24.
[185] J.F. Hainfeld et al., J. Pharm. Pharmacol. 60 (2008) 977–985.
[186] O. Zeiri, ACS Sensors 5 (12) (2020) 3806–3820.
[187] A. Heuer-Jungemann et al., Chem. Rev. 119 (8) (2019) 4819–4880.
[188] C. Singh et al., Phys. Rev. Lett. 99 (22) (2007) 1–4.
[189] M. Baranov, E. Nepomyashchaya, E. Velichko, Computer simulation of biomolecules around metallic nanoparticle for biomolecular electronics, in: Proc. 2021 Int. Conf. Electr. Eng. Photonics, EExPolytech 2021, 2021, pp. 171–174.
[190] R. Cappabianca et al., ACS Omega 7 (2022) 42292–42303.
[191] A.C.T. van Duin et al., Catalysis 14 (2014) 223–243.
[192] O.T. Unke et al., Chem. Rev. 121 (16) (2021) 10142–10186.
[193] G. Trezza et al., npj Computational Materials 8 (1) (2022) 1–14.
[194] E. Chiavazzo et al., Proc. Nat. Acad. Sci. 114 (28) (2017) E5494–E5503.
[195] O.T. Unke et al., Chem. Rev. 121 (16) (2021) 10142–10186.
[196] S. Bag et al., J. Chem. Theory Comput. 17 (11) (2021) 7195–7202.
[197] T. Seki, K. Takeda, A. Nakayama, J. Phys. Chem. C 126 (7) (2022) 3404–3410.
[198] S. Wei et al., J. Mol. Liq. 319 (2020) 114135.
[199] K. Leung, J.A. Greathouse, Commun. Chem. 5 (1) (2022) 1–8.
[200] D. Chen, C. Shang, Z.-P. Liu, J. Chem. Phys. 156 (9) (2022) 094104.
[201] I. Tomé, V. Francisco, H. Fernandes, et al., APL Bioeng. 5 (3) (2021) 1–12.
[202] J.E. Dahlman et al., Proc. Natl. Acad. Sci. U.S.A. 114 (8) (2017) 2060–2065.
[203] X. Mao et al., npj Comput. Mater. 7 (1) (2021) 1–9.
[204] N. Jeliazkova et al., Beilstein J. Nanotechnol. 6 (1) (2015) 1609–1634.
[205] N. Jeliazkova et al., Nat. Nanotechnol. 16 (6) (2021) 644–654.
[206] N. Jeliazkova, V. Jeliazkov, J. Cheminform. 3 (1) (2011) 1–18.
[207] N. Jeliazkova, Expert Opin. Drug Metab. Toxicol. 8 (7) (2012) 791–801.
[208] N. Kochev et al., Nanomaterials 10 (10) (2020) 1–23.
[209] OECD, Guidance on Grouping of Chemicals, Second Edition, 2017.
[210] ECHA, Guidance on information requirements and chemical safety assessment Chapter R.6: QSARs and grouping of chemicals, 2008.
[211] A. Mech et al., Nanotoxicology 13 (1) (2019) 119–141.
[212] S. Halappanavar et al., Part. Fibre Toxicol. 17 (1) (2020) 1–24.
[213] G.T. Ankley et al., Environ. Toxicol. Chem. 29 (3) (2010) 730–741.
[214] A. Giusti et al., NanoImpact 16 (2019) 100182.
[215] L. Lamon et al., Nanotoxicology 13 (1) (2019) 100–118.
[216] A.G. Oomen et al., NanoImpact 9 (2018) 1–13.
[217] A. Bahl et al., NanoImpact 15 (2019) 100179.
[218] A. Bahl et al., NanoImpact 19 (2020) 100234.
[219] L. Lamon et al., Part. Fibre Toxicol. 15 (1) (2018) 1–17.
[220] I. Karkossa et al., Part. Fibre Toxicol. 16 (1) (2019) 1–19.
[221] I. Karkossa et al., Sci. Total Environ. 801 (2021) 149538.
[222] A. Bannuscher et al., Nanotoxicology 14 (2) (2020) 181–195.
[223] A. Bannuscher et al., Nanotoxicology 14 (6) (2020) 807–826.
[224] L.A. Saarimäki et al., Sci. Data 8 (1) (2021) 1–10.
[225] V. Fortino et al., Nat. Commun. 13 (1) (2022).
[226] OECD, Transcriptomic Reporting Framework (TRF), 2021.
[227] OECD, Metabolomic Reporting Framework (MRF), 2021.
[228] A. Bahl, C. Ibrahim, K. Plate, A. Haase, J. Dengjel, P. Nymark, V.I. Dumit, Proteomas: a workflow enabling harmonized proteomic meta-analysis and proteomic signature mapping, Journal of Cheminformatics 15 (1) (2023) 1–17.