NanoImpact 25 (2022) 100366

Contents lists available at ScienceDirect

NanoImpact

journal homepage: www.elsevier.com/locate/nanoimpact

How can we justify grouping of nanoforms for hazard assessment? Concepts and tools to quantify similarity

Nina Jeliazkova [a], Eric Bleeker [b], Richard Cross [c], Andrea Haase [d], Gemma Janer [e], Willie Peijnenburg [b,f], Mario Pink [d], Hubert Rauscher [g], Claus Svendsen [c], Georgia Tsiliki [h], Alex Zabeo [i], Danail Hristozov [i], Vicki Stone [j], Wendel Wohlleben [k,*]

[a] Ideaconsult Ltd, Sofia, Bulgaria
[b] National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands
[c] UKRI Centre for Ecology and Hydrology, MacLean Building, Benson Lane, Wallingford OX10 8BB, UK
[d] German Federal Institute for Risk Assessment (BfR), Department of Chemical and Product Safety, Berlin, Germany
[e] LEITAT Technological Center, Barcelona, Spain
[f] Institute of Environmental Sciences (CML), Leiden University, Leiden, the Netherlands
[g] European Commission, Joint Research Centre (JRC), Ispra, Italy
[h] Athena-Research and Innovation Center in Information, Communication and Knowledge Technologies, Marousi, Greece
[i] GreenDecision, Venice, Italy
[j] NanoSafety Research Group, Heriot-Watt University, Riccarton, Edinburgh EH14 4AS, UK
[k] BASF SE, Dept. Material Physics and Dept Experimental Toxicology & Ecology, Ludwigshafen, Germany

ARTICLE INFO

Editor: Bernd Nowack

Keywords:
Nanomaterials
Nanoform
Similarity
Clustering
Integrated approach to testing and assessment
Regulation
Nanomaterial grouping
Read-across
Nanoinformatics
Safe(r) by design

ABSTRACT

The risk of each nanoform (NF) of the same substance cannot be assumed to be the same, as they may vary in their physicochemical characteristics, exposure and hazard. However, neither can we justify a need for more animal testing and resources to test every NF individually. To reduce the need to test all NFs, (regulatory) information requirements may be fulfilled by grouping approaches. For such grouping to be acceptable, it is important to demonstrate similarities in physicochemical properties, toxicokinetic behaviour, and (eco)toxicological behaviour.

The GRACIOUS Framework supports the grouping of NFs, by identifying suitable grouping hypotheses that describe the key similarities between different NFs. The Framework then supports the user to gather the evidence required to test these hypotheses and to subsequently assess the similarity of the NFs within the proposed group.

The evidence needed to support a hypothesis is gathered by an Integrated Approach to Testing and Assessment (IATA), designed as decision trees constructed of decision nodes. Each decision node asks the questions and provides the methods needed to obtain the most relevant information. This White paper outlines existing and novel methods to assess similarity of the data generated for each decision node, either via a pairwise analysis conducted property-by-property, or by assessing multiple decision nodes simultaneously via a multidimensional analysis.

For the pairwise comparison conducted property-by-property we included in this White paper:

- A Bayesian model assessment which compares two sets of values using nested sampling. This approach is new in NF grouping.
- A Arsinh-Ordered Weighted Average model (Arsinh-OWA) which applies the arsinh transformation to the distance between two NFs, and then rescales the result to the arsinh of a biologically relevant threshold before grouping using OWA based distance. This approach is new in NF grouping.
- An x-fold comparison as used in the ECETOC NanoApp.
- Euclidean distance, which is a highly established distance metric.

The x-fold, Bayesian and Arsinh-OWA distance algorithms performed comparably in the scoring of similarity between NF pairs. The Euclidean distance was also useful, but only with proper data transformation. The x-fold

* Corresponding author.
E-mail address: wendel.wohlleben@basf.com (W. Wohlleben).

https://doi.org/10.1016/j.impact.2021.100366
Received 9 August 2021; Received in revised form 15 October 2021; Accepted 12 November 2021
Available online 20 November 2021
2452-0748/© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

method does not standardize data, and thus produces skewed histograms, but has the advantage that it can be implemented without programming knowhow.

A range of multidimensional evaluations, using for example dendrogram clustering approaches, were also investigated. Multidimensional distance metrics were demonstrated to be difficult to use in a regulatory context, but from a scientific perspective were found to offer unexpected insights into the overall similarity of very different materials.

In conclusion, for regulatory purposes, a property-by-property evaluation of the data matrix is recommended to substantiate grouping, while the multidimensional approaches are considered to be tools of discovery rather than regulatory methods.

1. Introduction

1.1. Introduction: motivation on similarity of nanoforms

It is increasingly difficult to test, on a case-by-case basis, the large number of nanoforms (NFs) (NF, Box 1) which potentially can exist for a single substance. For example, silica is commercially available from at least three very different production processes, in a variety of specific surface areas and many different types and extents of surface treatments, while carbon nanotubes are available with a variety of wall numbers, of lengths, of catalysts used for their production and many different types and extents of functionalisation. Instead of (animal) testing of each NF, alternative approaches including grouping and read-across (Box 1) are needed to fill data gaps on hazard information.

Within the EU Regulation on Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), grouping and read-across are among the most commonly used alternative approaches, and are also considered in cosmetics and biocidal product EU regulations (Giusti et al., 2019). In fact, REACH foresees two different levels of grouping for NFs; grouping to generate sets of similar NFs (ECHA, 2019b; Janer et al., 2020), and grouping for the purpose of read-across (ECHA, 2019a). In the case of the sets of similar NFs, similarity is required to justify that

Box 1
Terminology

This box provides explanations for the most important generic terms used in the GRACIOUS Framework and in this article. For harmonized GRACIOUS terminology related to specific physicochemical, environmental, eco-toxicological, exposure-related and human health endpoints/properties please refer to the GRACIOUS wiki https://terminology-harmonizer.greendecision.eu/Gracious

Term	Explanation
Nanomaterial	A natural, incidental or manufactured material containing particles, in ither an unbound state, as an aggregate or as an agglomerate and where, for 50% or more of the particles in the number size distribution, one or more external dimensions is in the size range 1 nm - 100 nm. […]. By derogation from the above, fullerenes, graphene flakes and single wall carbon nanotubes with one or more external dimensions below 1 nm should be considered as nanomaterials. For this purpose, “particle” means a minute piece of matter with defined physical boundaries; “agglomerate” means a collection of weakly bound particles or aggregates where the resulting external surface area is similar to the sum of the surface areas of the individual components and “aggregate” means a particle comprising of strongly bound or fused particles.
Nanoform (Commission, E, 2018)	On the basis of the Commission Recommendation of 18 October 2011 on the definition of nanomaterial, a NF is a form of a natural or manufactured substance containing particles, in an unbound state or as an aggregate or as an agglomerate and where, for 50% or more of the particles in the number size distribution, one or more external dimensions is in the size range 1 nm–100 nm, including also by derogation fullerenes, graphene flakes and single wall carbon nanotubes with one or more external dimensions below 1 nm. […] A NF shall be characterised in accordance with REACH Annex VI, section 2.4. A substance may have one or more NFs, based e.g. on differences in their number based particle size distribution, shape, aspect ratio, crystallinity, assembly structure, specific surface area and surface functionalisation or treatment (REACH Annex VI, points 2.4.2. – 2.4.5).
Set of similar NFs (Commission, E, 2018)	A group of NFs characterised in accordance with section 2.4 of REACH Annex VI where the clearly defined boundaries in the parameters in the points 2.4.2 to 2.4.5 (of Annex VI) of the individual NFs within the set still allow to conclude that the hazard assessment, exposure assessment and risk assessment of these NFs can be performed jointly. A justification shall be provided to demonstrate that a variation within these boundaries does not affect the hazard assessment, exposure assessment and risk assessment of the similar NFs in the set. A NF can only belong to one set of similar NFs.
Grouping	The OECD defines grouping as the general approach for assessing more than one chemical at the same time (OECD, 2014). According to OECD (2014), the rationale underpinning grouping of substances may be based on similarity due to: • Common functional group(s); • Common constituents or chemical classes, similar carbon range numbers; • A common mode or mechanism of action or adverse outcome pathway; • The likelihood of common precursors and/or breakdown products via physical or biological processes that result in structurally similar chemicals; • An incremental and constant change across the category. The EU builds on this OECD approach. Annex XI to REACH (Commission, E, 2018) addresses grouping and read-across between different substances and establishes that structural similarity is a prerequisite for any

(*continued*)

Term	Explanation
	grouping and read-across approach. However, for grouping different NFs [or sets of NFs] of the same substance the molecular structural similarities alone cannot serve as a justification (ECHA, 2019a). Where technically and scientifically justified, grouping and read-across can be applied within a registration dossier to two or more NFs for the purposes of one or more information requirements. However, for grouping different NFs of the same substance, consideration of the molecular structural similarities alone is not sufficient to serve as a justification (REACH Annex XI) (Commission, E, 2018).
Grouping hypothesis	A description of the similarities in key properties which can be linked to a certain hazard endpoint and route of exposure (or environmental compartment), and which define the NF(s) of concern as member(s) of a group (Stone et al., 2020). The concept of similarity underpinning grouping and read-across for NFs of the same substance possibly includes physicochemical information on "what they are", "where they go" and "what they do" (European and C. The, 2006). ECHA identified two basic grouping hypotheses: (ECHA, 2019a) • grouping of chemicals that have the same type of effect(s), and • grouping of chemicals that (bio)transform to common compound(s). In principle these hypotheses are considered applicable to NFs (Worth et al., 2017; Lamon et al., 2019).
Read-across	The OECD defines read-across as a technique to fill in data gaps, where the test information concerning a certain endpoint for one chemical, referred to as a source chemical, is used to predict the test information concerning the same endpoint for another chemical, referred to as a target chemical, which is considered to be similar based on a scientific justification (OECD, 2014). Annex XI to REACH (Commission, E, 2018) addresses grouping and read-across between different substances and establishes that structural similarity is a prerequisite for any grouping and read-across approach aimed to fulfil the standard information requirements. Apart from a structural similarity basis, ECHA requires a read-across hypothesis to be provided that establishes why a prediction is possible (ECHA, 2017). Annex XI to REACH was recently revised to include specific provisions for NFs and extend the applicability of the concept of grouping and read-across to NFs of the same substance. Accordingly, read-across is a technique for predicting endpoint specific information for one NF or set of NFs (designated as target), by using data on the same endpoint from another form of the substance (i.e. NFs, sets of NFs or non-NFs) (designated as source). ECHA released guidance on how to apply grouping and read-across to NFs of the same substance (ECHA, 2019a).
Category approach (ECHA, 2017; European and C. The, 2006)	The term category approach is used when read-across is employed between several substances that have structural similarity or other similarity characteristic. These substances are grouped together on the basis of defined similarity and differences between the substances. The approach provides a basis on which to identify possible trends in properties across the category. As the number of possible chemicals being grouped into a category increases, the potential for developing hypotheses for specific endpoints and making generalisations about the trends within the category will also increase.
Analogue approach (ECHA, 2017; European and C. The, 2006)	The term analogue approach is used when read-across is employed between a small number of structurally (or otherwise) similar substances; there is no trend or regular pattern in the properties. As a result of the similarity, a given (eco)toxicological or fate property of one substance (the source) is used to predict the same property for another substance (the target). The simplest case of an analogue approach is read-across from a single source substance to a target substance.
Endpoint	Hazard or toxicological endpoints are values derived from toxicity tests. They are the results of specific measurements made during or at the conclusion of the test. Examples for such endpoints are acute toxicity (oral toxicity: LD50, short-term aquatic toxicity: LC50) and repeated dose toxicity (LOAEL). Endpoints are the recorded observation coming from an *in chemico* method, an *in vitro* assay or an *in vivo* assay.
Property (intrinsic and extrinsic)	A property of a NF can be a basic physicochemical parameter (e.g. size, mass) required to identify a NF, or it can describe an aspect of the NF interaction with the immediate surroundings (e.g. reactivity, attachment efficiency). In the latter case the property depends on both the NF and its surroundings (extrinsic property), whereas in the former case the property is independent of the surroundings (intrinsic property).
(Scalar) descriptor	A single number, accompanied by units of measurement (e.g. nm). A scalar descriptor is the result of a reduction of a two-dimensional distribution of data points that characterises the data field in a way which is assumed sufficient for a specific purpose. Examples of scalar descriptors are D50 (median) for particle size distribution or LOAEL (Lowest-observed-adverse-effect level) for the dose response curve in inhalation toxicity.
Data requirement	Information needed to determine whether a specific grouping hypothesis is applicable.
Substance	A chemical element and its compounds in the natural state or obtained by any manufacturing process, including any additive necessary to preserve its stability and any impurity deriving from the process used, but excluding any solvent which may be separated without affecting the stability of the substance or changing its composition (European and C. The, 2006).
Application range	One can assess the similarity of NFs in one property only within the application range, which is given by the overlap of the biologically relevant range and the measurable range of that property.
Applicability domain	The applicability domain of a grouping hypothesis describes the ranges of values of an endpoint within which reliable estimations for an endpoint can be made for the members of the group (ECHA, 2017).
Data matrix (matrix of data availability)	A matrix consisting of the group members/group candidates vs. the corresponding set of available data for all relevant physicochemical, toxicological and ecotoxicological properties/endpoints for a specific IATA. The data matrix is the evidence base used to formulate or decide a grouping or read-across decision. A data matrix contains all and only the evidence required by the IATA that applies to a specific hazard. Missing values are indicated by 'NA'. The matrix therefore helps highlighting the data gaps. The data matrix provides the data needed to evaluate similarity between NFs for each hazard endpoint.

(*continued*)

Term	Explanation
Integrated Approach to Testing and Assessment (IATA)	An IATA is an approach based on multiple information sources used for the hazard identification, hazard characterization and/or safety assessment of chemicals. An IATA integrates and weights all relevant existing evidence and guides the targeted generation of new data, where required, to inform regulatory decision-making regarding potential hazard and/or risk. Within an IATA, data from various information sources are evaluated and integrated to draw conclusions on the hazard and/or risk of chemicals. Within this process, the incorporation of data generated with non-animal testing and non-testing methods is expected to contribute considerably to a reduction of testing in animals. In general, the output of an IATA is a conclusion that, along with other considerations, informs regulatory decision making (OECD, 2017). The IATAs generated by GRACIOUS guide the user through acquisition of the documentation (data matrix and information) required to accept or reject a specific grouping and read-across hypothesis.
Safe(r)-by-design (SbD)	The SbD concept for nanomaterials was initially formulated in NANoREG. (Gottardo et al., 2017) According to OECD, the SbD (Safe-by-Design, Safer-by-Design, or Safety-by-Design) concept (OECD, 2020) refers to identifying the risks and uncertainties concerning humans and the environment at an early phase of the innovation process so as to minimize uncertainties, potential hazard(s) and/or exposure. The SbD approach addresses the safety of the material/product and associated processes through the whole life cycle: from the Research and Development (R&D) phase to production, use, recycling and disposal. For SbD in nanotechnology, three pillars of design can be specified: I. Safe(r) material/product: minimizing, in the R&D phase, possible hazardous properties of the nanomaterial or nano-enabled product while maintaining function; II. Safe(r) production: ensuring industrial safety during the production of nanomaterials and nano-enabled products, more specifically occupational, environmental and process safety aspects; and III. Safe(r) use and end-of-life: minimizing exposure and associated adverse effects through the entire use life, recycling and disposal of the nanomaterial or nano-enabled product. This can also support circular economy. Grouping hypotheses can be formulated based on similarity of hazard (I), similarity of exposure rates (II), similarity of forms of release (III).
Distance or metric	A function that defines how far apart two data points are. Typical examples are the Euclidean, Manhattan and Minkowski distances. A distance is a metric if it is nonnegative and symmetric, while the identity principle and the triangle inequality holds. The latter means that distance between points A and B is less or equal to the sum of the distances between points A and C and between points B and C.
Data standardisation	In statistics, "standardised" means that a data scaling transformation is applied per property to have variance 1 and mean 0.
Supervised and unsupervised machine learning methods	Machine learning algorithms generally can be divided into unsupervised or supervised. Regression and classification are supervised algorithms (because the training / fitting is supervised by the Y values). Clustering is unsupervised – clusters are identified solely by X data, without taking into account any Y data. Examples are hierarchical and non-hierarchical unsupervised algorithms such as Hierarchical Cluster Analysis (HCA), k-means algorithm, Density-Based Spatial Clustering (DBSCAN), spectral clustering. Dimensionality can also be reduced by other methods (e.g. Principal Component Analysis (PCA)),
Benchmark materials and Representative Test Materials (RTM)	All materials used in GRACIOUS as benchmark or reference materials are representative test materials (RTM) in the metrological sense (Roebben et al., 2013). They serve as a point of reference to support the interpretation and assessment of results obtained on a new test material. A representative test material is a material from a single batch, which is sufficiently homogeneous and stable with respect to one or more specified properties, and which implicitly is assumed to be fit for its intended use in the development of test methods which target properties other than the properties for which homogeneity and stability have been demonstrated. RTMs used in GRACIOUS are well-characterised nanomaterials, e.g. from the JRC repository. For some assays, they also serve as positive and negative controls, but controls could also be non-particulate chemicals.
Fingerprint/ fingerprinting	A unique set of descriptors indicating the presence of particular functionalities in or on a NF, as based on specialized analytic techniques.

hazard, exposure and risk assessment of the NFs within the set can be performed jointly (ECHA, 2019a). For a set of similar NFs, the justification of such similarity should apply to all hazard endpoints. If this is not possible, so that different hypotheses are required for different endpoints, a set of similar NFs cannot be created. In such a case, grouping for the purpose of read-across might be explored. This White paper, and the publications in the associated issue of NanoImpact will focus on similarity assessment methods to support grouping and read-across. Such similarity assessment should include an assessment on structural similarity together with a hypothesis to establish why a prediction of hazard is possible. To justify the grouping hypothesis the similarity assessment may need to go beyond structural similarity alone. This is further elaborated in the next paragraphs.

The general concepts on grouping of chemicals are also applicable to NFs. For 'conventional' chemicals, structural similarity is the key element in establishing chemical categories or finding analogues, but this is not sufficient when dealing with NFs. To address this, ECHA have generated guidance on grouping and read-across between NFs, or between NFs and non-NFs of the same substance (ECHA, 2019a). The ECHA guidance clarifies the need to consider similarities of not just

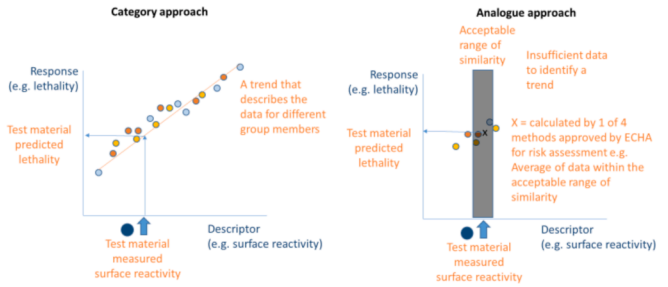

Fig. 1. Category approach and analogue approach. The different colour spots represent data for different nanoforms or substances. When combining the data from the different substances a clear trend can be observed in the left graph allowing a category approach to grouping to be applied. For the category approach the applicability range can be predicted by the pattern or trend in the groups data. In the right graph there is insufficient data available to identify a trend in the data, and therefore an analogue approach has to be applied by comparing the target NF(s) to one or more source NFs that are similar with respect to the parameter mewasured. The applicability range for an analogue approach is restricted to the range of values for which data exists for the group members.

physicochemical properties, but also toxicokinetic behaviour and fate, and (eco)toxicological behaviour between different NFs. The guidance indicates that it is possible to use physicochemical parameters and/or in vitro screening methods to develop a robust scientific explanation of why different forms of the substance are sufficiently similar to be grouped when considering their hazard (ECHA, 2019a). Although this guidance addresses only read-across for different forms of the same substance, it does not preclude read-across between NFs of different substances.

For each hazard endpoint, the explanation of why different NFs can be grouped needs to be written as a hypothesis, including identification of the key properties that can be mechanistically linked to the endpoint. Such properties should then be the focus of the similarity assessment. The arguments such as common breakdown products and a common mode or mechanism of action (as suggested by OECD 2014) could be used to build hypotheses. When accompanied by data demonstrating a sufficient level of similarity, these hypotheses would support grouping and read-across.

The EU H2020 funded project GRACIOUS has generated a Framework to support grouping and read-across (Stone et al., 2020), which includes 40 predefined hypotheses as well as a template for the user to design their own hypotheses. Each hypothesis is accompanied by a tailored Integrated Approach to Testing and Assessment (IATA), which consists of a series of questions designed to identify the information needed to test the hypothesis. This information includes a matrix of properties such as dissolution rate, surface reactivity, or ability to induce inflammation, which when combined allow the user to accept or reject the grouping hypothesis. The IATA specifies the most relevant tests required to provide the evidence for each property, and supports the generation of a data matrix (as recommended by ECHA) that can be used to evaluate similarity between NFs (and non-NFs) for each hazard endpoint.

Besides the support of regulatory applications of grouping, the GRACIOUS Framework also supports other purposes, including development of precautionary measures and safe(r) by design (SbD) during product innovation. A similarity assessment of NFs can also be of help when considering changes in properties through the NF life cycle, including when released into the environment. It is important to highlight that similarity between two or more NFs is context-dependent (e.g. in an occupational setting, versus in an aquatic environment) and hazard endpoint specific (e.g. sensitization versus mutagenicity). The underlying hypothesis and key determinants of similarity may however be common to several hazard endpoints (e.g. surface reactivity could influence endpoints such as inflammation, fibrosis and genotoxicity).

The criteria needed to assess the level of similarity required for each property will depend on the purpose of the assessment. For example, a lower similarity may be acceptable in a safe-by-design context than in a regulatory context. The level of required similarity will depend on whether the source material is an example of a worst-case (most hazardous) NF or non-NF. In the case of a category approach (see details on the category versus analogue approach (ECHA, 2017)), it is assumed that a trend in one or more physicochemical properties is associated with a trend in (eco)toxicological and/or environmental fate properties (Fig. 1). Therefore, read-across in the category approach context is supported by data that demonstrates such trends.

In contrast to the variety of approaches to estimate chemical structural similarity available for well-defined chemical structures (Willett et al., 1998; Kochev et al., 2003), there are, at the moment, no widely accepted quantitative methods to describe similarity between NFs. In fact, this is an issue that has received very little attention so far, partly because there were no widely accepted relevant descriptors of nanomaterials. The GUIDEnano project developed an approach to quantify similarity between exposure relevant NFs and tested NFs (Park et al., 2018). More recently, an approach was developed in the context of the formation of sets of similar NFs (Janer et al., 2020).

In this White paper, we present several types of algorithms to quantify similarity, along with different types of visual representations to facilitate decision making for simultaneous assessment of similarity for multiple properties. Since data sets are often complex (e.g. dose response curves for multiple NFs), the need for data reduction and selection of appropriate descriptors are addressed. We also describe different considerations and approaches that can be followed to establish acceptable similarity thresholds for a given parameter for a given endpoint. The methods described in this White paper provide the tools to allow quantitative rather than qualitative read-across to fill data gaps. Finally, we provide recommendations of which of the identified methods are appropriate for use in a regulatory or safe(r)-by-design context.

1.2. *Introduction: similarity assessment for substances*

Finding similar chemical structures is essential for pursuits such as drug design, with the first reports dating back to the mid-1980s. Screening large virtual libraries of molecular structures required the development of computational methods, allowing the user to go beyond an expert-based similarity assessment. Chemical structure has been traditionally represented using a variety of topological, physicochemical and electronic descriptors. This has provided the grounds for evaluating similarity between compounds by comparing numerical values of these descriptors. The seminal review by Peter Willett (Willett et al., 1998) discussed structure representations (descriptors calculated from chemical structure, e.g. chemical fingerprints), distance metrics (e.g. Euclidean or cosine distance) and similarity coefficients (e.g. Tanimoto index), as well as application of similarity searching like clustering of chemical structures and database screening. Chemical similarity assessment is now part of cheminformatics curricula and textbooks (Kochev et al., 2003; Bajorath, 2017), and there is a continuous stream of new publications, reporting new methods and/or applications. Different structure representations and different distances or similarity indices fit different purposes, for example different methods find different subsets of active compounds that lead to the same biological activity (Sheridan and Kearsley, 2002). It is important to realize that the "similarity principle", i.e. the expectation that similar (high-ranked) compounds are likely to have similar properties to the query compound is an assumption that has exceptions (Guney, 2017). In other words, proximity with respect to descriptors does not necessarily mean proximity with respect to activity, and does not represent a causal relationship per se. There is a decade of large and active research on "activity cliffs", developing methods to identify exactly these exceptions to the

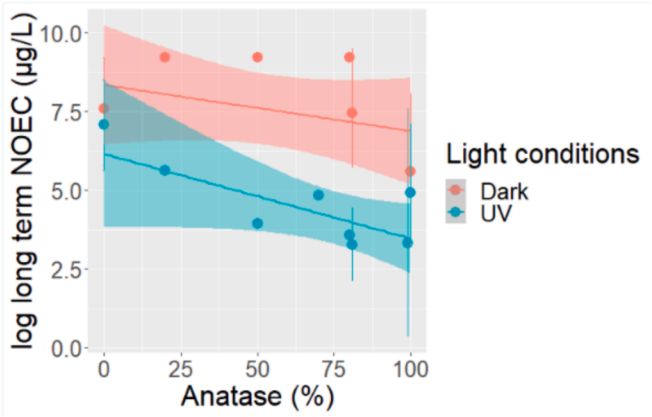

Fig. 2. The relationship between % anatase in a TiO_2 NF particle population, and log long term NOEC in Daphnia species under exposure to UV light or in the dark. Error bars represent the standard deviation where multiple data points were available for a NF. See the Supporting Information for a comprehensive summary detailing the data and statistical approach.

similarity principle.

Experiences in applying similarity methods for the grouping and read across of chemical substances have been reviewed by Patlewicz (Patlewicz et al., 2017). More recently, different similarity methods have been recommended for read across (Mellor et al., 2019; Floris and Olla, 2018). In addition such methods have been applied to develop efficient screening methods to prioritize for testing chemicals with a high potential of being hazardous (Wassenaar et al., 2021).

The similarity assessment workflow comprises of a 4-step approach:

1. Construction of datasets of substances for which data for a specific endpoint are available.
2. Physicochemical characterization of all substances in the datasets. Various options are available, ranging from experimentally determined to calculated properties covering a range of complexities. For example, the suite of descriptors potentially available ranges from extremely simple approaches such as counting the number of carbon atoms in a molecule, to generating binary fingerprints of molecules based on the presence/absence of specific functional groups, and

Box 2
Assessment of the similarity of different NFs based on property-response relationships: an environmental example for TiO_2 NFs varying in crystallinity

1.3. Establishing the mechanism of toxicity and the role of an intrinsic or extrinsic property

In the case of TiO_2 NFs, maximum photocatalytic activity is observed for a mixture of anatase and rutile, rather than a pure single phase of the material (Bacsa and Kiwi, 1998; Khataee et al., 2009). Therefore the ratio of anatase to rutile crystalline phases in the particle population is used to enhance the photocatalytic properties of the material in nano-enabled products such as self-cleaning glass. However, the photocatalytic activity of TiO_2 NFs is also suggested as a major driver behind its ecotoxicity (Clément et al., 2013). Photocatalysis and the generation of exogenous or endogenous reactive oxygen species may result in oxidative damage to cells, thus leading to toxicity (Dasari et al., 2013). Photocatalytic reactivity may therefore be considered a relatively specific extrinsic property for similarity assessment of TiO_2 NFs.

1.4. Identification of a regular pattern between a property and a specific endpoint

Whilst the literature concerning the ecotoxicity of TiO_2 NFs is rich (being one of the most frequently tested nanomaterials), it is multifaceted in nature, with toxicity data published for studies relating to assays with a range of species, test systems, endpoints and dose response descriptors. Normalization approaches have been employed to allow interrogation of this data, through extrapolation factors that convert short-term acute to chronic long-term values, and from other effect metrics to long term no observed effect concentrations (long term NOEC) (Sørensen et al., 2020).

Daphnia species are the preferred test species for short term toxicity testing of invertebrates (REACH Annex VII 9.1.1). For this reason, data for *Daphnia* species exposed to TiO_2 NFs were extracted from a species sensitivity database by Sørensen (Sørensen et al., 2020). A regular pattern is observed, in which increasing the proportion of anatase resulted in a decrease in *Daphnia* sp. long term NOEC, under exposure to ultraviolet light (Fig. 2). The lowest long term NOECs were observed for the ~80:20 mix of anatase/rutile TiO_2, a mixture designed to have the greatest theoretical photoreactivity. It is of note that the relationship between crystalline form and toxicity is less prominent, or even negligible, under dark conditions. Both lines of evidence support the hypothesis that photoreactivity is a significant driver of the toxicity of different NFs of TiO_2. TiO_2 NFs for which photoreactivity is the principle mechanism of toxicity would therefore be considered grouped for short term toxicity testing on invertebrates. Since a trend was detected the relationship between intrinsic (% anatase) or extrinsic properties (the photoreactivity of NFs) and their long term NOEC, a category approach to read-across for this endpoint can be applied to new NFs of TiO_2.

1.5. For endpoints that do not change over the range of the property, that property is of low environmental or biological relevance for the endpoint

The assembled species sensitivity database for TiO_2 (Sørensen et al., 2020) under dark conditions shows no systematic trend in toxicity for the different crystalline forms and hence the relationship forms a horizontal line (Fig. 3), as opposed to the systematic trend in the idealised Fig. 1. The absence of a relationship between crystallinity and NOEC in the dark is an example of where the property is of small influence (if any) on the endpoint. Endpoints for exposures under dark conditions (sediment and soil toxicity) could therefore consider all NF variants of TiO_2, with different ratios of anatase and rutile crystalline forms, as similar. On this basis, IATAs in the soil compartment would not include photoreactivity as decision node. Differences in crystalline form of TiO_2 NFs would not preclude grouping and read-across approaches between NFs for the NOEC endpoint.

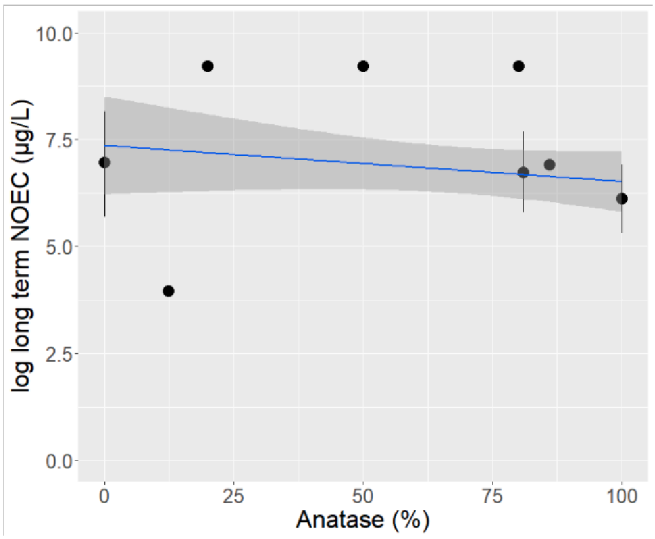

Fig. 3. The lack of contribution of crystallinity on toxicity (long term NOEC) under dark conditions (60 data points across 14 species). Where multiple data points are available for a crystalline form, mean and standard error are presented.

even to descriptors based on advanced quantum chemical calculations (Kochev et al., 2003; Bajorath, 2017).

3. Assessing the extent of chemical similarity between substances. Dependent on the descriptors available, a suite of methods can be used for this purpose. These methods range from comparison of fingerprints by means of similarity coefficients, to multivariate statistical analysis, or even more complex approaches such as quantum molecular similarity or pathway similarity (Willett et al., 1998).
4. Determining an endpoint-specific optimal similarity threshold (i.e. how similar do the chemical structures need to be for a particular purpose) and the predictive performance of either each descriptor generated or of each set of molecular descriptors available.

It is to be noted that the endpoint of assessment does not necessarily need to be a continuous scalar variable (e.g. size), as also binary endpoints (e.g. absence/presence of an amino-group, or yes/no induction of DNA-damage) can be used for similarity assessment.

2. Assessing the similarity of NFs based on individual NF properties

2.1. Dynamic range and relevant ranges of properties: how similar do NFs need to be for grouping?

2.1.1. What are the dynamic and relevant ranges of a parameter, and how do they limit the application range of an analogue approach when conducting a similarity assessment?

Many properties of NFs can be measured (e.g. size, shape, length, composition). For some of these properties, the measured values extend over a range (e.g. size, length). However, this full range might not be relevant to measure, either due to lack of biological relevance (e.g. particles exceeding a certain size cannot be inhaled), or due to a lack of method accuracy (Fig. 4). The application range describes the range of values within which the property is biologically relevant and can be measured reliably for the members of the group (see box on terminology) (ECHA, 2017). The grouping hypothesis may provide some context for this range (e.g. dissolution rate above or below a specific threshold). To assess the grouping hypothesis, an assessment of similarity must allow the user to distinguish between property values that fall within and outside the application range.

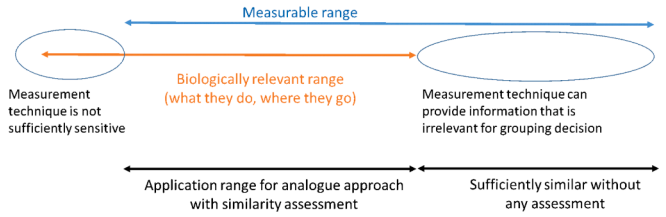

Fig. 4. Limitations to the application range. The range of descriptor values that is measurable and the biologically (or environmentally) relevant range may not map onto each other perfectly. Understanding of this relationship helps to inform data ranges suitable for similarity assessment of "where they go" and "what they do" for both analogue approaches and for category approaches. Similarity assessment is possible and required only for an analogue approach within the overlap of measurable and biologically (environmentally) relevant ranges – this overlap defines the application range. Ideally, representative test materials (RTMs) for the upper and lower limits are included in the measurement of a candidate NF group (e.g. Fig. 7, Table 3).

2.1.2. Can extrinsic properties of NFs be used to support a grouping hypothesis?

Box 2 exemplifies how the user can use an extrinsic property (e.g. photocatalytic activity) to refine the choice of potentially relevant intrinsic properties (e.g. crystallinity) relevant to a grouping hypothesis. This can be achieved if the link between the extrinsic property and the biological response is clearer than the link between the intrinsic property and the biological property. In box 2, the photocatalytic activity can be related to a toxic impact on organisms, and then the extent of toxicity can be calibrated against the proportion of anatase and rutile in the NF. Extrinsic properties have also been used to refine the role of intrinsic factors in previous grouping frameworks (Arts et al., 2015). In addition, extrinsic factors of non-nano chemicals have been used to refine the role of an intrinsic factor in justifying why substances can be grouped. For example, intrinsic properties such as atomic ratios, which are difficult to measure, can be compared to the extrinsic property of octanol-water partition coefficient (log k_{ow}) of a non-charged organic chemical, for which standardised methods exist. The octanol-water partition coefficient is therefore routinely tested as an extrinsic factor when investigating environmental fate. As data becomes available to allow hazard or fate to be predicted based on a regular pattern of that property, or when the underlying mechanism of toxicity is understood, it is likely that the relevance of an increasing number of extrinsic properties for NFs may be identified as useful for grouping and read-across (Box 2).

2.1.3. What happens if the application range encompasses the full dynamic range of a property?

To make a group, the application range would usually be a subset of the full dynamic range (concept in Fig. 4, detailed examples in Table SI_2); this particularly applies in an analogue approach. Only as part of a category approach, may a full dynamic range be applicable for a grouping hypotheses, e.g. when the full range of the property shows a regular pattern between the property and the endpoint response (e.g. crystalline structure under UV light conditions, see Box 2). In contrast, when the full range shows no relationship between the property and the endpoint (i.e. slope of the relationship is zero, e.g. crystalline structure under dark conditions, see Box 2), the property is not considered to have environmental or biological relevance. This means that the specific property will not distinguish group members from non-group members. Such properties were removed from the respective IATAs of the GRACIOUS Framework (Stone et al., 2020).

2.1.4. How can the application range be defined?

For both analogue and category approaches to read-across, the application range of a property relevant to the grouping hypothesis is closely linked to the idea of the "biologically or environmentally relevant range" (Fig. 4). In the case of the category approach, the trend may

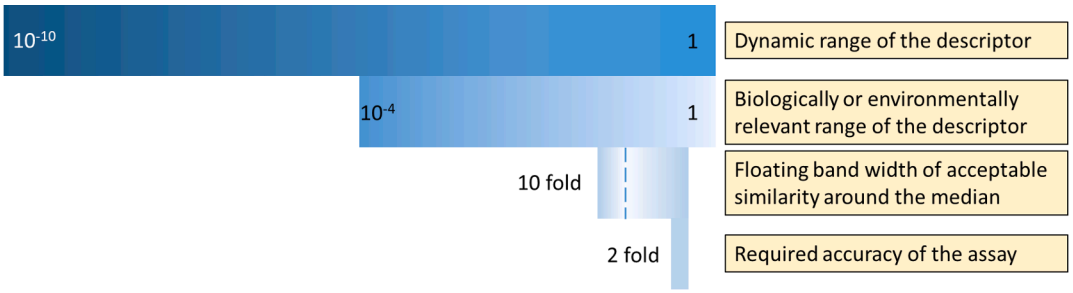

Fig. 5. Schematic representing how the different ranges of values of attachment efficiency relate to each other, dynamic range of the descriptor, biologically or environmentally relevant range, floating bands and required accuracy.

hold true across the full dynamic range of the property, for example, how *Daphnia* sp. long term NOEC scaled with the proportion of anatase and rutile in TiO_2 in Box 2. Alternatively, the application range may be related to a critical range where the trend holds true, such as the environmentally relevant range for attachment efficiency and its relationship to predicted environmental concentrations in water, as demonstrated in Box 3.

In the absence of a trend or regular pattern in a property (e.g. during an analogue approach), the application range is the region of the property in which similarity assessment by floating bands is appropriate. Floating bands describe a range, or a distance from the median value for the proposed group. The measured values for a NF property must fall within the floating band for the NF to be considered similar to the other group members. The width of these floating bands should be informed both by the dynamic range for the given property, and by how this range relates to a measurable change in the associated hazard endpoint, as demonstrated in Box 3. Once again, this application range is closely related to the biologically or environmentally relevant range, where beyond a cut-off, measurable differences between NFs are irrelevant for their biological behaviour and must be disregarded in the similarity analysis to prevent wrongly excluding NFs from a group (Fig. 4). For example, with respect to the EFSA cut-off of 10 min dissolution halftime (Hardy et al., 2018), NFs with 0.05 min halftime and those with a 5 min halftime are both considered soluble, and must be dealt as one group despite their 100-fold difference in this important descriptor. The 100-fold difference in dissolution halftime is deemed irrelevant for their biological behaviour below this cut-off of 10 min. Another example of a cut-off is fibre length, where the threshold of fibre length greater than 5 μm provides a justified cut-off to group fibres relevant to human inhalation (Murphy et al., 2021). Differences between fibres >5 μm in length are measurable, but are considered to all experience the same limited clearance in the lungs, and so are considered sufficiently similar without further assessment.

Box 3
Demonstrating the dynamic range, environmentally relevant range, floating band width and required accuracy of a property: an example of attachment efficiency in the aquatic environment

According to the NanoFASE project 'Attachment efficiency expresses the probability that upon collision of a particle with another surface or another particle, the two particles will stick to each other. In the case of favourable attachment, attachment efficiency equals 1, i.e. all collisions induce the growth of an agglomerate'. Attachment efficiency may be expressed on a logarithmic scale ranging from 1 down to 10^{-10}, where 10^{-10} represents low or no attachment. SimpleBox4Nano (SB4N) (Meesters et al., 2014) is a nanospecific adaptation of the SimpleBox exposure model which simulates environmental fate of chemicals as mass flows between environmental compartments (air, water, sediment and soil). The SB4N exposure model demonstrates that attachment efficiency is one of the most important descriptors for estimation of predicted environmental concentration (PEC) in freshwater systems, but that is only true when attachment efficiency is greater than a threshold of 1.1×10^{-4} (Meesters et al., 2019). Below this cut-off, whilst differences in attachment efficiency might be measurable, they are considered irrelevant for the particles environmental fate, as all particles will remain as unbound "free" particles. Above this critical attachment efficiency of 1.1×10^{-4} is considered to be the environmentally relevant range of this property. Within this range the predicted environmental concentrations (PECs) of unbound "free" particles of low solubility NFs decrease linearly with increasing attachment efficiency (i.e. they will form agglomerates). This trend represents a regular pattern within a category approach to grouping, where the applicability range of the group covers attachment efficiencies $>1 \times 10^{-4}$. The measurement technique must be able to measure in this environmentally relevant range to distinguish between dissimilar NFs when considering their fate in the aquatic compartment. The attachment efficiency measurement is sensitive to the method used and the conditions under which it is assessed, and so a single best approach to assessment of attachment efficiency has yet to be identified (Praetorius et al., 2020). It has been proposed that for attachment efficiency, an order of magnitude difference is sufficient to conclude that two NFs are similar. This is because, when attachment efficiency increases by an order of magnitude, a relevant difference in fate between NFs might be expected (Svendsen et al., 2020). Therefore the width of the floating band to conclude that NFs are similar with respect to attachment efficiency in the aquatic compartment would be within a 10 fold difference about the median of the group. The required accuracy of the assay should be a proportion of the width of the band that defines the limit for similarity. To illustrate, if the required accuracy is 20% of the floating band, the required accuracy for derivation of attachment efficiency for similarity assessment of the fate of NFs in water should be 2 times the median of the group (*Fig. 5*). Whilst this required accuracy is desirable, non-standard methods with lower accuracy might be acceptable. For example, if the property of a NF is closer to the median of the group (where there is room for error in the data range), the required accuracy is arguably lower than if the candidate NF is closer to the limits of acceptable similarity (where you need to be certain which side of the band limits the data lies).

This required accuracy of two-fold the median of the group would be sufficient to allow for grouping by either the category or analogue approaches. For the category approach, predicting PEC of free particles in water would rely on the relationship between attachment efficiency and PEC of "free"NFs. For an analogue approach, similarity could be justified if the attachment efficiency falls within the floating band about the median of the group.

Table 1
Data reduction from distributions to scalar descriptors. The distributions are generated by measuring a coordinate (e.g. % dissolved) in dependence of an ordinate (e.g. time t).

Endpoint/Property	Method examples	Two-dimensional distribution examples	Example Scalar descriptors used to assess and compare
Inflammation	OECD TGs, e.g. TG413 inhalation	Dose-response: neutrophil numbers if bronchoalveolar lavage	NOEC, LOAEC, EC50
Reactivity (abiotic) or Cytotoxicity (in vitro)	In vitro assays with e.g. LDH or MTT detection, DCFH assay, FRAS assay, EPR assay	Concentration (c) -response: Fluorescence (c), LDH(c), MTT(c), BOD(c), …	BMDx, LOAEL, mBOD,
Size	NanoDefine methods	Diameter distribution: Number (D)	D50
Surface	ISO9277:2012	N_2 adsorption isotherm	BET-surface area
Charge	Zetasizer with pH titration pH 4 to pH 10	Zeta potential: ζ (pH)	Iso-electric point (IEP)
Dissolution	ISO 19057:2017 and OECD draft TGs	Kinetics: % dissolved (t)	Half-time or rate
Dispersion stability	TG318:2017	Kinetics: % stable (t)	% stability at 6 h
Dustiness	EN17199:2019	Aerosol number (t)	Dustiness index DI_N

Abbreviations: c: concentration of NF in the test, t: time of the test, TG: Test Guideline, NOEC: No Observed Effect Concentration, LOAEC: Lowest-observed-adverse-effect Concentration, LDH: Lactate dehydrogenase, DCFH: Dichloro-dihydro-fluorescein, FRAS: Ferric Reduction Ability of Serum, EPR: Electron Paramagnetic Resonance, EC50: Half maximal effective concentration, BMD: Benchmark Dose, LOAEL: Lowest Observed Adverse Effect Level, mBOD: mass-metric Biological Oxidative Damage.

2.1.5. How can methods be identified to provide sufficient accuracy to support grouping?

For the similarity assessment to be meaningful, the assays should be of sufficient accuracy to identify biologically or environmentally relevant differences between NFs. The achievable accuracy for measurement of a given physicochemical property is limited by several factors, including an appropriate specification of the property, the sample preparation, the method of measurement (BIPM, 2008). In practice, it may not be necessary to apply the highest resolution or the best achievable accuracy of an assay, if such resolution is not necessary to identify biologically relevant differences between NFs. Also the polydispersity of each candidate NF in that property may give a practical limit of the required accuracy. For example, 80% of size bins are identical for one NF with D10 of 10 nm, D50 of 30 nm, and D90 of 60 nm, and another NF with all values shifted by 5 nm. A resolution of the size descriptors of better than 5 nm, which is 17% or 1.17-fold on D50, is not required, because of the high similarity by overlapping polydispersity. In general the required accuracy for a descriptor can be determined from the scale type (e.g. linear or logarithmic scale) and the range for the given property that is biologically relevant. Both parameters may be inferred using representative test materials (RTMs) that span the upper and lower limits of the biologically relevant range of an assay (Fig. 4). These serve as a point of reference or „benchmark“to support the interpretation and assessment of results for a new test material. For some relevant assays, benchmarks have been proposed (Wohlleben et al., 2019). The difference between RTMs (of known hazard) can also aid in understanding the achievable and required accuracy of a test method, (Cross, 2021a) and the biologically relevant ranges (Table SI_ 2).

2.1.6. Examples of how the application range can be determined

An example of the application range of a descriptor in an environmental compartment is provided in Box 3. The descriptor discussed is attachment efficiency, where attachment efficiency describes the probability for collisions between engineered nanoparticles and other particles (including both other engineered nanoparticles and natural particles) to result in aggregation. The example shows how derivation of attachment efficiency influences understanding of the environmental fate of the NF. A ten-fold change in attachment efficiency around the default value (0.01) in the SimpleBox4Nano (SB4N) (Meesters et al., 2014) model had a negligible effect on the model output (Salieri et al., 2019). For this reason, within a ten-fold range in attachment efficiency, NFs are considered similar for this parameter.

A different example of the level of similarity needed to define a group can be derived from (non-NF) mineral fibres. Fibres above the length cut-off (when rigid and biopersistent) are difficult to clear from the lung and surrounding pleural tissues, as they are larger than macrophages which are responsible for ingesting and clearing particles, as well as stomata (pores) that allow particles to drain from the pleural cavity around the lung (Murphy et al., 2021). In this case study, dissolution is the property that was most related to the hazard (Oberdörster, 2000; IARC, 2002). The World Health Organisation (WHO) considered fibres with a range of dissolution rates from 13 to 329 ng/cm^2/h (equivalent to a 25-fold range). Fibres with a high dissolution rate exhibited no significant hazard in terms of fibrosis or tumours in animal models, whilst fibres with a low dissolution rate induced both disease endpoints. Clear thresholds could be observed within the data, so that at a dissolution rate of 72 ng/cm^2/h (approximately 5-fold greater than the lowest dissolution rate, and 5-fold lower than the highest dissolution rate) only fibrosis was observed (IARC, 2002). This suggests that 5-fold ranges in dissolution rate are biologically relevant with respect to the ability of respirable fibres to induce fibrosis and tumours in animal models. In addition, 5-fold would be within the usual 10-fold width of the hazard classification codes (e.g. specific target organ toxicity (STOT) for repeated exposure (RE)) used for the Classification, Labelling and Packaging (*CLP*) *Regulation* ((EC) No 1272/2008).

2.1.7. Not all methods for assessing properties cover the full biologically relevant range for that property

For most properties, the biologically relevant range is narrower than the measurable range (Table SI_2). One exception is the method used to assess dustiness. Dustiness is important when estimating whether a powder is respirable. The lowest dustiness index that is measurable is about 10 mg of respirable particle released per kg of bulk particle (measurable by the EN17199 gravimetric methods). Considering all NFs with lower dustiness as similar in this property (Fig. 4) might not be sufficient for high toxicity materials, for which low exposures (and hence low dustiness index values) may induce relevant risk (Table SI_ 2). The biologically or environmentally relevant range of a certain descriptor also depends on the hazard endpoint included in the grouping hypothesis. RTMs (Box 1) with known toxicity serve as a point of reference to support the interpretation of results obtained for a new test material. For example, dissolution half-time indicates how quickly a NF can dissolve in a specific media. If the half-time for dissolution is shorter than the time that it takes the NF to reach viable cells, then the NF and released ions/molecules will induce equivalent biological effects and therefore can be considered similar. This criterion has been used to set the lower biologically relevant range for dissolution half time to 10 min. in the regulatory assessment of oral exposure to NFs (Hardy et al., 2018). By analogy, limits of the biologically relevant range of pulmonary exposure have been proposed (Keller et al., 2021).

In general, for each method, property and descriptor that are finally selected (Table 1), it is necessary to derive the biological relevance of any measurable differences, following the logic of the example in Box 3. The regulatory acceptable limits of similarity can be derived by a

calibration strategy (Nymark et al., 2020), as the GRACIOUS project has done for selected pre-defined hypotheses. In short, we first established NF groups with methods suitable for regulatory purposes (e.g. Tier 3 methods for human hazard assessment requiring in vivo testing). This did not mean every member of the group had data suitable for regulatory purposes, just that there were sufficient examples to define the group. The second step was to compare the data to Tier 1 (physicochemical and in vitro) results to assess how the in vitro and in vivo parameters align for a specific group. Tier 1 groups needed to be conservative, and so we narrowed the limits on acceptable similarity until application of Tier 1 methods resulted in the exclusion of some candidate NFs, even though such group members were acceptable according to Tier 3 testing (Fig. 6). The outlook section of this article lists case studies using this approach, but all depend on:

a) robust measurements, to enable quantitative comparison
b) known accuracy of these measurements, to exclude interpretation of insignificant differences
c) algorithms, to quantify similarity
d) calibration by case studies with available regulatory data

The next sections 2.3 and 2.4 explore items a and b respectively; section 4.1 and 4.2 explore items c and d respectively.

2.2. Robust measurements by data reduction to descriptors

Data is often available as a distribution (e.g. a concentration response curve for hazard) which is too complex to use routinely for a similarity assessment with large data sets. Instead data distributions can be converted to a single value known as a scalar descriptor (e.g. LC50), which is used to represent the response of an organism, or of a cell culture, to a range of concentrations of the NF. A similarity assessment can then be applied to the scalar descriptor values to determine the suitability of NFs for inclusion within the group. More than one type of scalar descriptor is available to assess dose response curves (e.g. Benchmark Dose (BMD) (Crump, 1984), No Observable Adverse Effect Level (NOAEL) (Brown and Erdreich, 1989)), allowing data reduction to be adapted according to the assay or the endpoint assessed. By applying the restriction that the same assays are to be used for data generation for all NFs within a proposed group, we ensure that NOAELs are not only numerically similar, but that they reflect the same type of response. Similarity assessment requires a robust data acquisition method and robust evaluation of raw data towards descriptors, and we can rely on established algorithms to generate descriptors by data reduction from the often two-dimensional distribution that are originally measured (Table 1). Algorithms exist that compare the two-dimensional and higher dimensional distributions. However, such data often varies between different sources, because laboratories using the same method can acquire data with different data spacings (e.g. concentrations ranges 1–10-100 vs. 1–5 – 25) and ranges of the ordinate axis (concentration, diameter, time, …). Only some algorithms can deal with such differences in data sets. Additionally, distributions can be noisy, scaled by detector sensitivity, or distributions are simply not accessible. Robust study summaries provide established scalar descriptors (Table 1) to assess substances, including those in the NF. Some of the scalar descriptors are required to register a NF under REACH, e.g. the median size D50 or the specific surface area (BET-surface). Others are harmonized by OECD Test Guidelines (TGs), e.g. the % dispersion stability in environmental medium. It is worth noting that the grouping of non-nano chemicals also uses scalar descriptors, but the choice of descriptors for non-nano vs NFs might be different, e.g. using log k_{ow} instead of dispersion stability, or molar mass M_w instead of size D50.

2.3. Experimental determination: achievable accuracy

A similarity assessment cannot be used for grouping if the measured differences are below the accuracy limits of the method. In another paper of this special issue (Cross, 2021a) we demonstrate that some standardised basic physicochemical properties can be reproduced across four experienced laboratories with just a few % accuracy limits, whereas others approach 2-fold and even 5-fold uncertainty. The European project PATROLS determined that an uncertainty factor of between two-fold and ten-fold (for well-dispersed NFs) can arise when predicting the deposited dose on *in vitro* submerged cell cultures (Keller et al., 2020). As a consequence, two-fold differences in the *in vitro* effect between two NFs are not sufficient to conclude a lack of similarity, and for some NFs this might extend to a ten-fold difference. Ongoing OECD TG projects will enhance the robustness of testing, but for the time being, the limits of acceptable similarity need to allow a factor of about two for measures of surface reactivity (Bahl et al., 2020) and in vitro inflammation (Krug, 2018; Elliott et al., 2017).

Often the measured values are below the limit of quantification (Fig. 4). For the human reader, it is recommended that the data matrix is then edited to replace any values that are beyond the limit of quantification with the highest or lowest value that can be accurately measured along with a < or > sign to indicate that it is probably higher or lower respectively. For the numerical assessment of similarity, the < or > sign is disregarded, resulting in a perfect similarity score for a pair of NFs that are both outside the measurable range (Fig. 4). One of the case studies implements such an editing or cropping strategy for very slowly dissolving, very low reactive NFs (Jeliazkova, 2021). The cropping to biologically relevant ranges is directly available in the browser-based similarity tool(Enanomapper similarity tool, 2021) and the GRACIOUS Blueprint.(Traas and Vanhauten, 2021)

3. Using IATAs to complete a data matrix and support a similarity assessment

GRACIOUS has established more than 40 IATAs that help the user to gather the evidence needed to assess environmental or human hazard hypotheses, as well as decision trees for the assessment of exposure. In the present paper, we do not discuss the testing, which is informed by guidance (ECHA, 2019a) and numerous correlative literature reports (Kühnel et al., 2019; Hund-Rinke et al., 2018; Drew et al., 2017; Bahl

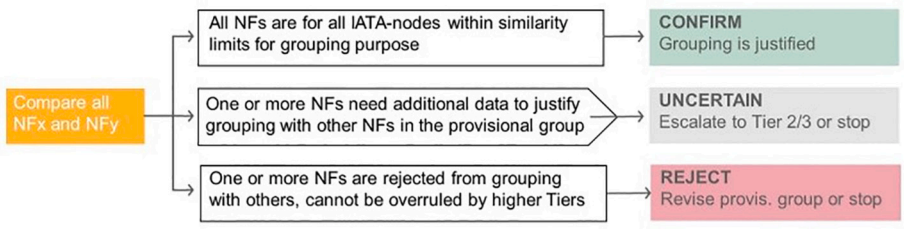

Fig. 6. Pairwise comparison of all NFs (NFs) in the candidate group needs to consider all properties that were identified as relevant decision nodes of an integrated approach to testing and assessment (IATA). The assessment can be based on limits of acceptable similarity of each property individually, or can assess a multidimensional distance, but in either case there are three outcomes: the similarity is sufficient to confirm grouping, or is uncertain, requiring more reliable methods from higher tiers, or is rejected, requiring revision of the group or abandoning grouping altogether. See section 4.3 for a data-based approach.

Table 2
Algorithms to assess similarity between descriptors, or (for Bayesian model and Arsinh models) to also assess similarity of dose-response curves.

Algorithm name	previous application (beyond NF)	Previous application to NF grouping	Appropriate descriptors	Pros & cons
x-fold difference	X-fold changes are commonly used to evaluate effects on most biological parameters, as most of these are known to follow log-normal rather than normal distributions.	ECETOC NanoApp (Janer et al., 2020)	Descriptors that follow log-normal distributions. If applied to multidimensional grouping, additional rules to combine scores are needed.	Pro: minimal mathematical processing. GRACIOUS implementation is available online.(Enanomapper similarity tool, 2021) Con: It does not correct for differences in the biological relevance over the dynamic range, such as noise and accuracy limits
Bayesian model assessment	BF model comparisons are used extensively in biology and pharmacology as an alternative to typical hypothesis testing. The method can deal with many types of distributions.	(Marvin et al., 2017; Furxhi et al., 2019)	Application to all descriptors, however the specific method is tailored to dose-response data.	Pro: Able to incorporate literature or previous knowledge from public data. Con: Implementation is difficult, as it needs adjustment for different statistical distributions depending on the data analysed.
Arsinh-OWA model	New proposal still to be published and validated.		For scalar descriptors with benchmarks.	Pro: Based on absolute distance metric, derived groups are not relative to the assessed entities. Con: Requires establishing a proper threshold for scaling.
Euclidean distance	Standard widely used method.	Many e.g. (Bahl et al., 2020)	Any numerical descriptor values.	Pro: Standard method, easy to implement, multi-dimensional; may need data pre-processing. GRACIOUS implementation is available online.(Enanomapper similarity tool, 2021) Con: Assumes data follows normal distribution; does not work with missing data, but variants are available.
Cluster analysis	May use any distance/similarity method, including ones above.	Many e.g. (Bahl et al., 2019; Cai et al., 2018)	Any descriptor values.	Pro: Many clustering methods available, easy to implement, visualization possible. Con: not easy to interprete. Data gaps on one descriptor require additional processing

et al., 2019; Karkossa et al., 2019), but focus instead on discussing how to use the generated data to assess similarity.

Each IATA selects the properties (e.g. basic information, extrinsic properties, in vitro assays, acute ecotoxicity studies) and the most appropriate scalar descriptors to assess NF grouping. Box 2 provides an example of the selection of the most relevant (extrinsic) properties, which often directly measure interactions between the NF and a well-controlled medium, instead of trying to predict such interactions from basic (intrinsic) properties. In the supporting information, Table SI_1 presents an example data matrix that demonstrates key notions:

- A data matrix contains (i) *all and only* the properties of the IATA that applies to a specific hazard, (ii) the NF physicochemical parameters that allow identification of the NF (size etc.), and (iii) several descriptors (e.g. content of different elemental impurities) per property (in this example: composition).
  - o In this manner, data from (ii) and (iii) must be known (European_Chemicals_Agency_(ECHA), 2019) to delimit the group boundaries, i.e. to include and exclude NFs from the group. Often they are only used as supporting parameters, e.g. the highly reproducible BET is used to *evaluate* a surface-based reactivity, which is a *decision* criterion, but BET is *not necessarily* a decision criterion for grouping in itself. Similarity must be assessed on data from (i) IATA decision nodes.
- The descriptors included in a data matrix are *consistent* for all NFs, and thus enables a *pairwise comparison* between all pairs of NFs in the group.
- Each method needs well-defined *control* materials, which enable testing laboratories to demonstrate proficiency.
- For some methods the control materials coincide with RTMs that represent certain biological behaviour, and thus define the biologically relevant range (Box 1, Table SI_2). The RTMs are in general different substances to those in the group, and are included in the data standardisation (i.e. data is transformed to a mean of 0 and variance of 1) before applying similarity algorithms.

Such a data matrix will be used to collate and organise the data, and to demonstrate the outcome of similarity assessments. It is also useful to quickly identify data gaps. The methods used for similarity assessments are introduced, applied and compared in the next section. The similarity assessment can be performed property-by-property, as demonstrated in section 4.2, or simultaneously for all of the data across all decision nodes, as demonstrated by multidimensional similarity analysis in section 5.2. In either case, the *methods* used to acquire the data for the decision nodes can be tiered. Since the same descriptor is required for all NFs to allow a similarity assessment, all NFs therefore require data for the same tier of testing. After each tier, a similarity assessment can support the decision to confirm grouping, to escalate, or to stop grouping (Fig. 6) (Stone et al., 2020).

It is beyond the scope of the present White paper to establish detailed rules for the escalation between tiers. However, the approach described in section 2.1 will ensure consistency, especially for the property-by-property assessment, which requires for each property and its descriptor, calibrated limits of similarity that are acceptable for this specific purpose of grouping, and for this specific hazard. Section 4.3 demonstrates this strategy without claiming to give a final answer. A conditional escalation to higher tier methods, based on the similarity of lower tier descriptors, is the same approach taken in the ECETOC NanoApp (Janer et al., 2020), which is consistent with “floating bands” (Wohlleben et al., 2019), but is different from all schemes of banding with predefined cut-offs.

As an illustration, sections 4.2, 4.3 and 5.2 include the properties and descriptors relevant to the hypothesis for respirable, very slowly dissolving NFs with low (acute) toxicity, for which accumulation of particles in the lungs can occur and can lead to increased likelihood for long-term toxicity after chronic exposure. Similarity in all the descriptors considered along the decision nodes in this IATA would support read-across for long-term toxicity after chronic exposure. The decision nodes that are included in this IATA are related to the degree of deposition in the distal region of the lung, dissolution in lysosomal fluids, NF reactivity, and induction of inflammatory markers. The IATA considers read-across towards a well characterised poorly-soluble low toxicity NF or towards another NF of the same substance under consideration, as long as similarity in all the descriptors described supports its use. The

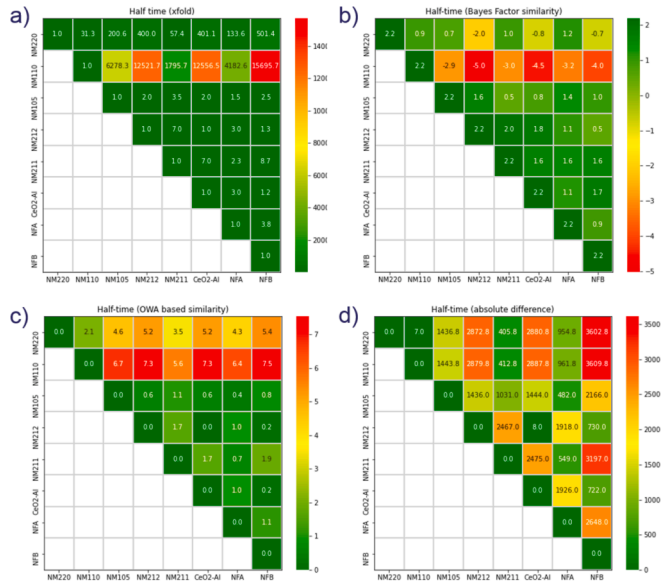

Fig. 7. Demonstration of different algorithms applied to the "half-time" descriptor of the "lysosomal dissolution" property. The original data is shown in Table SI_ 1. The assessment of similarity comprises the RTMs, which represent the biologically relevant range, and several NFs of CeO_2 and Fe_2O_3 respectively. The chemical composition of the materials from the JRC repository is CeO_2 (NM211, NM212), TiO_2 (NM105), ZnO (NM110), $BaSO_4$ (NM220). The scale of each panel is different, but dark green always indicates a pair of two very similar NFs, and red indicates a pair of two very different NFs. a) X-fold algorithm; b) Bayes Factor similarity approach; c) Arsinh-OWA approach; d) absolute distances (the one dimensional equivalent of Euclidean distance). For demonstration purposes, the original data was not restricted to the biologically relevant range, whereas this should be done in a full implementation of the similarity assessment. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

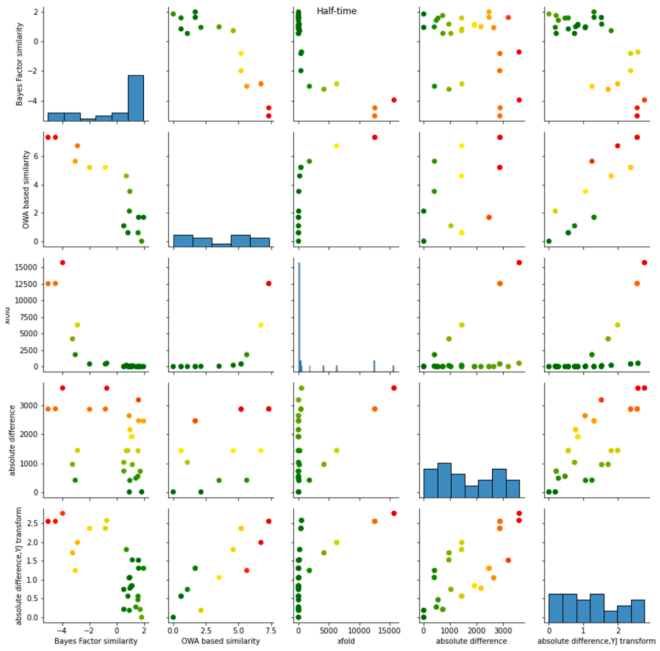

Fig. 8. Comparing the similarity assessment by different algorithms: Each dot is a pair of NFs assessed by two different algorithms (Fig. 7), on the basis of the "half-time" descriptor of the "lysosomal dissolution" property (Table SI_ 1). The assessment of similarity comprises the RTMs, which represent the biologically relevant range, and several NFs of CeO2 and Fe2O3 respectively. Appropriate data standardisation by a power transformation (see also Figure SI_1) significantly improves the consistency of the absolute difference (euclidean) metric with the other three metrics (Arsinh-OWA, x-fold, Bayes), which are mostly consistent with each other. Blue bar plots are histogram that illustrate the range of similarity scores for the whole dataset for one particular algorithm. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

data matrix (Table SI_1) holds experimentally determined data of all properties for two case studies and their RTMs, for this specific IATA.

4. Pairwise similarity assessed property-by-property

4.1. Tools to assess similarity

Here we outline algorithms (Table 2) that can be used to conduct a pairwise similarity comparison:

- The x-fold comparison algorithm is the simplest method and can be implemented even in *excel*, but is more elegantly supported a browser-based tool that also generates the graphical output.(Enanomapper similarity tool, 2021) It is also embedded in the publicly available GRACIOUS Blueprint pdf.(Traas and Vanhauten, 2021) When comparing descriptor values for two different NFs, the x-fold comparison divides the larger of two values by the smaller, and thus always generates an answer larger than one. Identical NFs measured by perfectly accurate methods would score 1 in this algorithm. It is most appropriate to descriptors that follow log-normal distributions, and which thus cover several orders of magnitude. This model is the basis for many of the criteria of the ECETOC NanoApp (Janer et al., 2020), and is explored for dissolution rates and half-times by Keller et al. in this special issue (Keller et al., 2021).
- The Bayesian model assessment compares two sets of values using nested sampling (Skilling, 2006). Depending on the format of the data (e. g. dose response data or single descriptors), the underlying distribution considered by the Bayesian model changes from normal to log-normal and exponential. The model assesses whether data from two NFs are derived from the same underlying distribution with the same parameters, as opposed to different distributions; this pairwise model comparison is performed by means of Bayes Factor (BF) calculations (Kass and Raftery, 1995) to show how the evidence in the data modifies our prior simplistic belief that the two NFs are not similar. Positive values support the assumption that the two NFs are derived from the same underlying distribution and so, they are estimated to be similar and can be grouped together. This model is further explained by Tsiliki et al. (2021) (Tsiliki, 2021) and applications for several reactivity assays are explored by Ag Seleci et al. (2021) (Ag-Seleci, 2021), both included in this special issue.
- The Arsinh-OWA model based clustering first applies the arsinh transformation to the distance between two NFs, and then rescales the result to the arsinh of a biologically relevant threshold, e.g. the range defined by the RTMs (Zabeo, 2021). This metric distance-based similarity allows the final aggregated distance between NFs to be an absolute distance preserving symmetry and triangular inequality, leading to groups which do not change if new members are included in the assessment. The rescaled similarity matrices are utilized for grouping by applying agglomerative hierarchical clustering in a multidimensional space. To evaluate the multidimensional distance, OWA aggregation is applied (Yager, 1996), where the highest distances among all dimensions are aggregated as the overall NF distance.
- Euclidean distance (the length of the line segment between two points) is widely used and usually the first choice of metric distance. It is supported by a browser-based tool that also generates the graphical output.(Enanomapper similarity tool, 2021) In one dimension (scalar descriptors) it is equal to the absolute value of the difference between the scalar values. Euclidean distance between two points on the plane (two dimensions) is calculated by applying

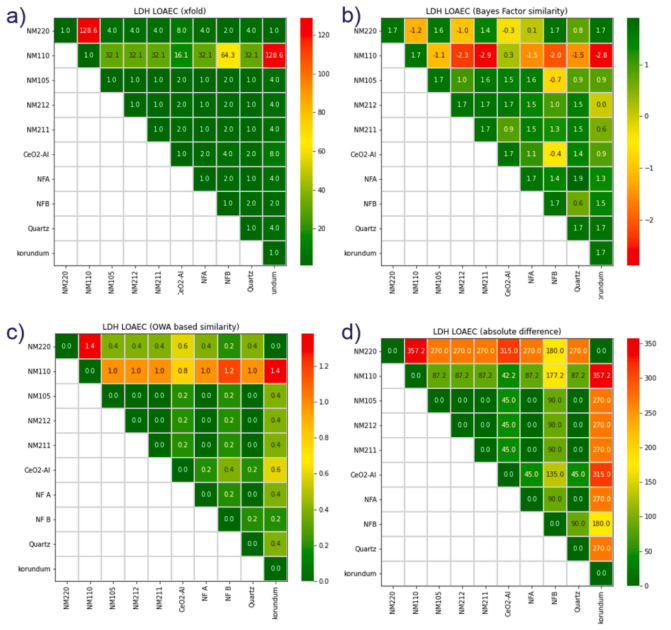

Fig. 9. Demonstration of different algorithms applied to the "LDH-LOAEC" descriptor of the "inflammation" property. The original data is shown in Table SI_ 1. The assessment of similarity comprises the RTMs, which represent the biologically relevant range, and several NFs of CeO_2 and Fe_2O_3 respectively. The chemical composition of the materials from the JRC repository is CeO_2 (NM211, NM212), TiO_2 (NM105), ZnO (NM110), $BaSO_4$ (NM220). The scale of each panel is different, but dark green always indicates a pair of two very similar NFs, and red indicates a pair of two very different NFs. a) X-fold algorithm; b) Bayes Factor similarity approach; c) Arsinh-OWA approach; d) absolute distances (the one dimensional equivalent of Euclidean distance). For demonstration purposes, the original data was not restricted to the biologically relevant range, whereas this should be done in a full implementation of the similarity assessment. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the Pythagorean theorem to a right angled triangle. The generic formula is the square root of the sum of squares of the differences between coordinates of points in a multidimensional space. One of the case studies, on inhalation of various substances and sizes of pigments, demonstrates the use of Euclidean distances (Jeliazkova, 2021).

- Dynamic Concentration Warping (DCW) method for dose response data comparison is introduced in analogy with the widely used Dynamic Time Warping (DTW) method for comparing time series. DTW was initially proposed for comparison of audio signals (Sakoe and Chiba, 1978) and has since been applied to sequences of video, audio and graphics in the context of applications as speech recognition, shape matching, financial markets. A nearest-neighbor classifier can achieve state-of-the-art performance when using DTW as a distance measure (Ding et al., 2008). The challenge in comparing dose response data gathered from different experiments is similar to comparing sequences, as different series of concentrations may be used in the dose response experiments being studied. The usual approach to fit a dose response curve (e.g. the Benchmark-Dose approach, Table 1) has the drawback that the shape of the curve has to be selected. For example, the sigmoidal shape is expected in cooperative binding of a ligand and enzyme to a receptor. Data from assays which do not rely on the same mechanism, such as acellular reactivity assays often have a different shape. The DCW method does not assume shape of the dose response data and enables automatic comparison without model selection.(Tsiliki, 2021) Given two dose response sets, a matrix is constructed with rows indexed by concentrations of the first experiment and columns indexed by concentrations used by the second experiment. Each cell of the matrix is a two-dimensional Euclidean distance between the "(concentration1,value1)" for the first material and "(concentration2, value2)" for the second material. Once the matrix is formed, the path between the top left corner and bottom right corner of the matrix is found by dynamic programming. The final distance between the dose response series is the sum of values along the path. The DCW method is used to compare dose response series in Braakhuis et al., this issue (Braakhuis, 2021).
- Clustering approaches can use any distance measure. Clustering is an unsupervised machine learning method and hence clusters may not be meaningful. Inclusion of RTMs solves the problem of scaling. Clustering analysis aims to discover two-dimensional patterns in the data matrix, searching for similarities between NFs and properties. NFs clustered together (i.e. grouped) are considered to be more similar between one another, compared to all other NFs belonging to other clusters.

4.2. *Demonstration of tools on two properties of two case studies and their RTMs*

If the distance metric is symmetrical (i.e. distance between A and B is the same as distance between B and A), as is the case for the algorithms selected here, the pairwise comparisons result in a triangular similarity matrix (e.g. Fig. 7 and Fig. 9) that allows pairwise comparison of NFs by reading along rows and down columns. The examples demonstrated here include the triangular similarity matrices used to assess the similarity of dissolution half-time data for each NF (lysosomal conditions, Fig. 7), and to assess similarity of in vitro toxicity data for each NF (LOAEC of LDH release, Fig. 9). Both of these properties are required by the GRACIOUS IATA on inhalation hazard. The data used to generate these triangular similarity matrices was extracted from the original data matrix, Table SI_1. Only a NOAEC was available for the NF $BaSO_4$ NM220, and so the LOAEC was estimated to be a factor 2 higher than the NOAEC (because this factor corresponds to the dose spacing that was

Table 3
Approaching limits of acceptable similarity: Comparing four algorithms for quantification of a similarity assessment of Tier 1 data from three decision nodes, for NFs and respective RTMs (as defined in Table SI_ 1). Note that for Bayes algorithm a lower values indicate less similarity, whereas it is the inverse for the three other algorithm, as highlighted by the RTM pair.

Surface reactivity: FRAS mBOD					
		x-fold	Bayes	OWA	Euclidean YJ
case Fe2O3	NF_A vs NF_B	3.6	1.5	3.7	0.9
case CeO2	NM212 vs NM211	1.5	2.1	1.3	0.6
case CeO2	NM212 vs CeO2_Al	5	1.5	4.9	1.2
RTMs	BaSO4_NM220 vs CuO	1033	-2	20.4	3.4
Dissolution: lysosomal half-time					
case Fe2O3	NF_A vs NF_B	3.8	0.9	1.1	1
case CeO2	NM212 vs NM211	7	2	1.7	1.3
case CeO2	NM212 vs CeO2_Al	1	1.8	0	0
RTMs	ZnO_NM110 vs NM105	6278	−2.9	6.7	2
in vitro toxicity: LDH LOAEC					
case Fe2O3	NF_A vs NF_B	2	1.4	0.2	0.8
case CeO2	NM212 vs NM211	1	1.7	0	0
case CeO2	NM212 vs CeO2_Al	2	1.7	0.2	0.6
RTMs	ZnO_NM110 vs. korundum	129	−2.8	1.4	3.6
respirable Dustiness Index					
case Fe2O3	NF_A vs NF_B	13	−1.8	0.7	2.9
case CeO2	NM212 vs NM211	2.2	1.5	0.2	1.1

used in the assay).

Colours are used in the triangular similarity matrices to indicate the degree of similarity between two NFs, with cool colours (green) indicating a high degree of similarity, and warmer colours (red) at the opposite end of the spectrum representing the NFs that are not similar. Within each figure, the same data has been analysed using the four different similarity algorithms, thereby allowing their comparison. When comparing either the partially dissolving benchmark $BaSO_4$ NM220 or the quickly dissolving benchmark ZnO NM110 with any other NFs, (Fig. 7) the yellow-red colour indicates a low degree of similarity for dissolution half time between the RTMs and the test NFs, confirming the suitability of these materials to span the relevant range on either side of the candidate group. The RTMs are not similar to the test materials according to any of the four algorithms, suggesting that the algorithms are in agreement for the RTM versus test material pairwise comparisons.

Within pairs of NFs of the same test substance, the absolute distance algorithm (Fig. 7d) finds relatively low similarity (yellow colours), which is contrary to literature findings that the substance primarily determines dissolution rates, while NF parameters modulate the dissolution rate to a much lesser extent. (Wohlleben et al., 2019; Koltermann-Jülly et al., 2019) In contrast, the other three algorithms exhibit a better consistency with such literature. For these three algorithms, green colours for the pairs of CeO_2 NFs and for pairs of Fe_2O_3 NFs, suggesting that the NFs are rather similar within each substance family.

The qualitative observations used above to compare the ability of different algorithms to assess similarity can be turned into a more detailed comparison by plotting each pair of NFs in a parity plot spanned by (and therefore comparing) two of the algorithms (Fig. 8). If the parity plot generates a random graph, as shown for absolute difference (1D Euclidean), then the consistency with the other algorithm(s) is poor. This can be attributed to the application of Euclidean distance, which assumes normally distributed data (Table 2), to a descriptor that is lognormally distributed spanning five orders of magnitude (Table SI_ 1). However, Euclidean difference can take into consideration the RTMs and by applying transformation to the data to generate a normal distribution, the results become more compatible with the other algorithms. The well-known transformations to generate approximate normality are log transform, Box-Cox transform and the more recent Yeo-Johnson (Y-J) transform (Yeo and Johnson, 2000) (the latter is used in Figure SI_ 1). The inconsistent results of the Euclidean algorithm on the non-transformed data matrix should not lead to the conclusion that a particular distance algorithm is not appropriate in general. Instead the well-known requirements of transformations and algorithms that make sense for a particular distribution of data must be respected.

Arsinh-OWA and Bayes factor algorithms take into account the range of values spanned by the RTMs, and try to normalize them (each in a different way), while the x-fold method does not include normalization. The parity plots provided in Fig. 6, can be used to identify consistency between the results generated by the different similarity algorithms. If the parity plots generate a clear monotonous pattern of any kind, then this is used to conclude that there is consistency between the algorithms. For the Arsinh-OWA and x-fold algorithms the monotonous parity plots indicate that the methods generate consistent conclusions regarding similarity. The Bayes statistical analysis is mostly consistent with Arsinh-OWA and x-fold, but indicates that small scores of x-fold and of Arsinh-OWA algorithm (greenish dots, representing differences among the most similar NFs) may not be statistically relevant. In fact, many of the half-times (Table SI_ 1) are above the biologically relevant range, and possibly even above the accuracy range of the method, because half-times above 54 weeks cannot be reliably evaluated from 1-week dissolution kinetics. The three algorithms can be brought to agreement by standardising data by scaling within the metrologically and biologically relevant ranges (Box 3). Numerical limits for each of the algorithms can then be chosen such that one obtains identical groupings. For convenience, one could adapt the colour coding of each algorithm's similarity score such that the acceptable similarity is e.g. yellow in each plot. This was not done here.

Triangular similarity matrices can also be made to assess the similarity of in vitro toxicity data for different NFs (Fig. 9). Again the triangular similarity matrices identify comparisons with the test NFs as red, suggesting they are not similar. Comparison of the test NFs suggest some are sufficiently similar to be grouped (green), but not all (yellow). The parity plots allowing comparison of the similarity assessments for the in vitro toxicity data using different algorithms are provided in the supplementary information (Figure SI_ 2). The Arsinh-OWA algorithm and the x-fold algorithm are monotonously consistent. For the most similar NFs, represented by dark green dots, the Bayesian statistics show that any differences between the NFs are not statistically relevant (Figure SI_ 2), supporting the conclusion that they are similar. The Bayesian statistics also differentiates a few NF pairs as even less similar than apparent by their yellowish/reddish colour in the x-fold and Arsinh-OWA approaches. The absolute difference (1D-Euclidean) algorithm is not comparable to the other algorithms, however appropriate transformation would improve the results.

4.3. Approaches to quantitative limits of acceptable similarity

It is important to reiterate from section 2.1 that the regulatory acceptable limits of similarity need to be derived by a calibration strategy, as is usual for alternative methods. To achieve this, case studies need to first establish NF groups with methods that are accepted by regulators (e.g. Tier 3 in vivo testing). Then the Tier 1 similarity plots (such as Fig. 9) can be evaluated for each of the properties of the respective IATA, such that a Tier 1 limit of acceptable similarity results in a conservative grouping., In other words, Tier 1 may exclude candidate NFs from the group although Tier 3 justifies their grouping, but Tier 1 must not include candidate NFs that are excluded by Tier 3. Such a validation can be performed independently for each of the properties that were selected for a specific IATA. A practical example to demonstrate the process is provided by CeO_2 NM211 and CeO_2 NM212, described in (Keller et al., 2014). These two NFs are relatively similar in terms of their available Tier 3 in vivo inhalation data, and could therefore be grouped. Both are uncoated NFs of the same crystallinity and both have multimodal shapes, but they differ in size and surface area. An aluminium-doped NF, CeO_2_Al, may be considered as another candidate group member. Table 3 lists the decision nodes of the relevant

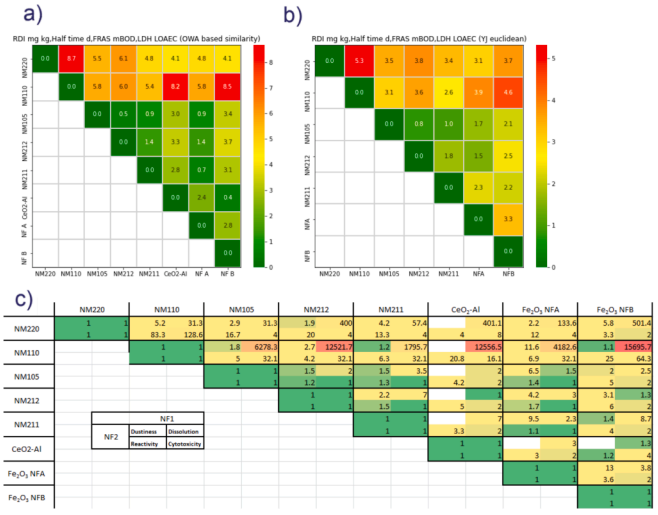

Fig. 10. Multidimensional pairwise comparison by the four decision nodes of the inhalation IATA a) Arsinh-OWA distance; b) Euclidean distance metric with Yeo-Johnson transformed data; c) comparison to the individual four decision nodes in the x-fold algorithm. The chemical composition of the materials from the JRC repository is CeO_2 (NM211, NM212), TiO_2 (NM105), ZnO (NM110), $BaSO_4$ (NM220).

inhalation IATA, and includes the similarity assessment of their Tier 1 data for the proposed group NFs and the RTMs (as defined in Table SI_ 1). Note that different RTMs are used for different decision nodes, as the RTM needs to be appropriate to the method employed and the property assessed.

In order for NM212 and NM211 to be grouped, we will now approach limits of acceptable similarity from the *lower* side accordingly. With the grouping decision in mind, we must accept (Table 3) at least a 1.5-fold difference in surface reactivity, and at least a 2.2-fold in the dustiness index. This is in line with our discussion of the measurable accuracy (section 2.3). The CeO_2 NF example is not useful to establish an acceptable similarity in dissolution, because both half-times are beyond the biologically relevant range (section 2.1 and, Table SI_ 2, Fig. 4).

The comparison of NM212 to the doped CeO_2_Al represents a case where grouping may not be appropriate. In order to prevent their grouping, we will now approach limits of acceptable similarity from the *upper* side. With the non-grouping in mind, 5-fold difference of surface reactivity should not be accepted. This rule ensures conservative decisions in Tier 1 (i.e. non-grouping). The other decision nodes of the specific IATA are not useful to ensure conservative decisions on this NF pair (CeO_2 NM212 vs CeO_2_Al) in Tier 1, because the two NFs differ even less in other properties.

For comparison, consider the least similar cases, represented by the RTM pairs: a similarity of 1033-fold ($BaSO_4$ NM220 vs CuO for reactivity), 6278-fold (ZnO_NM110 vs NM105 for dissolution), and of 129-fold (ZnO_NM110 vs korundum for in vitro toxicity) is obtained. In comparison, the acceptable similarity in the analogue approach is narrow compared to the biologically relevant range (Fig. 4). This outcome might differ for other IATAs.

In summary, for the calibration exercise using the two pairs of CeO_2 NFs, the limit of acceptable similarity for the x-fold algorithm will for many properties range at or below 5-fold, in line with the evaluation in sections 2.1 and 2.3, whereas the RTM pair (Table 3) and the biologically relevant range (Table SI_ 2, Fig. 4) is around 100-fold to 1000-fold for many properties. One can read from Table 3 the values of the other similarity algorithms that lead to the same grouping decision as the 5-fold limit in the x-fold algorithm: the acceptable similarity limit in the Bayes algorithm is approximately at or above 1.5, whereas the RTM pairs have scores around −2 to −3. The acceptable similarity in the Euclidean algorithm with Yeo-Johnson transformed data is approximately at or below 1.3, whereas the RTM pairs have scores around 2 to 3. The RTM scores in the Arsinh-OWA algorithm differ between different properties, which makes it more challenging to recommend limits of acceptable Arsinh-OWA similarities.

As explained in section 3, NF pairs that remain within acceptable similarity limits for all properties of the IATA, are “confirmed” for grouping (Fig. 6). If the limit of acceptable similarity is exceeded, that NF pair is “uncertain” and must be excluded from the candidate group or be re-assessed by higher tier methods. We did not yet establish rules for “rejection” (Fig. 6), but this may be triggered by a similarity score approaching the score of the RTM pair (i.e. the least similar case).

Case studies help to define acceptable similarity limits for many more IATAs and their relevant properties. Future users of the GRACIOUS Framework can justify their grouping decisions if their candidate NF group has quantitative tier 1 similarity scores consistent with the scores of acceptable similarity in the GRACIOUS case studies. Here, we exemplified the “calibration concept” on one pair with confirmed tier 3 (in vivo) similarity, from a borderline pair with confirmed tier 3 difference, and from the RTM pair with strong tier 3 difference (Table 3).

5. Multidimensional distances & cluster analysis

5.1. Tools for multidimensional distances & cluster analysis

The evaluation of NF similarity to support grouping and read-across for a given toxicological endpoint will generally need to consider several

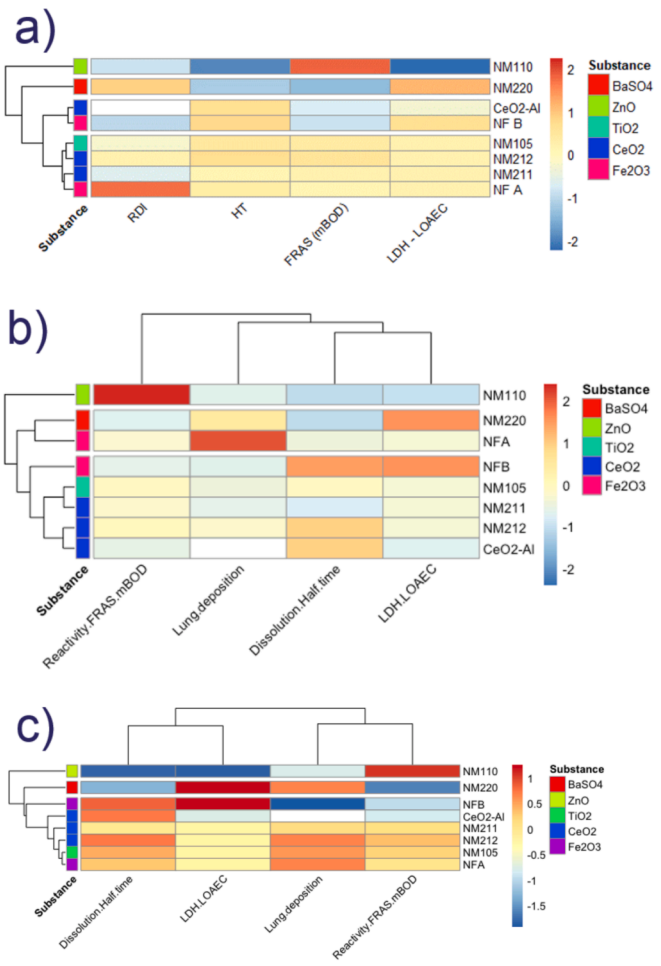

Fig. 11. Hierarchical clustering by the four decision nodes of the inhalation IATA a) Arsinh-OWA b) Minkowski distance metric; c) Euclidean distance metric on YJ-transformed data. The horizontal dendrogram shows the clustering of the NFs together with a colour code which categorizes them based on substance material.

decision nodes at the same time, but this can be implemented either by assessment of similarity property-by-property (section 4) or by an aggregated similarity measure using a multidimensional vector space (this section). Similarity assessment of NFs in multidimensional space could be achieved by multidimensional distances and/or by employing unsupervised statistical methods, such as clustering analysis, which aims to discover underlying patterns and relations in the dataset. The end result is an optimal number of clusters (or groups) of NFs, where NFs which are members of a specific group are considered to be more similar to one another than to those in other clusters. Similarity is measured by means of multidimensional distances and reveals the similarity for descriptors of the NF.

Clustering analysis is an unsupervised learning technique that is very useful to explore patterns of more than two NFs within the data matrix. In many occasions and under different settings such algorithms were employed in NF data sets. For instance, Liu et al. conducted multidimensional scaling analysis, along with hierarchical clustering, in order to illustrate the main underlying structure of a soil bacterial community dataset, whilst distance correlation coefficients were calculated to assess the consistency of NF impacts summarized at different taxonomic levels (Liu et al., 2015). Unsupervised clustering analysis is also employed with toxicogenomics data (Nikota et al., 2015; Tsiliki et al., 2017) to demonstrate that in vivo tissue ‘omics data can be used to effectively and efficiently screen new NFs and prioritize them in terms of their toxicity profiling. In an earlier attempt, Shaw et al. employed hierarchical

clustering to biological interaction profiles to support groupings for NFs, and further assign the identified clusters in domains, in an attempt to categorize them and use them in decision-making processes (Shaw et al., 2008).

5.2. Demonstration of tools on a real data matrix with two case studies and their RTMs

Using the data matrix, the multidimensional similarity for the different NFs is calculated for all decision nodes at once, i.e. the respirable dustiness index, the half-time of lysosomal dissolution, the surface reactivity in mass metrics (mBOD) from the FRAS assay, and the LOAEC of LDH release in macrophage cell cultures. These methods and descriptors represent all decision nodes of the GRACIOUS inhalation IATA and each forms a dimension of the multidimensional similarity assessment (Fig. 8a,b).

In the previous section we described the x-fold approach to assess similarity, which does not aggregate decision nodes to a single multidimensional distance. The property-by-property pairwise assessment for four descriptors is summarized into a single triangular similarity matrix (Fig. 8c). Using the x-fold approach for each property of the IATA indicates that CeO_2 NM212 and CeO_2 NM211 are relatively similar (up to 2.2-fold) in dustiness, reactivity and cytotoxicity, but they were not similar (7-fold) in dissolution half-time (*Fig. 10*c). However, the dissolution half-times of both materials is above 1 year (Table SI_1). If the dissolution half-times of both materials are considered to be beyond a biologically relevant range where solute release can no longer influence toxicity, any difference between them would not be considered relevant and dissolution should thus contribute to their assessment as being *similar* – this is easily achieved by restricting data to the biologically relevant range. The remaining Tier 1 criteria would therefore indicate sufficient similarity. Existing Tier 3 inhalation data can be used to challenge such an assessment, but here it indeed confirms the similarity (Keller et al., 2014).

The multidimensional comparison by Euclidean distance (Fig. 10b) scores the two Fe_2O_3 NFs as quite different, with similarity scores being comparable to either RTM. However, the multidimensional distance between the two NFs is driven by their difference in the dustiness, whereas their distance to the ZnO benchmark is driven by the more than 1000-fold dissolution half-time, and their difference to $BaSO_4$ benchmark is driven by reactivity and/or dissolution half-time. The relatively low multidimensional similarity score between the two Fe_2O_3 NFs (relative in comparison to the difference vs. the benchmark materials) persists after a restriction of all values of the data matrix to estimate the biologically relevant range, because the dustiness values are not restricted. This is a drawback of the multidimensional comparison in our particular implementation, e.g. in the safer-by-design purpose of grouping on the specific pair CeO_2 NM212 and CeO_2 NM211: the multidimensional distance conceals a case of high similarity in *some* of the descriptors (dimensions), which might justify a joint hazard assessment of both NFs, where the user would adapt only measures aimed at the different dustiness. With adjusted weighting of the descriptors based on more case studies, the high similarity would not be concealed.

Finally, one can again compare different algorithms (Fig. 8a,b; Figure SI_ 3). The similarity principle presumes the existence of a set of descriptors, such that molecules or NFs in the same local region (neighbourhood) of this descriptor space tend to have similar values of an endpoint. This is assumed to be the fundamental axiom of molecular similarity in descriptor space and is often called the neighbourhood principle or neighbourhood behaviour axiom (Patterson et al., 1996). The presence (or absence) of neighbourhood behaviour with respect to certain descriptors and properties may be revealed by examination of the plot of differences in descriptor values vs. differences in biological activities (Patterson et al., 1996). Each dot in Figure SI_ 3 represents a pair of NFs. Distances between descriptor values for a single descriptor are plotted on horizontal axis, while distances between property values are plotted on vertical axis. If a good neighbourhood behaviour holds, then the upper left triangle region (the region of activity cliffs (Cruz-Monteagudo et al., 2014)) would be empty, because there will be no large changes of the property due to small descriptor changes. These types of plots can be used to investigate if the neighbourhood assumption holds between different similarity assessment methods.

There may be future improvements where combinations of assays are validated to fulfil the data requirement of a specific decision node, e.g. because a single assay cannot screen for all possible mechanisms of action. A combination of cellular with an abiotic assay has been shown to achieve a slight improvement of predictivity compared to the single assays alone, especially when using surface metrics and when correcting for deposited dose(Bahl et al., 2020). Abiotic surface reactivity has the advantage of suspended test conditions with no concerns with accurately estimating the dose delivered to the target cells or tissues, but Ag-Seleci et al. in this special issue, demonstrate that the choice of a specific assay and concentration range profoundly influences the resulting similarity assessment (Ag-Seleci, 2021). It is beyond the scope of this White paper to recommend experimental methodologies, but the multidimensional evaluation of several assays may be a less defendable approach than the choice of one assay which is demonstrably sensitive and predictive for a specific class of NFs.

In summary, multidimensional distance metrics can offer unexpected insights into the overall similarity of very different materials, but data transformation and data standardisation must be performed for each property, and it is a major challenge to select a distance measure that is appropriate for all dimensions (i.e. all properties). When materials are identified as less similar, the user may need to additionally consider their ranking on the range spanned by the more toxic and less toxic RTMs, because the distance is always an absolute value which does not consider rankings. In summary, the multidimensional approaches are tools of discovery, but not recommended as routine tools for regulatory purposes.

5.3. Demonstration of multidimensional clustering tools on a real data matrix

Using different metrics, Fig. 11 presents the hierarchical clustering of all eight NFs and the same four properties as above (dustiness: respirable mass-based index; reactivity: FRAS mBOD; in vitro toxicity: LDH LOAEC; dissolution: lysosomal half-time). The Arsinh-OWA cluster (Fig. 11a) scores the quickly dissolving and toxic RTM ZnO NM110 as most different from all other NFs, and it also scores the partially dissolving and non-toxic RTM $BaSO_4$ NM220 as very different from all other NFs. The Arsinh-OWA cluster scores CeO_2 NM211, NM212 and Fe_2O_3 NF A in closely related branches, separate from CeO_2 Al-doped and Fe_2O_3 NF B.

The Minkowski distance cluster (Fig. 11b) also clearly separates the ZnO and $BaSO_4$ RTMs, and also scores the CeO_2 NFs in closely related branches, but finds more pronounced distance between the Fe_2O_3 NFs. The Euclidean distance cluster puts more weight (dark red and dark blue colours in Fig. 11c) on one property. The difference of the $BaSO_4$ and ZnO RTMs versus the Fe_2O_3 NF B and all other NFs in the *in vitro* toxicity (descriptor LDH LOAEC) divides the dendrogram, but all other NFs have limited difference.

Clustering analysis can be applied to all properties, methods, and all descriptors, as can be seen in Figure SI_ 4, where we expand the identical clustering tools from Fig. 11 to the data matrix in Table SI_ 1. Data in all cases were centred and scaled per descriptor onto values from −2 to 2, whereas missing values were not imputed (white cells). For each of the tested algorithms, the horizontal dendrograms in Figure SI_ 4 a,b,c are quite different when considering all available data *instead* of only the four descriptors selected by the IATA. Only one of the RTMs is always singled out, and the CeO_2 NFs now have a higher chance of ending up in the same branch. But the horizontal dendrograms are also very different *between* the algorithms, e.g. the relation of the Fe_2O_3 NFs to each other

and to the RTMs is not robust in Figure SI_ 4a,b,c. The distance metric considered for Figure SI_ 4c is the Euclidean distance, however other metrics such as the Manhattan distance or the Pearson's correlation coefficient could be alternatively considered.

The lack of robustness of the "all data multidimensional clustering" may be interpreted in favour of simpler, property-by-property similarity evaluations. It is important to note that the unsupervised clustering of "all data" contradicts the principles of grouping for regulatory purposes as introduced in sections 1 and 3.

The vertical dendrogram (correlating properties) has often been interpreted to prioritize descriptors for scientific purposes (Drew et al., 2017; Bahl et al., 2019). However, the vertical dendrogram changes completely between the Euclidean algorithm (Figure SI_ 4c) and the Minkowski algorithm (Figure SI_ 4b). In the Euclidean algorithm, size and dissolution rate are in one branch, and three measures of reactivity (carbonylation, ESR, FRAS) appear each as closely related. In the Minkowski algorithm (Figure SI_ 4b), reactivity by carbonylation, dissolution rate and reactivity by FRAS appear as closely related, and separate from each different branch of size and specific surface. The apparent lack of robustness needs to be considered when drawing conclusions on dendrograms.

In summary, relationships between NFs or between properties that have been discovered in the dendrograms of clustering approaches should not be taken as the basis for regulatory grouping of NFs of the same substance. Also for exploratory scientific purposes, the robustness should be challenged by carefully selecting the distance metric, and by comparing to other defendable distance metrics.

6. Outlook

6.1. Foreseeable methodical developments

Section 4 elaborated on robust descriptors, and on the use of data reduction. While this is well established, the advantage of comparing full distributions instead of descriptors is that one can account for the full shape of the distribution, including potential multimodality. The problems associated with comparison of data curves have been investigated in linear models, nonlinear models (Liu et al., 2009; Gsteiger et al., 2011; Liu et al., 2004) and in non-parametric models (Hall and Hart, 1990; Dette and Neumeyer, 2001). Normally, a simple regression model is set up to compare the effect of minimizing the sum-of-squares, with a traditional least squares fit (Liu et al., 2004). However model parameters can also be determined with a global optimization method; apart from minimizing the sum-of-squares metric, alternative methods include partial curve mapping (PCM), the area between two curves, the discrete Fréchet distance, the dynamic time warping (DTW) distance, and the curve length based similarity measures (Jekel et al., 2019). Interestingly Gsteiger et al. proposed a confidence band for the difference of two regression curves (Gsteiger et al., 2011), though their methodology can be easily expanded to non-parametric methods. Property distributions (e.g. size distributions) can be compared using well known statistical distances like Kullback–Leibler divergence, Hellinger distance, or Bhattacharyya distance.

Section 5 presented usable tools to assess similarity but is far from complete. Computational modelling such as clustering or dimensionality reduction (e.g. PCA) are suitable strategies to support grouping and its visualization. However, some methods require datasets without missing information. This can be challenging when applying such techniques to real-life datasets or even "historical" (literature) data. The challenge could be tackled either by applying methods which are robust to dealing with missing data and are able to work with sparse data matrices; or by filling in the missing values. The latter can be performed by e.g., listwise deletion of missing descriptors, calculation of missing values (e.g. wall number of MWCNT), expert judgment or by applying imputation modelling approaches. Imputation is a process of replacing missing data with substituted (estimated) values. Imputation can be performed by multiple statistical or machine learning methods, although a simple and commonly applied practice is to replace missing values for a specific descriptor with the arithmetic mean or the mode of the available data for that descriptor. Since the latter oversimplifies and perhaps crudely changes associations between descriptors, model-based approaches are preferred. Data imputation methods can be univariate or multivariate, two representative examples of the latter are the k-nearest neighbours imputation and the rich family of matrix completion algorithms (Keshavan et al., 2010).

Our demonstration in section 5 chose an assessment approach via a minimum number of most relevant properties with tiered measurement strategies given by the GRACIOUS IATAs (section 3), but contrary to this highly structured approach, one may explore much richer datasets. A quantitative approach that combines interspecies correlation analysis and self-organizing map analysis was developed (Sizochenko et al., 2018). The authors estimated patterns of toxicity among metal and silica oxide nanoparticles aiming to categorize in different groups and also account for missing values. Ha et al. presented a meta-analysis of 216 published articles on oxide nanoparticles and showed that mean imputation by applicability domain and physicochemical property-based scoring (Ha et al., 2018) improved their findings in terms of risk assessment. Furxhi et al. investigated the robustness of several machine learning methods on generated versions of the dataset by removing values artificially (Furxhi, 2018). Additionally, an integration of two imputation filling techniques to predict neurotoxicity for non-NFs was implemented by Pradeep et al., which demonstrated the capacity of integrating methodologies (Pradeep et al., 2019).

In view of the above trends, new tools beyond the ones implemented here can be expected, and will require validation on increasingly larger data sets.

6.2. Immediate next steps to validation

This special issue includes several case studies for which higher-tier data is available. The aim is to calibrate similarity assessments using lower-tier data by comparison to the higher-tier data, where the higher-tier data needs to indicate that NFs are sufficiently similar to be grouped. For the grouping hypothesis that addresses the environmental degradation of organic surface treatments, Cross et al. use the Tier 3 results in mesocosm studies, derived from a literature review, to calibrate the similarity limits of OECD test methods for assessing ready biodegradation that represent tier 2 of the GRACIOUS tiered testing strategy, and the similarity limits of Tier 1 screening methods (Cross, 2021b). For inhalation, Braakhuis et al. use Tier 3 inhalation literature for different silica NFs to calibrate the limits of a similarity assessment of Tier 1 data (Braakhuis, 2021), while Jeliazkova et al. do the same for several classes of organic pigments, across different substances (Jeliazkova, 2021). Peijnenburg et al. (to be submitted) assessed the similarity of uptake and effects in Tier 3 environmental model organisms (lettuce, bacteria, sediment oligochaete) to calibrate the limits of acceptable Tier 1 similarity of soluble NFs.

One cannot presuppose that such limits of acceptable Tier 1 similarity of a certain property in a certain IATA, derived from RTMs and case studies, will apply to all NFs in the same IATA. The calibrated limits will apply to NFs whose intrinsic toxicity (via ions and via surface) is within the range spanned by the RTMs. Also another instance of the same property in another IATA may require different limits of acceptable similarity, e.g. when lysosomal dissolution is a primary criterion in the inhalation and HARN IATA (Murphy et al., 2021; Keller et al., 2021; Braakhuis et al., 2016)(Braakhuis, 2021), but a subordinate criterion in the oral IATA (Di Cristo et al., 2021), where less strict limits might apply. The similarity of oral uptake and hazard is explored by another case study by di Christo et al. (Di Cristo, 2021).

Finally, the calibration for regulatory purposes results in limits that are too strict for the application of grouping concepts during early stages of industrial development, where SbD NFs can be identified by grouping

across different substances, and/or with less strict requirements on similarity, and/or by multidimensional similarity assessments. The GRACIOUS Framework explicitly targets such applications (Stone et al., 2020), and the demonstration blueprint of a GRACIOUS software e-tool allows for these options.(Traas and Vanhauten, 2021)

7. Conclusions

This White paper outlines and demonstrates a range of approaches to assess the similarity of different NFs, in order to support grouping and read-across approaches via the GRACIOUS Framework. These approaches included property-by-property assessments (equivalent to a single decision node of an IATA), as well as multidimensional evaluations (allowing simultaneous comparisons of different NFs across all properties derived from an IATA). For both types of assessment, the methods require data collection for all decision nodes in a data matrix. By default, the data matrix containing the NF values for the decision nodes is evaluated property-by-property, and NFs are discarded from the candidate group until a hypothesis is confirmed. Afterwards, the basic physicochemical properties of the confirmed NFs describe the boundaries of the group via the min and max values of composition, size, BET, aspect ratio, surface treatment. For the category approach, uncertainty is linked to uncertainty in the trend description, while for the analogue approach, uncertainty is linked to response variability within the acceptable range of similarity (Fig. 1).

We recommend that only data obtained by the same method (i.e. the same SOP or TG) must be applied to all of the NFs within one similarity assessment. Often such data is available as a distribution (e.g. a dose response curve of in vitro test data). Since such curves can be difficult to compare to assess similarity, they can be reduced to scalar descriptors (e.g. LC50 or LOAEL) to simplify the process. Representative Test Materials (RTM) help to identify the lower and higher values for a biological response and so assess the biologically relevant range. The biologically relevant range is typically narrower than the measurable range; in consequence, sensitive measurement beyond that range may provide information that is irrelevant for grouping decisions. Justified groups by the analogue approach must have a defined similarity that is within the measurable and biologically relevant range in each of the properties of the IATA. An exception would be if a regular pattern between a property and response in an endpoint is established for a category approach to read-across, in which case, the applicability domain of the group may extend further. Since the IATAs are designed to capture the key information required to test the associated hypothesis, the similarity assessment must be conducted for all decision nodes within the IATA – and only for these. The limits of acceptable similarity for properties or endpoints also need to consider method reproducibility. For example the reproducibility can include factors of less than 1.5-fold (another notation for 50% deviation) for properties that describe a NF (such as size, surface area), but rather factors of two- to three-fold for descriptors of surface reactivity, in vitro inflammation, and dissolution half-time. Standardised test methods (such as for size, surface area) may be associated with better reproducibility whilst non-standard methods (such as for reactivity) are required by many IATA decision nodes. Their standardisation is a priority.

One of the aims of grouping and read-across is to reduce the need for testing, especially when using animals. The use of tier 1 data to support grouping is possible, providing that the tier 1 (e.g. *in vitro*) method has been calibrated by comparison to available tier 3 (e.g. *in vivo*) results for one or more of the group members for the same endpoint. Case studies delivering such calibration are presented in the same special issue.

The GRACIOUS Framework foresees that the methods used to determine the data for the IATA decision nodes can be tiered, and after each tier, similarity assessment can support the decision to confirm grouping, to escalate to a higher tier, or to stop grouping. Also the ECETOC NanoApp (Janer et al., 2020; Janer et al., 2021) uses such a conditional escalation to higher tiers, but the NanoApp triggers *more properties in higher tier*, whereas the GRACIOUS approach triggers *more reliable methods in higher tier* for the same properties, and ultimately in vivo testing if required. The pairwise assessment of similarity after Tier 1 is consistent with the concept of "floating bands", but is different from all schemes of banding with predefined cut-offs. We described and explored two established and two novel algorithms to conduct a pairwise comparison property-by-property:

- The x-fold comparison as used in the ECETOC NanoApp.
- The novel Bayesian model assessment which compares two sets of values using nested sampling.
- The novel Arsinh-OWA model which applies the arsinh transformation to the distance between two NFs, and then rescales the result to the arsinh of a biologically relevant threshold.
- Euclidean distance which is the length of the line segment between two points, and is widely used and usually the first choice of distance metric.

Two of the algorithms are freely available as browser-based tool, (Enanomapper similarity tool, 2021) and the property-by-property x-fold algorithm is also embedded in the publicly available GRACIOUS Blueprint pdf.(Traas and Vanhauten, 2021) Based on the similarity scores of RTMs and orientating case studies, we concluded that the x-fold, Bayesian and Arsinh-OWA distance algorithms are mutually consistent in scoring NF pairs. The very popular Euclidean distance is also useful, but only with Yeo-Johnson data transformation, which enhances consistency with the other algorithms, albeit not perfectly. The Tier 1 score of a NF pair with known Tier 3 similarity can be indicatively set at or below 1.3 (Yeo-Johnson Euclidean) and at or above 1.5 (Bayesian). The x-fold metric does not standardize data, but has the advantage of being implemented without programming knowhow, and being easily compared to parameters such as experimental reproducibility (Cross, 2021a); for example, acceptable similarity can be indicatively set at or below 5-fold, whereas the comparison of opposite controls (i.e. the pair of representative test materials) scores between 100-fold to 1000-fold.

The similarity scores of a pair of biologically similar NFs that are given here for the different similarity algorithms are indicative only and need confirmation by more case studies. Beyond the present example, cases need to explore the decision nodes relevant for other environmental and human hazard endpoints, and ideally each case uses known Tier 3 similarity to identify NF pairs with acceptable similarity and NFs pairs with borderline similarity, which then define limits of acceptable similarity in the respective Tier 1 methods of the same IATA decision nodes.

A range of multidimensional evaluations, for example dendrogram clustering approaches identify relationships between NF properties, and were also explored. Multidimensional distance metrics were found to offer unexpected insights into the overall similarity of very different materials, but it is a major challenge to select a distance metric that is appropriate for all dimensions (i.e. all properties), and inappropriate data transformation can lead to false conclusions. The multidimensional tools are therefore difficult to use in a regulatory context. If materials are identified as less similar when using these methods, the user may need to additionally consider their ranking in individual properties, because rankings are not represented by distances, but may be important to justify read-across. When used for exploratory scientific purposes, the robustness should be challenged by carefully selecting the distance metric, and by comparing to other defendable distance metrics. The multidimensional approaches are not generally recommended for regulatory purposes, instead they are primarily tools of discovery.

In conclusion, for regulatory purposes, a property-by-property evaluation of the data matrix is recommended to substantiate grouping. This means that for one grouping hypothesis, all decision nodes of the associated IATA need to be assessed individually using the property-by-property evaluation. If any NF for any decision node is not

found to be sufficiently similar for that property, then that NF should be considered for removal from the group. If for any decision node there is no evidence of similarity between the NFs then the whole hypothesis should be rejected.

Interestingly, even in the property-by-property evaluation, the same method and algorithm are always applied across different substances, because both the NFs of the candidate group and the RTMs, hence overall materials of three substances, are included in the assessment. Although the current regulatory guidance limits grouping and read-across to materials of the same substance, the similarity tools are thus applicable beyond this limitation. The same tools can support SbD decisions during industrial development of innovative NFs, where comparison between NFs of different substances is often required.

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

The GRACIOUS project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 760840. We sincerely thank Frank Le Curieux and colleagues at ECHA for detailed comments on an earlier version of the manuscript. We acknowledge critical reading, comments and discussion by the entire similarity team of the GRACIOUS consortium.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.impact.2021.100366.

References

Ag-Seleci, D., et al., 2021. Determining nanoform similarity via assessment of surface reactivity by abiotic and in vitro assays. NanoImpact under review.
Arts, J.H., et al., 2015. A decision-making framework for the grouping and testing of nanomaterials (DF4nanoGrouping). Regul. Toxicol. Pharmacol. 71 (2), S1–S27.
Bacsa, R.R., Kiwi, J., 1998. Effect of rutile phase on the photocatalytic properties of nanocrystalline titania during the degradation of p-coumaric acid. Appl. Catal. B Environ. 16 (1), 19–29.
Bahl, A., et al., 2019. Recursive feature elimination in random forest classification supports nanomaterial grouping. NanoImpact 15, 100179.
Bahl, A., et al., 2020. Nanomaterial categorization by surface reactivity: a case study comparing 35 materials with four different test methods. NanoImpact 19, 100234.
Bajorath, J., 2017. In: Keith, J.M. (Ed.), Molecular Similarity Concepts for Informatics Applications BT - Bioinformatics: Volume II: Structure, Function, and Applications. Springer New York, New York, NY, pp. 231–245.
BIPM, 2008. Evaluation of Measurement Data — Guide to the Expression of Uncertainty in Measurement.
Braakhuis, H., et al., 2021. Similarity assessment of silicas to evaluate the GRACIOUS grouping approach. NanoImpact under review.
Braakhuis, H.M., Oomen, A.G., Cassee, F.R., 2016. Grouping nanomaterials to predict their potential to induce pulmonary inflammation. Toxicol. Appl. Pharmacol. 299, 3–7.
Brown, K.G., Erdreich, L.S., 1989. Statistical uncertainty in the no-observed-adverse-effect level. Fund. Appl. Toxicol.: Off. J. Soc. Toxicol. 13 (2), 235–244.
Cai, X., et al., 2018. Multi-hierarchical profiling the structure-activity relationships of engineered nanomaterials at nano-bio interfaces. Nat. Commun. 9 (1), 4416.
Clément, L., Hurel, C., Marmier, N., 2013. Toxicity of TiO2 nanoparticles to cladocerans, algae, rotifers and plants – Effects of size and crystalline structure. Chemosphere 90 (3), 1083–1090.
Commission, E, 2018. In: E. Commission (Ed.), Commission Regulation (EU) 2018/1881 of 3 December 2018 amending Regulation (EC) No 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) as regards Annexes I, III,VI, VII, VIII, IX, X, XI, and XII to address nanoforms of substances (Text with EEA relevance.). C/2018/7942.
Cross, R., 2021a. Reproducibility of methods required to identify and characterize nanoforms of substances. NanoImpact under review.
Cross, R., 2021b. Similarity of nanoforms with different organic surface treatments based on coating material biodegradation. NanoImpact under review.
Crump, K., 1984. A new method for determining allowable daily intakes*1. Fundam. Appl. Toxicol. 4 (5), 854–871.
Cruz-Monteagudo, M., et al., 2014. Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? Drug Discov. Today 19 (8), 1069–1080.
Dasari, T.P., Pathakoti, K., Hwang, H.-M., 2013. Determination of the mechanism of photoinduced toxicity of selected metal oxide nanoparticles (ZnO, CuO, Co3O4 and TiO2) to E. coli bacteria. J. Environ. Sci. 25 (5), 882–888.
Dette, H., Neumeyer, N., 2001. Nonparametric analysis of covariance. Ann. Stat. 29 (5), 1361–1400.
Di Cristo, L., et al., 2021. Grouping of orally ingested silica nanomaterials via use of an Integrated Approach to Testing and Assessment to streamline risk assessment. NanoImpact under review.
Di Cristo, L., et al., 2021. Grouping hypotheses and an integrated approach to testing and assessment of nanomaterials following oral ingestion. Nanomaterials 11 (10), 2623.
Ding, H., et al., 2008. Querying and mining of time series data: experimental comparison of representations and distance measures. Proceed. VLDB Endow. 1 (2), 1542–1552.
Drew, N.M., et al., 2017. A quantitative framework to group nanoscale and microscale particles by hazard potency to derive occupational exposure limits: proof of concept evaluation. Regul. Toxicol. Pharmacol. 89 (Supplement C), 253–267.
ECHA, 2017. Read-Across Assessment Framework (RAAF). Helsinki.
ECHA, 2019a. Appendix R.6-1 for Nanoforms Applicable to the Guidance on QSARs and Grouping of Chemicals. Helsinki.
ECHA, 2019b. Appendix for nanoforms applicable to the Guidance on Registration and Substance Identification. Helsinki.
Elliott, J.T., et al., 2017. Toward achieving harmonization in a nanocytotoxicity assay measurement through an interlaboratory comparison study. ALTEX-Alternat. Animal Exper. 34 (2), 201–218.
Enanomapper similarity tool, 2021. https://search.data.enanomapper.net/projects/gracious/similarity. (Accessed 30 November 2021).
European, P., C. The, 2006. Regulation (EC) No 1907/2006 on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) {Brussels}.
European_Chemicals_Agency_(ECHA), 2019. Appendix for nanoforms applicable to the Guidance on Registration and Substance Identification. In: ECHA (Ed.), ECHA-19-H-14-EN.
Floris, M., Olla, S., 2018. Molecular similarity in computational toxicology. Methods Mol. Biol. 1800, 171–179 (Clifton, N.J.).
Furxhi, I., et al., 2018. Predicting Nanomaterials Toxicity Pathways based on Genome-wide Transcriptomics Studies using Bayesian Networks. IEEE 18th International Conference on Nanotechnology (IEEE-NANO).
Furxhi, I., et al., 2019. Application of Bayesian networks in determining nanoparticle-induced cellular outcomes using transcriptomics. Nanotoxicology 13 (6), 827–848.
Giusti, A., et al., 2019. Nanomaterial grouping: Existing approaches and future recommendations. NanoImpact 16, 100182.
Gottardo, S., et al., 2017. NANoREG Framework for the Safety Assessment of Nanomaterials. Publications Office of the European Union, Luxembourg.
Gsteiger, S., Bretz, F., Liu, W., 2011. Simultaneous confidence bands for nonlinear regression models with application to population pharmacokinetic analyses. J. Biopharm. Stat. 21 (4), 708–725.
Guney, E., 2017. Reproducible Drug Repurposing: When Similarity Does not Suffice (Pacific Symposium on Biocomputing), 22, pp. 132–143.
Ha, M.K., et al., 2018. Toxicity classification of oxide nanomaterials: effects of data gap filling and pchem score-based screening approaches. Sci. Rep. 8 (1), 3141.
Hall, P., Hart, J.D., 1990. Bootstrap test for difference between means in nonparametric regression. J. Am. Stat. Assoc. 85 (412), 1039–1049.
Hardy, A., et al., 2018. Guidance on risk assessment of the application of nanoscience and nanotechnologies in the food and feed chain: Part 1, human and animal health. EFSA J. 16 (7).
Hund-Rinke, K., et al., 2018. Grouping concept for metal and metal oxide nanomaterials with regard to their ecotoxicological effects on algae, daphnids and fish embryos. NanoImpact 9, 52–60.
Janer, G., Landsiedel, R., Wohlleben, W., 2020. Rationale and decision rules behind the ECETOC NanoApp to support registration of sets of similar nanoforms within REACH. Nanotoxicology 1–22.
IARC, 2002. Man-Made Vitreous Fibres. World Health Organization, pp. 1–433.
Janer, G., et al., 2021. Creating sets of similar nanoforms with the ECETOC NanoApp: real-life case studies. Nanotoxicology 1–19.
Jekel, C.F., et al., 2019. Similarity measures for identifying material parameters from hysteresis loops using inverse analysis. Int. J. Mater. Form. 12 (3), 355–378.
Jeliazkova, N., 2021. Possibilities to group nanomaterials across different substances – A case study on organic pigments. NanoImpact under review.
Karkossa, I., et al., 2019. An in-depth multi-omics analysis in RLE-6TN rat alveolar epithelial cells allows for nanomaterial categorization. Particle Fibre Toxicol. 16 (1), 38.
Kass, R.E., Raftery, A.E., 1995. Bayes Factors. J. Am. Stat. Assoc. 90 (430), 773–795.
Keller, J., et al., 2014. Time course of lung retention and toxicity of inhaled particles: short-term exposure to nano-Ceria. Arch. Toxicol. 88 (11), 2033–2059.
Keller, J., et al., 2021. Variation in dissolution behavior among different nanoforms and its implication for grouping approaches in inhalation toxicity. NanoImpact 23, 100341.
Keller, J.G., et al., 2020. Dosimetry in vitro – exploring the sensitivity of deposited dose predictions vs. affinity, polydispersity, freeze-thawing, and analytical methods. Nanotoxicology 1–14.
Keshavan, R.H., Montanari, A., Oh, S., 2010. Matrix completion from a few entries. IEEE Trans. Inf. Theory 56 (6), 2980–2998.
Khataee, A.R., Aleboyeh, H., Aleboyeh, A., 2009. Crystallite phase-controlled preparation, characterisation and photocatalytic properties of titanium dioxide nanoparticles. J. Exp. Nanosci. 4 (2), 121–137.
Kochev, N., Monev, V., Bangov, I., 2003. Searching Chemical Structures, pp. 291–318.

Koltermann-Jülly, J., et al., 2019. Addendum to "Abiotic dissolution rates of 24 (nano) forms of 6 substances compared to macrophage-assisted dissolution and in vivo pulmonary clearance: grouping by biodissolution and transformation"[NanoImpact 12 (2018) 29–41]. NanoImpact 14, 100154.
Krug, H.F., 2018. The uncertainty with nanosafety: Validity and reliability of published data. Colloids Surf. B: Biointerfaces 172, 113–117.
Kühnel, D., et al., 2019. Closing gaps for environmental risk screening of engineered nanomaterials. NanoImpact 15, 100173.
Lamon, L., et al., 2019. Grouping of nanomaterials to read-across hazard endpoints: a review. Nanotoxicology 13 (1), 100–118.
Liu, R., et al., 2015. Analysis of soil bacteria susceptibility to manufactured nanoparticles via data visualization. Beilstein J. Nanotechnol. 6 (1), 1635–1651.
Liu, W., Jamshidian, M., Zhang, Y., 2004. Multiple comparison of several linear regression models. J. Am. Stat. Assoc. 99 (466), 395–403.
Liu, W., et al., 2009. Assessing nonsuperiority, noninferiority, or equivalence when comparing two regression models over a restricted covariate region. Biometrics 65 (4), 1279–1287.
Marvin, H.J.P., et al., 2017. Application of Bayesian networks for hazard ranking of nanomaterials to support human health risk assessment. Nanotoxicology 11 (1), 123–133.
Meesters, J.A.J., et al., 2014. Multimedia modeling of engineered nanoparticles with simplebox4nano: model definition and evaluation. Environ. Sci. Technol. 48 (10), 5726–5736.
Meesters, J.A.J., et al., 2019. A model sensitivity analysis to determine the most important physicochemical properties driving environmental fate and exposure of engineered nanoparticles. Environ. Sci.: Nano 6 (7), 2049–2060.
Mellor, C.L., et al., 2019. Molecular fingerprint-derived similarity measures for toxicological read-across: Recommendations for optimal use. Regul. Toxicol. Pharmacol. 101, 121–134.
Murphy, F., et al., 2021. An integrated approach to testing and assessment of high aspect ratio nanomaterials and its application for grouping based on a common mesothelioma hazard. NanoImpact 22, 100314.
Nikota, J., et al., 2015. Meta-analysis of transcriptomic responses as a means to identify pulmonary disease outcomes for engineered nanomaterials. Particle Fibre Toxicol. 13 (1), 1–18.
Nymark, P., et al., 2020. Toward rigorous materials production: new approach methodologies have extensive potential to improve current safety assessment practices. Small 16 (6), 1904749.
Oberdörster, G., 2000. Determinants of the pathogenicity of man-made vitreous fibers (MMVF). Int. Arch. Occup. Environ. Health 73 (1), S60–S68.
OECD, 2014. Series on Testing and Assessment, No. 194. Guidance on Grouping of Chemicals, second edition. Paris.
OECD, 2017. Guidance Document on the Reporting of Defined Approaches to be Used Within Integrated Approaches to Testing and Assessment, OECD Series on Testing and Assessment. OECD Publishing. No. 255.
OECD, 2020. Series on the Safety of Manufactured Nanomaterials, No. 96. Moving Towards a Safe(r) Innovation Approach (SIA) for More Sustainable Nanomaterials and Nano-enabled Product. ENV/JM/MONO(2020)36/REV1. Paris, France.
Park, M.V., et al., 2018. Development of a systematic method to assess similarity between nanomaterials for human hazard evaluation purposes - lessons learnt. Nanotoxicology 12 (7), 652–676.
Patlewicz, G., et al., 2017. Navigating through the minefield of read-across tools: a review of in silico tools for grouping. Comput. Toxicol. 3, 1–18.
Patterson, D.E., et al., 1996. Neighborhood behavior: A useful concept for validation of "Molecular Diversity" descriptors. J. Med. Chem. 39 (16), 3049–3059.
Pradeep, P., et al., 2019. Integrating data gap filling techniques: a case study predicting TEFs for neurotoxicity TEQs to facilitate the hazard assessment of polychlorinated biphenyls. Regul. Oxicol. Pharmacol.: RTP 101, 12–23.
Praetorius, A., et al., 2020. Strategies for determining heteroaggregation attachment efficiencies of engineered nanoparticles in aquatic environments. Environ. Sci.: Nano 7 (2), 351–367.
Roebben, G., et al., 2013. Reference materials and representative test materials: the nanotechnology case. J. Nanopart. Res. 15 (3), 1455.
Sakoe, H., Chiba, S., 1978. Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust. Speech Signal Process. 26 (1), 43–49.
Salieri, B., et al., 2019. Fate modelling of nanoparticle releases in LCA: an integrative approach towards "USEtox4Nano". J. Clean. Prod. 206, 701–712.
Shaw, S.Y., et al., 2008. Perturbational profiling of nanomaterial biologic activity. Proceed. National Acad. Sci. 105 (21), 7387–7392.
Sheridan, R.P., Kearsley, S.K., 2002. Why do we need so many chemical similarity search methods? Drug Discov. Today 7 (17), 903–911.
Sizochenko, N., et al., 2018. How the toxicity of nanomaterials towards different species could be simultaneously evaluated: a novel multi-nano-read-across approach. Nanoscale 10 (2), 582–591.
Skilling, J., 2006. Nested sampling for general Bayesian computation. Bayesian Anal. 1 (4), 833–859.
Sørensen, S.N., et al., 2020. Comparison of species sensitivity distribution modeling approaches for environmental risk assessment of nanomaterials – A case study for silver and titanium dioxide representative materials. Aquat. Toxicol. 225, 105543.
Stone, V., et al., 2020. A framework for grouping and read-across of nanomaterials-supporting innovation and risk assessment. Nano Today 35, 100941.
Svendsen, C., et al., 2020. Key principles and operational practices for improved nanotechnology environmental exposure assessment. Nat. Nanotechnol. 15 (9), 731–742.
Tsiliki, G., 2021. Bayesian based grouping of nanomaterials and Dose Response similarity models. NanoImpact under review.
Traas, Lion, Vanhauten, Ralph, 2021. GRACIOUS framework blueprint. zenodo. https://doi.org/10.5281/zenodo.5497615.
Tsiliki, G., et al., 2017. Enriching nanomaterials omics data: an integration technique to generate biological descriptors. Small Methods 1 (11), 1700139.
Wassenaar, P.N.H., et al., 2021. Evaluating chemical similarity as a measure to identify potential substances of very high concern. Regul. Toxicol. Pharmacol. 119, 104834.
Willett, P., Barnard, J.M., Downs, G.M., 1998. Chemical similarity searching. J. Chem. Inf. Comput. Sci. 38 (6), 983–996.
Wohlleben, W., et al., 2019. The nanoGRAVUR framework to group (nano) materials for their occupational, consumer, environmental risks based on a harmonized set of material properties, applied to 34 case studies. Nanoscale 11 (38), 17637–17654.
Worth, A., et al., 2017. Evaluation of the Availability and Applicability of Computational Approaches in the Safety Assessment of Nanomaterials. Publications Office of the European Union.
Yager, R.R., 1996. Quantifier guided aggregation using OWA operators. Int. J. Intell. Syst. 11 (1), 49–73.
Yeo, I.K., Johnson, R.A., 2000. A new family of power transformations to improve normality or symmetry. Biometrika 87 (4), 954–959.
Zabeo, A., 2021. Ordered weighted average based grouping of nanomaterials with arsinh and dose response similarity models. NanoImpact 100370.