Grouping of advanced multi-component nanomaterials: can machine learning help?

Georgia Tsiliki[1], Alex Zabeo[2], Vicki Stone[3], Danail Hristozov[2]

1. Introduction

Multi-component nanomaterials (MCNMs) can be manufactured featuring many different physico-chemical properties. Unlike mono-component nanomaterials (NM), MCNMs consist of multiple components that may interact with one another, and for that reason their safety profile should be clearly characterised. Traditionally evaluating MCNMs is expensive and time-consuming. Similarity assessment methodologies specifically designed for MCNMs could be valuable tools to justify that existing safety-related information can be re-used between group members, thereby reducing the need to generate new hazard data and improve their sustainability. In the case of MCNMs, it is expected to evaluate the differences and similarities to their mono-components NMs aiming to assess their functionality and potential hazard.

We suggest a two-step grouping approach for identifying similar MCNMs, which allows researchers to quantify similarities between nanomaterials and then use this information to read-across safety information from well-studied materials to new ones. The two procedures can be used as standalone approaches for grouping and read-across. Depending on the quality and the quantity of the data available, the method can adjust to various scenarios using both scalar and dose-response data. The methodology is applied to two use cases from industry to demonstrate the method's effectiveness.

2. Methods

A sequential workflow is presented to identify groups in the data and read-across unknown hazard or toxicity endpoint values. The method first normalizes intrinsic and extrinsic properties, calculates MCNMs similarity scores for each of the properties, and finally suggest a unique grouping ranking across data sets. The output is then used for reading across toxicity endpoints. Specifically, Gaussian mixture modelling is employed to describe parameter- and MCNM-specific distributions and compare them in a probabilistic manner. The agility of the method is that it can adopt to any format or range of the raw data by breaking the input data into distinct components and then combine them to describe a data-specific mixture distribution. The method primarily calculates pairwise comparisons of interest to allow researchers to focus on specific parameters of interest and evaluate their biological relevance. To determine the mixture model, we initially consider an arbitrary set of *k* Gaussian distributions and calculate the mean vector and covariance matrices from all the data points. An iterative process using an

[1] Purposeful IKE, 144 Tritis Septemvriou, 11251 Athens, Greece; georgia@purposeful.eu
[2] Greendecision srl, Canareggio 5904, 30121 Venice, Italy
[3] Institute of Biological Chemistry, Biophysics and Bioengineering, Heriot-Watt University, Edinburgh, UK

Expectation-Maximisation (EM) algorithm is used to find the optimal *k* and corresponding distributions' parameters. Preliminary work on pairwise similarity analysis is used as the basis of model comparison to identify the optimal *k* groups in the data (Tsiliki et al., 2022). Additionally, a clustering topology is included in the analysis to visualize the groups of similar MCNMs formed.

We demonstrate our method's performance comparatively to hierarchical agglomeration clustering and to k Nearest Neighbours (kNN) algorithm. Applications to use case data following integrated approaches to testing and assessment (IATA) for inhalation exposure are shown (Tsiliki et al., 2024).

The workflow includes a final step for data gap filling for toxicity or hazard endpoints using the grouping results. The data gaps are filled individually using a weighted average between the neighbours' mean value and the class mean value. To help the user judgement for the acceptable uncertainty for read-across regulatory purposes, we enforced the algorithm with an uncertainty estimate.

3. Conclusions

Toxicological properties of a material are often driven by their physico-chemical properties. This principle is the basis for many existing in silico predictive models. However, for grouping advanced materials, more information should be collected and analysed for a robust similarity assessment and a grouping decision. We propose a similarity assessment method and grouping approach that can be applied to one- and two-dimensional data and identifies common groups across data sets accompanied with an uncertainty score for ease of decision making in terms of regulatory purposes. Although the suggested method is less computationally intensive compared to most Machine Learning (ML) algorithms (e.g. Support Vector Machine) and it is easy to implement, it is highly influenced by the size of the available data. When rich toxicity data or high-quality biological activity data are available, more consistent and accurate will be produced.

The workflow is demonstrated on MCNMs, however the method can be extended to any advanced materials reducing the burden of testing the safety of chemicals and therefore contributing to improving their sustainability.

4. References

1. Tsiliki G., Zabeo A., Di Battista V., Hristozov D., Stone V. (2024) Similarity of Multicomponent Nanomaterials: a computational justification, *Under review in Environmental Science: Nano*.
2. Tsiliki G., Ag Seleci D., Zabeo A., Basei G., Hristozov D., Jeliazkova N., Boyles M., Murphy F., Peijnenburg W., Wohlleben W., Stone V. (2022), Bayesian based similarity assessment of nanomaterials to inform grouping, *NanoImpact*, 25. https://doi.org/10.1016/j.impact.2022.100389.