Apellis: an online tool for read-across model development

Dimitra-Danai Varsou and Haralambos Sarimveis
School of Chemical Engineering, National Technical University of Athens, 157 80 Athens, Greece
dimitra.varsou@gmail.com, hsarimv@central.ntua.gr

November 5, 2020

Abstract

In this study, we present Apellis, a web application developed for training grouping/ read-across models for the prediction of engineered nanomaterials (ENMs) toxicity-related endpoints. The application applies a generic and novel computational workflow for estimating the endpoint of interest, which can be either categorical or numerical. During the training procedure, the application selects the most important ENM properties of concern that affect their toxic behaviour. In the process of grouping ENMs for performing read-across predictions, the multi-perspective characterization of ENMs can be taken into account, by defining more than one similarity criteria. The workflow converges to the grouping hypothesis that leads to the most accurate read-across estimations. Visualisation tools are included in the application, which offer better and more clear understanding of grouping and similarities among ENMs. The trained models can be saved in an electronic format, so that they can be easily retrieved, for calculating new predictions. In addition, this allows model developers to disseminate and share the produced models with the community. Apellis is free to use and accessible at `apellis.jaqpot.org`.

Keywords: Engineered nanomaterials, nanoinformatics, read-across, grouping

1 Introduction

The use of nanoinformatics for the *in silico* assessment of adverse effects of engineered nanomaterials (ENMs) as an alternative option to experimental procedures, has already been established. The extensive development of computational tools, has been guided by the REACH Regulation, which promotes the development of *in silico* techniques, as surrogates to animal testing. [1]

Nanoinformatics offers a plethora of software methods and tools that are based either on data-driven approaches or on dynamics simulations for nanotoxicity assessment. [2] The

computational approaches include QSAR-type (Quantitative Structure-Activity Relationship) models for the prediction of ENMs adverse effects. However questions are raised about the reliability of the produced predictions, due to the absence of sufficiently "large" datasets. [3]

Grouping and read-across techniques are alternative methods for data-gap filling, that can be applied for nanotoxicity prediction, even when only small datasets are available. [4] There are two main read-across approaches, namely the category and the analogue approach. In the category approach, the ENMs are treated as part of a group. The properties of all ENMs belonging to a specific group are similar or follow a regular trend. In the analogue approach, the toxicity endpoint of an untested ENM is estimated from the endpoint value of "neighbouring" tested ENMs. The grouping or neighbourhood boundaries in the two approaches are defined, by quantifying the similarity or dissimilarity among ENMs.

In order to harmonise the emerging practices and workflows for grouping and read-across, ECHA has provided guidelines for producing and justifying robust data-gap filling of toxicity endpoints during hazard assessment of ENMs. [5] These guidelines are based on the definition of a grouping hypothesis that is tested in terms of its ability to produce reliable predictions. However, the guidelines do not include a systematic method for searching for a successful hypothesis. The researcher may try many different hypotheses and select the one that produces the most accurate and robust predictions. This is a time consuming trial and error procedure, which may not converge to the optimal solution.

In order to overcome this limitation, in a previous study, Varsou *et al.* (2019) [6] presented a systematic workflow, which was able to automate the grouping definition process by selecting from the space of alternative hypotheses, the one that produces the most reliable predictions. The method was based on the formulation of a rigorous mixed integer non-linear mathematical optimisation problem (MINLP) and was designed to include one or multiple similarity criteria corresponding to different types of ENM descriptors, like physicochemical or biological descriptors.

In this work, we are presenting an extension of the read-across workflow proposed by Varsou *et al.* (2019), which includes an adjustment of the objective function (OF) with the addition of a regularisation parameter and the adaptation of the methodology for the prediction of categorical endpoints. The methodology is offered to the scientific community as a free web-tool, named Apellis (`apellis.jaqpot.org`), which is also presented in this work. Apellis allows training of predictive models with numerical or categorical toxicity endpoints, and generation of predictions for untested ENMs. It includes tools for data visualisation and for exporting, saving and sharing the produced models and results. The intuitive and friendly graphical user interface (GUI) allows potential users in the nanosafety community to apply the tool, even if they do not have deep knowledge of the background methodology.

2 Methods

In the present work, the process of developing a grouping/read-across model for predicting a toxicity-related numerical or categorical endpoint, is automated using a genetic algorithm

(GA) scheme which has a twofold objective: the removal of noisy descriptors (variable selection) and the definition of a neighbourhood of similar ENMs around each query ENM using universal thresholds.

The GA scheme assumes that a dataset with known descriptors and toxicity endpoint values is available. For validation purposes, the dataset is partitioned into training and test sets. The training set is used in the optimisation process, whereas the test set is used to assess the performance of the final model.

A subset of the available descriptors along with the thresholds, encoded in a hybrid array of binary and real values (*genes*), defines precisely a potential solution of the optimisation problem (*chromosome*). Different combinations of variables and thresholds are encoded in different *chromosomes* that comprise a *population*. The *population* evolves through a number of cycles (*generations*) of "biological" operations (*selection*, *crossover* and *mutation*) between the potential solutions, leading to an optimal solution (*genome*). All the biological operations are controlled by user-defined probability values.

Each *chromosome* in every level of the evolutionary process is tested for read-across prediction (Fig. 1) and, depending on the accuracy and robustness of the generated predictions, a score number is assigned to it. When all *chromosomes* of a *population* are scored, the ones with the highest scores are selected in pairs and are combined in order to exchange *genes* in random *crossover* points. The two new *chromosomes* are subject to *mutation*, where the *gene* values are altered according to a uniform or non-uniform scenario. The above process of *selection*, *crossover* and *mutation* is repeated until a new *population* is created. The process ensures that if a *chromosome* of the old *population* has higher score than all the *chromosomes* of the new *population*, the *chromosome* with the highest score is included in the new *population* (*elitism*). It also ensures that at least one variable will be selected, and that the combination of variables and threshold(s) will produce predictions for at least a predefined number of training samples (*predFactor*).

In the next paragraphs we describe the steps that lead from a *chromosome* to the calculation of the read-across predictions and the assessment of their quality. The steps are also presented schematically in Fig. 1. This workflow is applied to every *chromosome* of the *populations* of all *generations*, as well as to the final *genome*. As discussed before, the neighbouring selection and the read-cross predictions can be performed using more than one similarity criteria. For brevity the approach is presented only for one similarity criterion, but can be easily extended to two or more similarity criteria.

2.1 Variable selection

As described before, the key element in the genetic algorithms is the *chromosome* (Table 1). *Chromosomes* contain *genes* in a specific order. Each *gene* except the last one corresponds to a specific descriptor and takes a binary value that encodes the selection or not of the corresponding descriptor to the prediction workflow. Through the *generations*, *populations* of *chromosomes* are subject to the biological operators of *selection*, *crossover*, *mutation* and *elitism*, leading to the *genome* which contains the optimal combination of selected descriptors to produce reliable predictions. The total number of selected variables is controlled by a

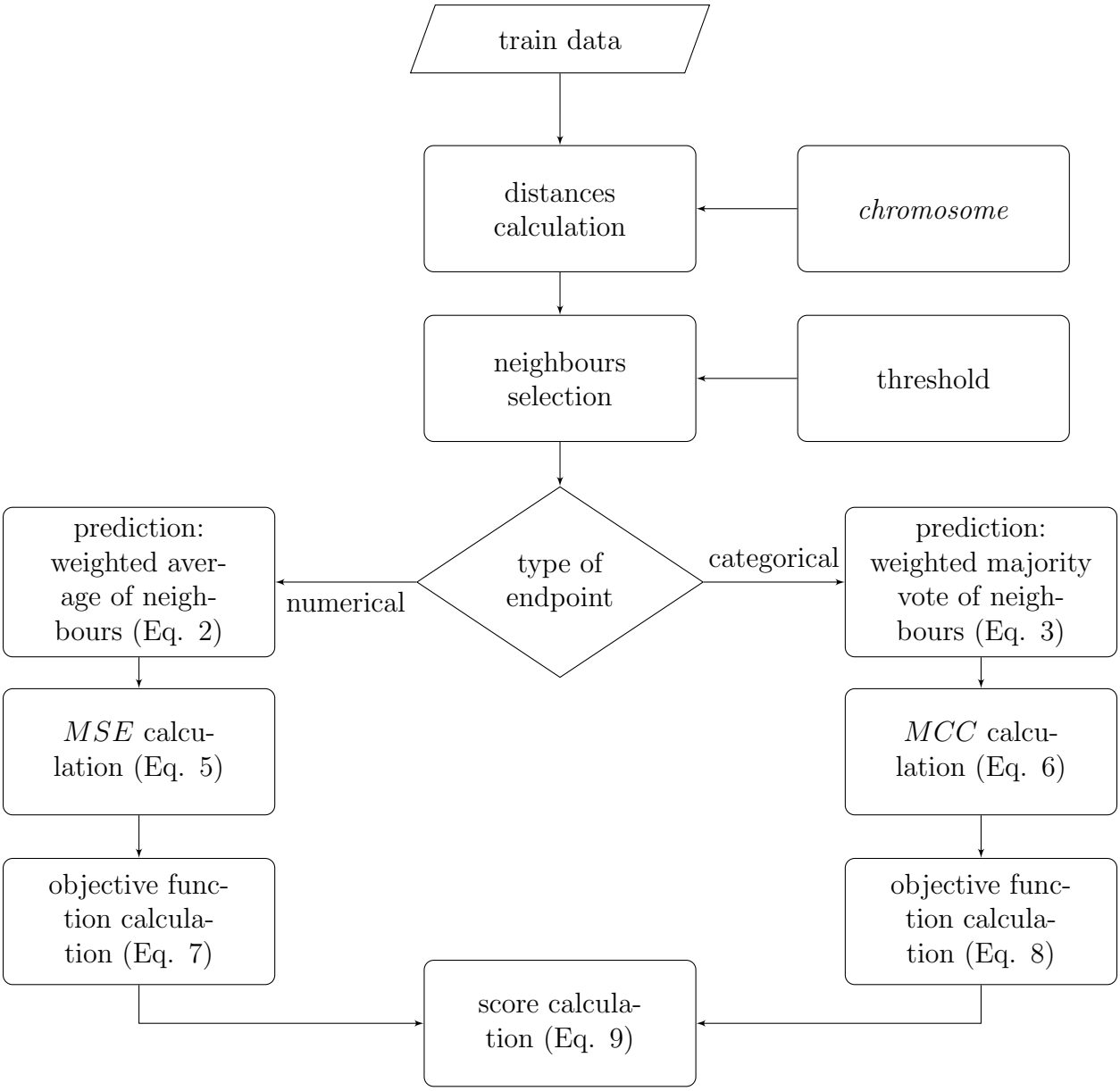

Figure 1: The main steps of a *chromosome*'s evaluation process during training.

Table 1: Examplary chromosome.

variable 1	variable 2	variable 3	variable 4	variable 5	variable 6	variable 7	threshold
0	1	0	1	0	1	1	0.05

regularisation factor. *Mutation* for binary *genes* is performed by inverting the value of the selected *genes*: 0 becomes 1 and vice versa (uniform *mutation*).

2.2 Definition of neighbours

In order to define the neighbours of a query ENM (i), the Euclidean distances between the query and all training ENMs is calculated, considering only the selected descriptors of a particular *chromosome*. Next, from the pool of j training ENMs, the ones with distance $dist_{i,j}$ equal or lower than the threshold value - which is the last element of the *chromosome* - are selected as the neighbours of the query ENM (Fig. 2). In this case the binary variables $neib_{i,j}$ take the value of 1, otherwise they take the value of 0. During training when $i = j$, the value of $neib_{i,j}$ is automatically set equal to zero. [6]

When different categories of descriptors are available (e.g. physicochemical, image, theoretical, bioinformatics descriptors), it is possible to use more than one similarity criteria (thresholds) for the selection of neighbours. In this case, Euclidean distances between ENMs are calculated independently for each type of descriptors (considering only the selected ones indicated by the *chromosome*). Two ENMs are considered similar, only if all distances are below the respective thresholds. [6, 7]

Threshold values are also subject to genetic operators and converge to their optimal value during the evolutionary process, as parts of the *chromosomes*. Unlike the rest of the *genes*, thresholds are continuous and the non-uniform *mutation* operator described in (Eq. 1) is applied:

$$thr_{new}^{GA} = \begin{cases} thr_{old}^{GA} + (thr_{max}^{GA} - thr_{old}^{GA}) \cdot (1 - r^{(1-g/generations)bGA}) \\ \text{if a random digit is 0} \\ \\ thr_{old}^{GA} - (thr_{old}^{GA} - thr_{min}^{GA}) \cdot (1 - r^{(1-g/generations)bGA}) \\ \text{if a random digit is 1} \end{cases} \quad (1)$$

where, thr_{old}^{GA} is the old threshold value,
thr_{new}^{GA} the threshold value that results from the non-uniform *mutation*,
thr_{max}^{GA} and thr_{min}^{GA} are the upper and the lower bounds of the threshold values,
r is a random number between 0 and 1,
g is the number of the current *generation* and,
bGA is a parameter which determines the degree of dependency on the *generation* number.

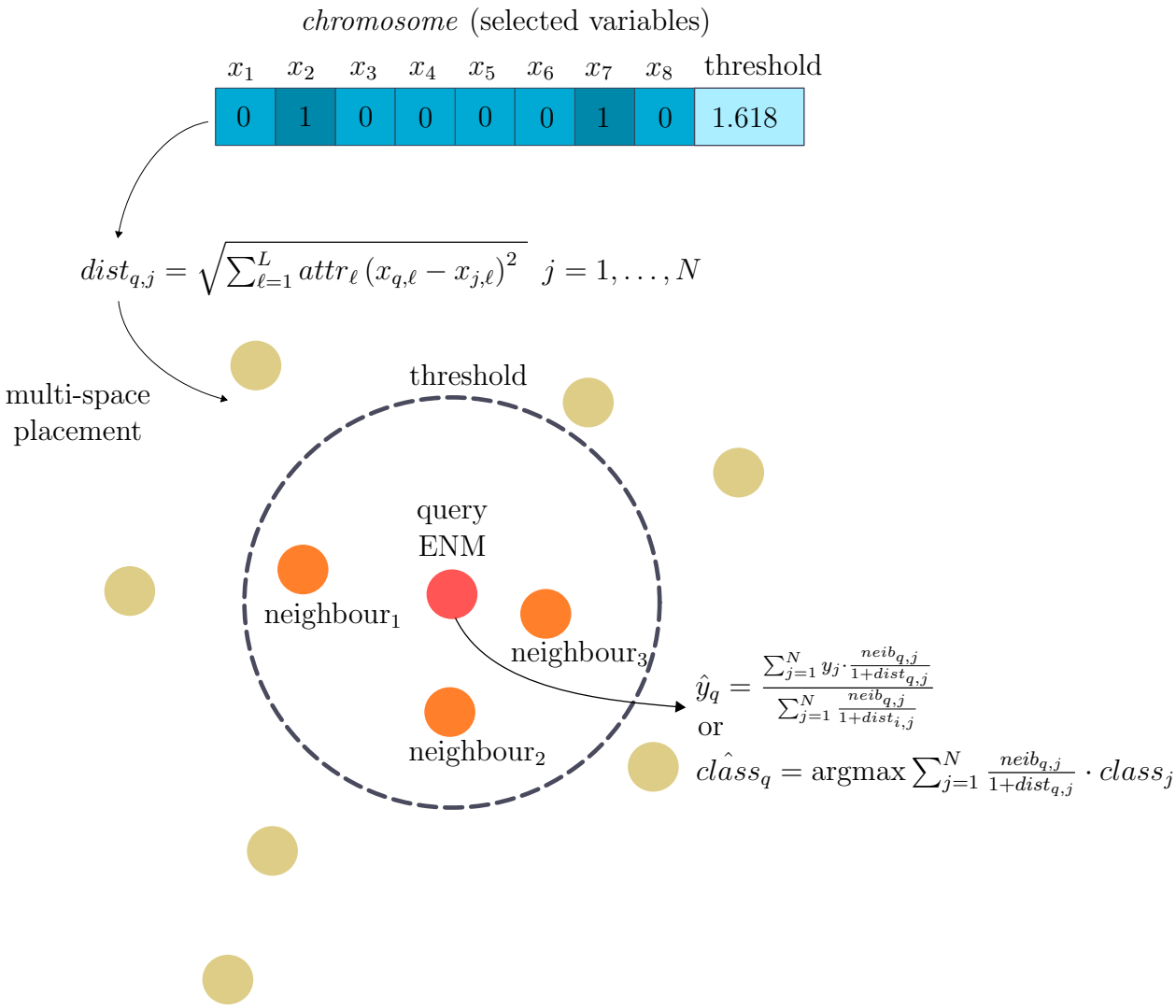

Figure 2: A schematic representation of the background read-across approach using the GAs optimisation scheme: the selected variables are used to compute the Euclidean distances in the multi-dimensional space and the optimal threshold value defines a hyper-sphere around a query ENM (red particle). ENMs inside the hyper-sphere are considered as neighbours $neib_{i,j} = 1$ (orange particles) whereas the rest ENMs (light yellow particles) do not belong to the reference ENM neighbourhood and are not involved in the read-across prediction $neib_{i,j} = 0$. In case of more than one types of available descriptors, neighbours are selected when all distance thresholds are satisfied. $attr_\ell$ is a binary variable equal to the respective *gene* values.

2.3 Read-across prediction

Each *chromosome* generated through the GA process corresponds to a read-across model. The predicted read-across value for the ith ENM in the training set is calculated using only the neighbour ENMs ($neib_{i,j} = 1$). This calculation assumes the existence of at least one neighbour ENM. For numerical endpoints, the prediction is computed as the weighted average of the endpoint values of neighbour ENMs (Eq. 2). For categorical endpoints, the prediction is the distance-weighted majority vote of the closest training neighbours (Eq. 3). All *class* arguments are of a binary type TRUE/FALSE.

$$\hat{y}_i = \begin{cases} \frac{\sum_{j=1}^{N_{tr}} \frac{neib_{i,j}}{1+dist_{i,j}} \cdot y_j}{\sum_{j=1}^{N_{tr}} \frac{neib_{i,j}}{1+dist_{i,j}}}, \text{ if } pred_i \neq 0 \\ \\ NA, \text{ if } pred_i = 0 \end{cases} \qquad \forall i = 1, \ldots, N_{tr} \tag{2}$$

$$\hat{class}_i = \begin{cases} \underset{class}{\text{argmax}} \left(\sum_{j=1}^{N_{tr}} \frac{neib_{i,j}}{1+dist_{i,j}} \cdot class_j \right), \text{ if } pred_i \neq 0 \\ \\ NA, \text{ if } pred_i = 0 \end{cases} \qquad \forall i = 1, \ldots, N_{tr} \tag{3}$$

where, N_{tr}, is the number of ENMs in the training set ,
$\hat{y}_i$, is the predicted endpoint value for the ith training ENM with at least one neighbour,
y_j is the actual endpoint value of the jth training ENM,
$\hat{class}_i$ is the predicted categorical endpoint value of the ith training ENM with at least one neighbour
$class_j$ is the actual categorical endpoint value of the jth training ENM
$neib_{i,j}$ is a binary variable taking the value of 1 if ENMs i and j are neighbours and 0 if they are not,
$dist_{i,j}$ is the Euclidean distance between ENMs i and j,
$pred_i$ is a binary variable that becomes equal to 1, when a read-across prediction is achieved for the ith ENM, and 0, if no prediction is possible.

A *chromosome* is accepted if the corresponding read-across model satisfies Eq. 4, which means that read-across predictions can be computed for at least a percentage of the training samples, defined by the parameter *predFactor*. In other words, this percentage of training samples have at least one neighbour. If a *chromosome* does not satisfy Eq. 4, it is rejected, and is substituted by its best *parent chromosome*.

$$\sum_{i=1}^{N_{tr}} pred_i \geq \ predFactor \cdot \ N_{tr} \tag{4}$$

where, $pred_i$, is a binary variable that becomes equal to 1, when a read-across prediction is achieved for the ith ENM, and 0, if no prediction is possible and,

predFactor, is a user-defined percentage of the N_{tr} training ENMs.

2.4 Evaluation of chromosomes

Each *chromosome* is evaluated for its ability to produce accurate predictions using a score (fitness value). *Chromosomes* with higher fitness values are more probable to survive during the "roulette-wheel" *selection* process of the evolutionary algorithm. The definition of the score function is based on the mean squared error (MSE) for numerical endpoints or the Matthews correlation coefficient (MCC) (6) for categorical endpoints. These metrics are computed over all training ENMs with at least one neighbour.

$$MSE = \frac{1}{N_{pred}} \sum_{i=1}^{N_{tr}} pred_i \, (y_i - \hat{y}_i)^2 \tag{5}$$

where, N_{tr} is the number of ENMs in the training set,
y_i and $\hat{y}_i$ are the actual and predicted endpoint values for the ith ENMs,
$pred_i$ is a binary variable that becomes equal to 1, when a read-across prediction is achieved for the ith ENM, and 0, if no prediction is possible and,
N_{pred} is the total number of ENMs with a successful prediction, $N_{pred} = \sum_{i=1}^{N_{tr}} pred_i$.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

where, TP (true positive) is the frequency of class TRUE ENMs correctly classified as "TRUE",
TN (true negative) is the frequency of class FALSE ENMs correctly classified as "FALSE",
FP (false positive - Type I error) is the frequency of class FALSE ENMs incorrectly classified as "TRUE" and,
FN (false negative - Type II error) is the frequency of class TRUE ENMs correctly classified as "FALSE".

Different objective functions (OF) to be minimised are defined next for the numerical and categorical endpoints. In particular, for numerical endpoints, the OF is defined as follows:

$$OF = MSE + wf_{OF} \cdot \sum_{\ell=1}^{L} attr_\ell \tag{7}$$

while for categorical endpoints, the following definition of the OF is used:

$$OF = |0.5 - 1/(1 + MCC)| + wf_{OF} \cdot \sum_{\ell=1}^{L} attr_\ell \tag{8}$$

where $attr_\ell$, is a binary variable indicating if the descriptor ℓ is selected and L is the total number of available descriptors.

In the OF definition for categorical endpoints, the first term is $|0.5 - 1/(1 + MCC)|$, which becomes zero when $MCC = 1$, i.e. in the case of perfect match between predicted and actual values. At this point we have to highlight that in case of $MCC = -1$ (total disagreement between predictions and actual endpoint values), *chromosome*'s score (Eq. 9) is automatically set to zero.

In both OF definitions, the second term is a "regularisation" term, which controls the influence of the number of selected variables. The addition of the regularisation term was inspired by the penalisation of the coefficient estimates in LASSO and ridge regression in order to avoid the risk of overfitting. Higher wf_{OF} values lead to the selection of fewer variables and this reduces the complexity of the produced model.

Finally, the fitness value of the chromosome is computed by just inverting the value of the OF (Eq. 9). The constant term 10^{-5} in the denominator ensures that the *score* value will not become infinite in the case of $OF = 0$.

$$score = 1/(OF + 10^{-5}) \tag{9}$$

2.5 Validation of the produced read-across model

An external validation scheme [8] is used to test the performance of the produced read-across model in terms of predicting accurately the endpoint on ENMs that have not been used during the training process. To this end, the read-across model which is the final outcome of the training evolutionary workflow, is used to compute endpoint predictions for the ENMs belonging to the test set. These calculations are performed using Eqs. 2 and 3, adapted for test samples. [6]

For numerical endpoints, the following validation metrics are computed: the q^2_{ext} statistic (Eq. 10) [8], the mean absolute error (*MAE*, Eq. 11) and the *MSE* metric (Eq. 12), which is adapted to the test samples:

$$q^2_{ext} = 1 - \frac{\sum_{i=1}^{N_{pred}} (y_i - \widehat{y_i})^2}{\sum_{i=1}^{N_{pred}} (y_i - \bar{y}_{tr})^2} \tag{10}$$

$$MAE = \frac{1}{N_{pred}} \sum_{i=1}^{N_{test}} |pred_i\,(y_i - \hat{y}_i)| \tag{11}$$

$$MSE = \frac{1}{N_{pred}} \sum_{i=1}^{N_{test}} pred_i\,(y_i - \hat{y}_i)^2 \tag{12}$$

where, y_i and $\hat{y}_i$ are the actual and predicted endpoint values over the test set,
$\bar{y}_{tr}$ is the averaged value of the endpoint over the N_{tr} training ENMs and,
N_{test}, is the number of ENMs in the test set ,
N_{pred} is the number of test ENMs with $pred_i \neq 0$.

In case of categorical endpoints, validation results are displayed in a confusion matrix (Table 2), where TP, TN, FP, FN have been defined in Eq. 6. The proportion of actual TRUE-class ENMs that are correctly classified as "TRUE" is measured by sensitivity (Sn, Eq. 13) and the proportion of actual FALSE-class ENMs that are correctly classified as "FALSE" is measured by specificity (Sp, Eq. 14). The overall success rate is measured by accuracy (Ac, Eq . 15). [9]

Table 2: Confusion matrix

		Predicted class	
		TRUE	FALSE
Actual class	TRUE	TP	FN
	FALSE	FP	TN

$$Sn = \frac{TP}{TP + FN} \tag{13}$$

$$Sp = \frac{TN}{TN + FP} \tag{14}$$

$$Ac = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

The MCC metric is also calculated for the test set, according to the Eq. 6, which is adapted for the test samples.

2.6 Use of the read-across model to predict the endpoints of untested ENMs

Fully validated read-across models can be used for predicting the endpoint of interest on untested ENMs, provided that the selected input descriptors indicated by the corresponding *genome* are available. The endpoint estimation process is similar to the training-validation process: for each untested ENM its neighbours are identified from the previously used training set, according to the Euclidean distances and the optimised threshold(s). The endpoint prediction is computed using Eq. 2 or Eq. 3, depending on the type of the endpoint (numerical/categorical). If no neighbours are identified for an untested ENM, this ENM is considered to be located outside the domain of applicability of the model and a read-across prediction is not possible.

3 Web Implementation of the grouping/read-across workflow

The grouping/read-across workflow described in the previous sections, has been implemented in the user-friendly Apellis R shiny web-application (`apellis.jaqpot.org/`). The applica-

tion is also available through a Docker Hub: `hub.docker.com/r/demetradanae/apellis`. Users of the Apellis tool can develop and validate predictive read-across models and can apply the produced models to reliably predict toxicity-related endpoints.

3.1 Read-across model development

The Apellis application offers four options for read-across model development according to the type of the endpoint (numerical or categorical) and the use of one or two similarity criteria for neighbour selection. For training a read-across model, the following input is required by the user:

- The user uploads the dataset as one or two csv files (depending on the availability of more than one types of properties as described in "Methods" section). The dataset contains the input descriptors for a number of ENMs and the corresponding toxicity endpoint values, that can be either numerical or categorical (TRUE/FALSE). The user has the option to scale the data in the range of [0,1].

- For splitting the full dataset into training and test sets, the user selects the partition method (random partition or Kennard-Stone partition [10]) and the train:test ratio.

- The user selects the hyperparameters of the algorithm, i.e. the operational parameters related to the evolutionary process and the objective function, and the probability values that control the biological operations of the genetic algorithm. Users are advised to select the default values for the hyperparameters and only if the results are not satisfactory, apply different combinations. The number of *generations* affects the computational time required by the algorithm and should be increased if the algorithm is not converging using the default value.

After the training procedure has been completed, the application returns the results to the user in the form of automatically created plots and tables:

- A scatter-plot depicting the actual and the predicted toxicity endpoint for all training and test ENMs (for numerical endpoints), or the confusion matrix containing the TP, TN, FP and FN frequencies for the test set (for categorical endpoints).

- A table containing the actual and the predicted endpoint values for the test set.

- A table containing information about the trained model: the optimised threshold(s), the *generation* that produced the model, the number of selected descriptors, the total number of test samples with a prediction, the score value, the external validation metrics, i.e. q^2_{ext}, *MSE* and *MAE* (for numerical endpoints) or the *MCC*, accuracy, sensitivity and specificity values (for categorical endpoints).

- The optimal set of descriptors selected by the model during the evolutionary process.

- The neighbours heatmap, where the neighbours of the test ENMs are depicted in colour code. The value of 1 (red) is used to denote that two ENMs are neighbours and the value of 0 (beige) is used in the opposite case.

All training results and the produced read-across model, can be downloaded for future use in the application. Figs. 3 and 4 contain screenshots of the Apellis application corresponding respectively to the development of read-across models for numerical and categorical endpoint predictions.

3.2 Using the model for obtaining predictions

This part of the application allows users to apply already developed read-across models to compute endpoint predictions for one or more untested ENMs. The user just needs to upload a csv file containing the input descriptors to the read-across model, for the untested ENMs. The analysis produces:

- A table that contains the predicted toxicity index value for all untested ENMs. In case that no neighbours are found in the training set, no predictions are produced.
- The neighbours heatmap, depicting the neighbours of the untested ENMs in the training set.

All results can be downloaded for future analysis.

3.3 Documentation and training material

In order to support and facilitate the use of the Apellis application by the nanosafety community, particular emphasis has been given to the development of documentation material and tutorials. These include a detailed user-guide, a video tutorial, quick-start guides and informative images. The documentation and training material is freely available through the web application.

4 Case studies

4.1 Numerical endpoints

The proposed read-across workflow has been previously applied on a dataset of gold surface-modified ENMs, which was presented by Walkey *et al.* (2014) [11] and was filtered by Varsou *et al.* (2017) [6] using the toxFlow web application. The dataset contains 40 physicochemical descriptors and 63 biological descriptors in the form of protein corona fingerprints (PCF) and the respective cell association values for 84 gold ENMs. From the available physicochemical descriptors, *zp_serum_sign* was removed because it has the same value for all samples.

The cell association of the ENMs with A549 human lung epithelial carcinoma cells was calculated as a pseudopartition coefficient, considering the magnesium content of the cells,

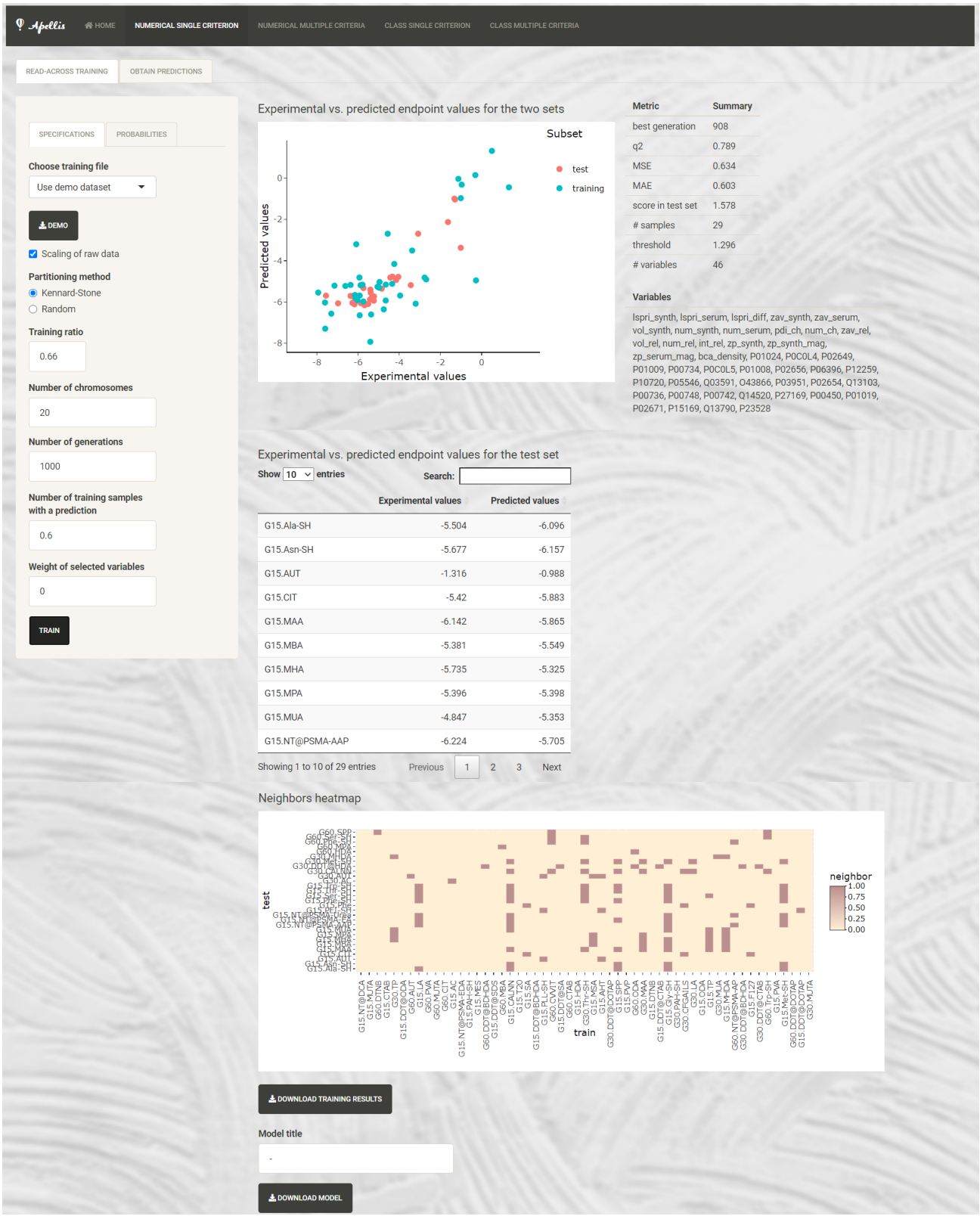

Figure 3: Training results for a numerical endpoint and a single similarity criterion, including a scatter-plot depicting the actual and the predicted toxicity endpoint values for all ENMs, the corresponding accuracy metrics, the predicted endpoint for each test sample and the neighbours heatmap.

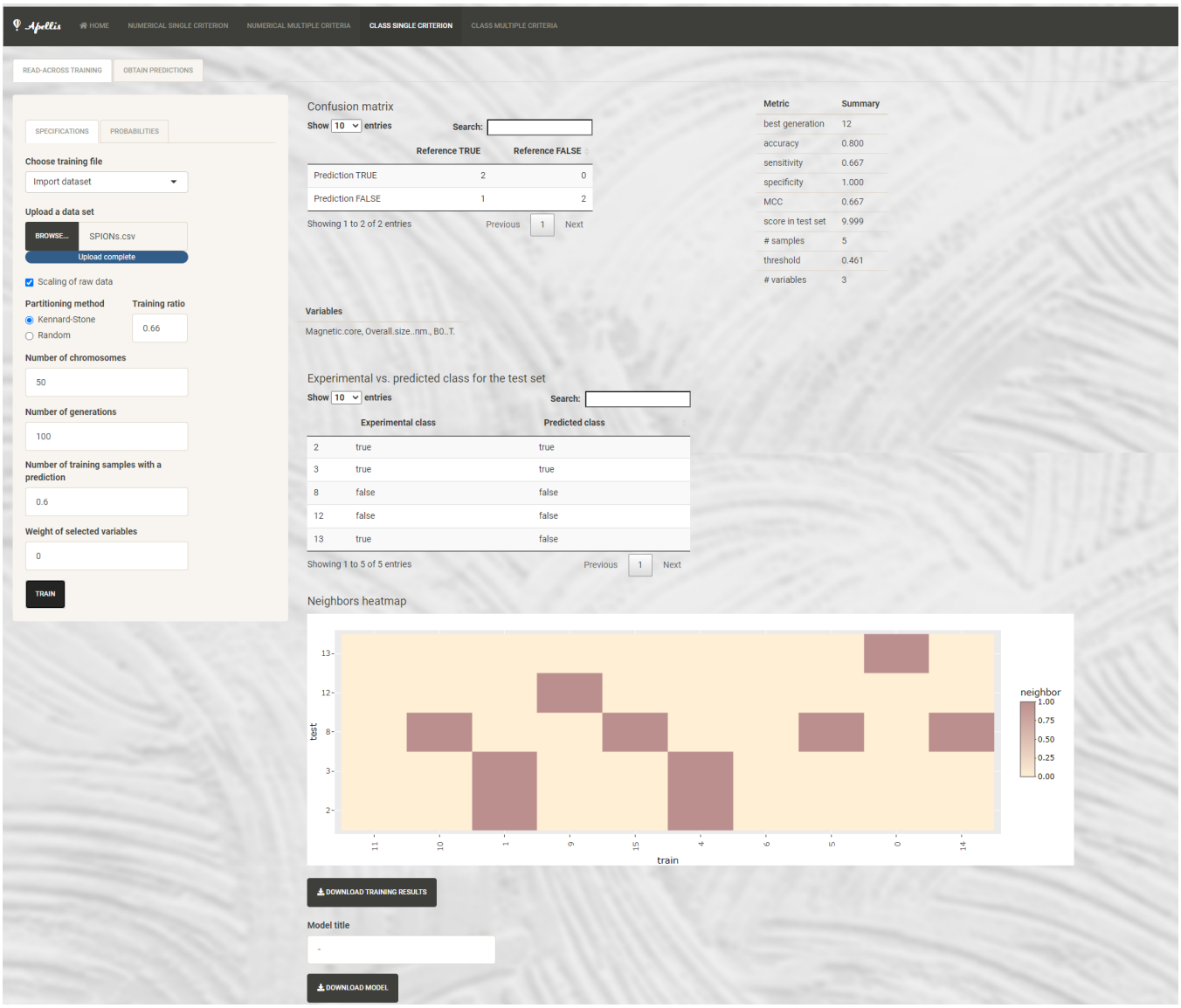

Figure 4: Training results for a categorical endpoint and a single similarity criterion, including the confusion matrix for the test set, the corresponding accuracy metrics, the predicted class for each test sample and the neighbours heatmap.

Table 3: Training parameters for the PCF read-across model.

Specifications	
Partitioning method	Kennard & Stone
Training ratio	0.66
# chromosomes	20
Generations	1000
Initial variable selection probability	0.6
Uniform *mutation* probability	0.01
Non-uniform *mutation* probability	0.1
Min threshold value	0.1
Max threshold value	mean(max(Dist))
bSA	1
Crossover probability	0.7
predFactor	0.6

and used as an endpoint in predictive models for the toxicity of gold ENMs. [11] The ENM-cell association -especially when studied in relevance to protein corona formation- can be considered as an interesting endpoint, as it is an important initiating event provoking disperse biological interactions such as inflammatory responses, biodistribution and toxicity. [11, 12, 13]. A detailed case study regarding this dataset can be found in the publication of Varsou *et al.* (2019). [6]

Table 3 presents the training specifications, and Table 4 summarises the produced results and statistics from the Apellis tool, with the presence or absence of the regularisation factor. Clearly, selection of a nonzero value for the regularisation factor wf_{OF} resulted to a read-across model that has a similar predictive power compared to the case where wf_{OF} is set to 0, but using significantly less input variables.

In the *Supplementary Information* file, analytical results for the PCF read-across model for $wf_{OF} = 0.05$ are presented. Table S1 presents the selected variables for the prediction of cell association and Table S2 depicts the training and test pairs of ENMs which are considered as neighbours by the algorithm. Pairs of neighbouring ENMs are marked with number 1. We can observe that in most cases neighbour ENMs share the same type of coating (either "anionic" or "cationic") thus, we can conclude that gold ENMs with the same surface charge modifications behave similarly as far as cell association levels are concerned. Future read-across studies could confirm the hypothesis of grouping the surface-modified ENMs based on their surface charge.

4.2 Categorical endpoints

4.2.1 MeOx dataset

The proposed read-across workflow is demonstrated through the Apellis application on the dataset for metal oxide (MeOx) ENMs, which was extracted by the publications of Zhang

Table 4: Results for the PCF read-across models. The number of predicted samples, and the *MSE*, *MAE* and q^2 validation metrics refer to the test set.

Optimised parameters and validation metrics		
wf_{OF}	0	0.05
threshold	1.336	0.996
# variables	57	27
# predicted samples	29	29
MSE	0.692	0.662
MAE	0.647	0.612
q^2	0.769	0.779

et al. (2012) [14] and Liu *et al.* (2013). [15] The original dataset consists of 24 MeOx ENMs with a detailed multiperspective toxicity profile, from seven different assays for two different cell lines (human bronchial epithelial (BEAS-2B) cells and murine myeloid (RAW 264.7) cells). The toxicity studies included a single-parameter cytotoxic assay, where cell viability was measured, and a multi-parameter toxicity assay for oxidative stress responses assessment of cells. [14] The toxicity profile of the ENMs was summarised to a toxicity class ("toxic"/"non-toxic") based on dose-response analysis and consensus Self-Organising Map clustering. [15]

From the original dataset the F_3O_4 sample was excluded due to high rate of impurities. Finally, 24 descriptors were selected for modelling purposes including ENMs size descriptors (the actual ENMs diameter and the hydrodynamic diameter in different media (water, BEGM and DMEM)), surface charge descriptors (the zeta potential at pH=7.4 and the isoelectic point), fundamental MeOx descriptors (the number of metal and oxygen atoms, the atomic mass of metals and the molecular weight, the metal electronegativity and the ionic index), MeOx energy descriptors (atomisation energy and sublimation energy, standard molar enthalpy of formation, lattice energy, ionisation energy and first molar ionisation energy) and ENMs energy descriptors (conduction band and valence band energies, chemical potential, hardness, electrophilicity and electronegativity of MeOx). All the descriptor values were obtained from the aforementioned toxicity studies.

Table 5 summarises the training specifications, and Table 6 the produced results and statistics from the Apellis tool using different regularisation factor values. The metrics are similar, however the inclusion of the regularisation factor has produced a model that uses only 2 descriptors and produces predictions for all test samples, in contrast to the model developed by setting the regularisation factor to zero, which requires 8 descriptors and fails to produce predictions for all ENMs.

Table S3 presents the two selected variables for toxicity prediction using the read-across model trained with wf_{OF}=0.05. The "atomic mass of metal" can be easily derived from the Periodic Table, whereas computations of the "chemical hardness" (η, Eq. 16) requires only the values of energy of conduction band (E_c) and energy of valence band (E_v). Table S4 presents the neighbour ENMs in this case study. In most pairs of neighbour MeOx, metals belongs to the same period.

Table 5: Training parameters for the MeOx models.

Specifications	
Partitioning method	Kennard & Stone
Training ratio	0.66
# *chromosomes*	50
Generations	1000
Initial variable selection probability	0.6
Uniform *mutation* probability	0.01
Non-uniform *mutation* probability	0.1
Min threshold value	0.1
Max threshold value	mean(max(Dist))
bSA	1
Crossover probability	0.7
predFactor	0.6

Table 6: Results for the MeOx read-across models. The number of predicted samples and the validation metrics refer to the test set.

Optimized parameters and validation metics		
wf_{OF}	0	0.05
threshold	0.69	0.38
# variables	8	2
# predicted samples	7	8
MCC	0.65	0.75
Accuracy	0.86	0.88
Specificity	0.83	0.83
Sensitivity	1.00	1.00

$$\eta = \frac{E_c + E_v}{2} \tag{16}$$

4.2.2 SPIONs dataset

Finally, our read-across methodology was applied in the dataset of 16 super-paramagnetic iron oxide nanoparticles (SPIONs) included in the publication of Kotzabasaki *et al.* (2020). [16] The dataset consists of six physicochemical descriptors and additional measurements of the SPIONs cell viability. From the available descriptors, zeta potential was excluded from the study due to high rate of missing values.

From the SPIONs cell viability, a categorical toxicity endpoint was derived; a cut-off value of cell viability was defined (75%) and samples with cell viability value less than the threshold value, were characterised as "toxic", otherwise were characterised as "non-toxic". Cell viability is a commonly evaluated biological endpoint in *in vitro* and *in silico*

Table 7: Training parameters for the SPIONs model.

Specifications	
Partitioning method	Kennard & Stone
Training ratio	0.66
# chromosomes	50
Generations	100
Initial variable selection probability	0.6
Uniform *mutation* probability	0.01
Non-uniform *mutation* probability	0.1
Min threshold value	0.1
Max threshold value	mean(max(Dist))
bSA	1
Crossover probability	0.7
wf_{OF}	0
predFactor	0.6

nanotoxicology and is usually the percentage of live cells in comparison to a control sample, after exposure to ENMs. [16, 17]

For this dataset an extra preprocessing step was applied: we removed the sample with ID 7, because it differentiates sample with ID 6 only in the type of coating which is not used as a training descriptor. In total 15 SPIONs were used. The missing relaxivity values were calculated following the clustering steps performed in the original study.

Due to the small size of the SPIONs dataset, one read-across model was trained, using a regularisation factor wf_{OF} equal to zero. Table 7 summarises the training specifications, and Table 8 the produced results and validation statistics extracted from the Apellis application.

Table S5 presents the selected variables of the SPIONs model. It is observed that two out of the three selected variables, namely the "Magnetic core" and the "Overall size", are also considered as the two important variables in the study of Kotzabasaki *et al.*. Table S6 presents the neighbour ENMs for this particular case study. It is observed that all neighbouring SPIONs share the same core (either magnetite or maghemite). It addition, in most cases, neighbour ENMs belong to the same hierarchical cluster according to the clustering analysis of Kotzabasaki *et al.*. Therefore we can conclude that there is an agreement between the two studies.

5 Conclusions

In this study, a previously proposed methodology for ENM grouping and read-across has been extended, to include model development functionalities for the prediction of categorical toxicity-related endpoints. The workflow offers the option to consider the multiperspective characterisation of ENMs using multiple similarity levels and, produces in an automatic way a grouping hypothesis that leads to trustworthy results in terms of prediction accuracy.

Table 8: Results for the SPIONs read-across model. The number of predicted samples and the validation metrics refer to the test set.

Optimized parameters and validation metrics	
threshold	0.43
# variables	3
# predicted samples	5
MCC	0.67
Accuracy	0.80
Specificity	1.00
Sensitivity	0.67

The proposed workflow also performs variable selection, indicating the most informative descriptors for the endpoint predictions.

Additionally, a web application named Apellis has been developed in this work, which implements the workflow and offers it as a user-friendly tool to the nanosafety community for supporting risk assessment. Through Apellis, users can train and validate predictive read-across models using their own specific data and share the produced models with the community as ready-to-use web applications. The Apellis tool is equipped with a number of visualisation tools and services, which present graphically the results in the forms of easy-to-interpret plots and tables. All modelling functionalities and services are offered through an integrated user-friendly environment, which can be used even by users that do not have any prior computational skills.

6 Abbreviations

- ENM: engineered nanomaterial
- FN: false negative
- FP: false positive
- GA: genetic algorithm
- GUI: graphical user interface
- *MAE*: mean absolute error
- *MCC*: Matthews correlation coefficient
- MeOx: metal oxide
- MINLP: mixed integer non-linear mathematical optimisation problem
- *MSE*: mean squared error

- *OF*: objective function
- PCF: protein corona fingerprints
- SAR: Structure-Activity relationship (models)
- SPIONs: super-paramagnetic iron oxide nanoparticles
- TN: true negative
- TP: true positive

Conflicts of interest

The authors declare no conflicts of interest.

Acknowledgements

The research work of D.-D.V was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the HFRI PhD Fellowship grant (Fellowship Number: 637).

H.S. acknowledges support by the NanoCommons project, which has received funding from the EU Horizon 2020 Programme (H2020) under grant agreement no. 731032 and by the NanoSolveIT project, which has received funding from the EU Horizon 2020 Programme (H2020) under grant agreement no. 814572.

References

[1] G Basei, D Hristozov, L Lamon, A Zabeo, N Jeliazkova, G Tsiliki, A Marcomini, and A Torsello. Making use of available and emerging data to predict the hazards of engineered nanomaterials by means of in silico tools: A critical review. *NanoImpact*, 13:76–99, 2019.

[2] Antreas Afantitis, Georgia Melagraki, Panagiotis Isigonis, Andreas Tsoumanis, Dimitra Danai Varsou, Eugenia Valsami-Jones, Anastasios Papadiamantis, Laura-Jayne A Ellis, Haralambos Sarimveis, Philip Doganis, et al. Nanosolveit project: Driving nanoinformatics research to develop innovative and integrated tools for in silico nanosafety assessment. *Computational and structural biotechnology journal*, 18:583–602, 2020.

[3] Agnieszka Gajewicz, K Jagiello, MTD Cronin, J Leszczynski, and Tomasz Puzyn. Addressing a bottle neck for regulation of nanomaterials: quantitative read-across (nano-qra) algorithm for cases when only limited data is available. *Environmental Science: Nano*, 4(2):346–358, 2017.

[4] L Lamon, K Aschberger, D Asturiol, A Richarz, and A Worth. Grouping of nanomaterials to read-across hazard endpoints: a review. *Nanotoxicology*, 13(1):100–118, 2019.

[5] ECHA. Guidance on information requirements and chemical safety assessment, appendix r.6-1 for nanomaterials applicable to the guidance on qsars and grouping of chemicals, December 2019.

[6] Dimitra-Danai Varsou, Antreas Afantitis, Georgia Melagraki, and Haralambos Sarimveis. Read-across predictions of nanoparticle hazard endpoints: a mathematical optimization approach. *Nanoscale Advances*, 1(9):3485–3498, 2019.

[7] Dimitra-Danai Varsou, Georgia Tsiliki, Penny Nymark, Pekka Kohonen, Roland Grafström, and Haralambos Sarimveis. toxflow: a web-based application for read-across toxicity prediction using omics and physicochemical data. *Journal of chemical information and modeling*, 58(3):543–549, 2017.

[8] Alexander Tropsha, Paola Gramatica, and Vijay K Gombar. The importance of being earnest: validation is the absolute essential for successful application and interpretation of qspr models. *QSAR & Combinatorial Science*, 22(1):69–77, 2003.

[9] Dimitra-Danai Varsou, Antreas Afantitis, Andreas Tsoumanis, Georgia Melagraki, Haralambos Sarimveis, Eugenia Valsami-Jones, and Iseult Lynch. A safe-by-design tool for functionalised nanomaterials through the enalos nanoinformatics cloud platform. *Nanoscale Advances*, 1(2):706–718, 2019.

[10] Ronald W Kennard and Larry A Stone. Computer aided design of experiments. *Technometrics*, 11(1):137–148, 1969.

[11] Carl D Walkey, Jonathan B Olsen, Fayi Song, Rong Liu, Hongbo Guo, D Wesley H Olsen, Yoram Cohen, Andrew Emili, and Warren CW Chan. Protein corona fingerprinting predicts the cellular interaction of gold and silver nanoparticles. *ACS nano*, 8(3):2439–2455, 2014.

[12] Rong Liu, Wen Jiang, Carl D Walkey, Warren CW Chan, and Yoram Cohen. Prediction of nanoparticles-cell association based on corona proteins and physicochemical properties. *Nanoscale*, 7(21):9664–9675, 2015.

[13] S Palchetti, L Digiacomo, D Pozzi, G Peruzzi, E Micarelli, M Mahmoudi, and G Caracciolo. Nanoparticles-cell association predicted by protein corona fingerprints. *Nanoscale*, 8(25):12755–12763, 2016.

[14] Haiyuan Zhang, Zhaoxia Ji, Tian Xia, Huan Meng, Cecile Low-Kam, Rong Liu, Suman Pokhrel, Sijie Lin, Xiang Wang, Yu-Pei Liao, et al. Use of metal oxide nanoparticle band gap to develop a predictive paradigm for oxidative stress and acute pulmonary inflammation. *ACS nano*, 6(5):4349–4368, 2012.

[15] Rong Liu, Hai Yuan Zhang, Zhao Xia Ji, Robert Rallo, Tian Xia, Chong Hyun Chang, Andre Nel, and Yoram Cohen. Development of structure–activity relationship for metal oxide nanoparticles. *Nanoscale*, 5(12):5644–5653, 2013.

[16] Marianna I Kotzabasaki, Iason Sotiropoulos, and Haralambos Sarimveis. Qsar modeling of the toxicity classification of superparamagnetic iron oxide nanoparticles (spions) in stem-cell monitoring applications: an integrated study from data curation to model development. *RSC Advances*, 10(9):5385–5391, 2020.

[17] Valérie Forest, Jean-François Hochepied, and Jérémie Pourchez. Importance of choosing relevant biological end points to predict nanoparticle toxicity with computational approaches for human health risk assessment. *Chemical research in toxicology*, 32(7):1320–1326, 2019.