 doi:10.20944/preprints202210.0402.v1

*Article*

# q-RASAR Modeling of Cytotoxicity of $TiO_2$-based Multi-component Nanomaterials

**Arkaprava Banerjee[1], Supratik Kar[2]*, Agnieszka Gajewicz-Skretna[3] and Kunal Roy[1]***

[1] Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India
[2] Chemometrics & Molecular Modeling Laboratory, Department of Chemistry, Kean University, 1000 Morris Avenue, Union, New Jersey 07083, USA
[3] Laboratory of Environmental Chemoinformatics, Faculty of Chemistry, University of Gdansk, Wita Stwosza 63, 80-308 Gdansk, Poland
*Corresponding author: skar@kean.edu (S.K.); kunal.roy@jadavpuruniversity.in (K.R.)

**Abstract:** Read-Across Structure-Activity Relationship (RASAR) is an emerging cheminformatic approach that combines the usefulness of a QSAR model and similarity-based Read-Across predictions. In this work, we have generated a simple, interpretable, and transferable quantitative-RASAR (q-RASAR) model which can efficiently predict the cytotoxicity of $TiO_2$-based multi-component nanomaterials. The data set involves 29 $TiO_2$-based nanomaterials which contain specific amounts of noble metal precursors in the form of Ag, Au, Pd, and Pt. The data set was rationally divided into training and test sets and the Read-Across-based predictions for the test set were generated using the tool Read-Across-v4.1 available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. The hyperparameters were optimized based on the training set data and using this optimized setting, the Read-Across-based predictions for the test set were obtained. The optimized hyperparameters and the similarity approach, which yields the best predictions, were used to calculate the similarity and error-based RASAR descriptors using the tool RASAR-Desc-Calc-v2.0 available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. These RASAR descriptors were then clubbed with the physicochemical descriptors and were subjected to features selection using the tool Best Subset Selection v2.1 available from https://dtclab.webs.com/software-tools. The final set of selected descriptors was used to develop multiple linear regression based q-RASAR models, which were validated using stringent criteria as per the OECD guidelines. Finally, a random forest model was also developed with the selected descriptors. The final machine learning model can efficiently predict the cytotoxicity of $TiO_2$-based multi-component nanomaterials superseding previously reported models in the prediction quality.

**Keywords:** QSAR; q-RASAR; random forest; machine learning; $TiO_2$-based nanoparticles

## 1. Introduction

In the present day, nanoparticles have become critical components of our everyday life. They have a wide array of applications ranging from the textile to the electronics industry (1,2). As mentioned in the project on Emerging Nanotechnologies (https://www.nanotechproject.org/inventories/consumer/), the annual turnover of the newly developed nano-products is expected to be around $2.6 trillion (3). Nanoparticles also find their use as cosmetics, nanomedicine, and dental fillings (4,5). Nanoparticles can be broadly classified into two distinct classes namely carbon nanoparticles and metal/metal oxide nanoparticles (6). The metal oxide nanoparticles possess certain unique properties like optical, electronic, and magnetic, which are responsible for their benefits as well as their toxic effects on the ecosystem (1,7,8). Humans have used the cytotoxic properties of nanoparticles to make recent advances in the treatment of cancer and the development of potential nano-anticancer agents (9). A common mechanism of toxicity of the metal oxide nanoparticles lies in the generation of reactive oxygen species (ROS),

 doi:10.20944/preprints202210.0402.v1

which leads to the oxidation and death of the cells (10). Titanium dioxide is one such metal oxide nanoparticle that apart from its utility in photodegradation, photocatalysis, energy conversion, and reduction of $CO_2$ in organic fuels, also shows cytotoxicity activity in humans and animals (11). The present study is to model the cytotoxic properties of hybrid $TiO_2$ nanoparticles using in-silico approaches.

The emergence of various Machine Learning (ML) algorithms and their potential applicability in cheminformatics has led to their adoption by researchers from all over the world (12). Many informatics and omics researchers believe that ML, a form of Artificial Intelligence (AI), can effectively substitute in-vivo tests shortly thus obviating animal experimentation. The most promising features of ML algorithms lie in their ability to produce fast, reliable, and accurate results with minimal manpower. Among the earliest and most popular ML applications used to predict the activity/property/toxicity of substances are Quantitative Structure-Activity/Property/Toxicity Relationships (QSAR/QSPR/QSTR). Classical QSAR approaches involve a set of independent variables known as descriptors used to predict the response variable by using a simple and interpretable mathematical model represented as an equation (13). The prediction results generated by the QSAR approach are accepted by regulations such as EU-REACH (https://echa.europa.eu/regulations/reach/understanding-reach) for data gap filling in case of the non-availability of experimental data (14). The recent trend is to adopt similarity-based approaches like Read-Across, which does not involve the development of a mathematical model (15). The basic theory behind the Read-Across-based predictions is that a molecule with an unreported experimental endpoint value should possess an endpoint value similar to the molecules which are structurally and/or biologically similar to the query molecule. Read-Across-based predictions are generated by the interpolation or extrapolation of the properties of one or more source compounds to one or more target compounds concerning their similarity values. Chatterjee et al. (16) developed a java-based tool (Read-Across-v4.1) that quickly computes Read-Across-based predictions based on the Euclidean Distance-based similarity, the Gaussian Kernel-based similarity, and the Laplacian Kernel-based similarity approaches. This tool also computes various similarity and error-based metrics for each query compound which helps one to estimate the uncertainty in the Read-Across-based predictions (17). Although Read-Across is originally an unsupervised approach, the tool Read-Across-v4.1 (16) (available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) utilizes a set of optimized hyperparameters generated from the tool Auto_RA_Optimizer-v1.0 (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) which incorporates a supervised learning algorithm in the Read-Across-based predictions. In many cases, it has been observed that the predictive ability of the similarity-based Read-Across algorithm supersedes the QSAR model; however, the only drawback of Read-Across is its lack of interpretability of the essential features. Aiming to overcome this, Luechtefeld et al. (18) introduced the concept of Read-Across Structure-Activity Relationship (RASAR) which combines the advantages of both the QSAR and Read-Across algorithms performed in an ML-based approach. They performed classification-based models which yielded graded predictions. Later, Banerjee and Roy (19) introduced the quantitative RASAR (q-RASAR) modeling using a set of similarity and error-based descriptors. They have shown that the q-RASAR models possess a much higher predictive ability and lower MAE as compared to their QSAR analysis and Read-Across-based predictions. The advantages of q-RASAR approach lie in the consideration of the similarity and error-based measures as descriptors yet generating simple, interpretable, transferable, and reproducible models with enhanced predictivity. The RASAR descriptor RA function is a specialized descriptor derived from Read-Across and a composite of all the physicochemical descriptors. This descriptor involves either Euclidean Distance-based similarity, Gaussian Kernel Similarity, or Laplacian Kernel Similarity depending on the choice of input in the descriptor calculator tool RASAR-Desc-Calc-v2.0 (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home). The RA function calculated using the Gaussian Kernel and the Laplacian Kernel Similarity approaches involves a nonlinear function, although the q-

 doi:10.20944/preprints202210.0402.v1

RASAR models consisting of the RA function are developed within a linear model generation framework (20).

The current study is designed to model the cytotoxicity of $TiO_2$ nanoparticles hybridized with Ag, Au, Pd, and Pt. The experimental toxicity data was obtained from Mikolajczyk et al (11). The same set of descriptors used in the previous publication was initially employed to model the cytotoxicity. We have made cytotoxicity predictions using the q-RASAR approach after optimizing the required hyperparameters. Finally, we have used two additional descriptors, one among them being generated using the VASP package (https://www.vasp.at), and developed another q-RASAR model. Lastly, we have implemented Random Forest regression using the descriptors selected in the final q-RASAR model to achieve the model's enhanced predictivity.

## 2. Materials and Methods

### *2.1. Collection of cytotoxicity data of hybrid $TiO_2$ nanoparticles*

The data set containing cytotoxicity data in the form of pEC50 for 29 hybrid $TiO_2$ nanoparticles have been obtained from Mikolajczyk et al (11), and made available in an excel sheet in **Supplementary Materials SI-1**. We have initially considered this data set consisting of experimental toxicity data along with two descriptors $\chi_{mix}$ (additive electronegativity) and A (electron affinity) for the generation of a q-RASAR model. Additionally, we have also considered two additional descriptors BET surface area (Brunauer, Emmett, and Teller surface area) and $\Delta H_f$ (formation energy concerning the constituent elements of the multicomponent $TiO_2$-based nanomaterials [eV/atom], also represented as X_H_eVatom in some Figures of this manuscript) for the generation of another q-RASAR model. The descriptor denoting the formation energy has been retrieved from the Open Quantum Materials Database (OQMD) (21,22).

### *2.2. Dataset division*

Due to the absence of a true external set of data, the available data set consisting of data for 29 hybrid $TiO_2$-based nanomaterials has been rationally divided into training and test sets to check the robustness and the predictive ability of the model. The division of the data set was performed by employing the sorted response-based division algorithm (13).

### *2.3. Read-Across predictions*

Read-Across, being a similarity-based approach, does not require a statistically reliable model for the prediction of query compounds. For efficient predictions, there arises a need to optimize the hyperparameters associated with the computation of similarities. The optimization of the hyperparameters ($\sigma$, $\gamma$ and the number of close source compounds) as well as the similarity-based approach (Euclidean Distance-based, Gaussian Kernel Similarity-based, and Laplacian Kernel Similarity-based) was done by the java-based tool Auto_RA_Optimizer-v1.0 available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. The training set data were further divided into sub-training and sub-test sets, and these were taken as inputs in the Auto_RA_Optimizer-v1.0 tool. This tool generated predictions from all possible combinations of $\sigma$, $\gamma$ (in the range of 0.25 to 2 at the interval of 0.25) and the number of close source compounds, and a suitable combination was selected. This optimized setting was then used to generate the Read-Across-based predictions using the java-based tool Read-Across-v4.1 available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home taking the training and test set files as inputs.

### *2.4. Calculation of RASAR Descriptors*

Read-Across Structure-Activity Relationship (RASAR) is a novel cheminformatic approach that combines the concepts of Read-Across and QSAR together (**Figure 1**) (18,19). To perform the q-RASAR analysis, it is essential to calculate the similarity and error-based

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2022 doi:10.20944/preprints202210.0402.v1

RASAR descriptors for the training and test sets. The optimized setting of the Read-Across hyperparameters was used to calculate the RASAR descriptors using the java-based tool RASAR-Desc-Calc-v2.0 available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. The training set was used as the source set to compute the RASAR descriptors of the test set (query set) and training set itself (which was also used as the query set). It is essential to note that this tool identifies the training set when it is used as a query set input and, leaves out the identical compound in the source set for the computation of the RASAR descriptors thus enabling the use of a Leave-same-out (LSO) algorithm and prevents overfitting. This algorithm does not apply to the test set as no identical compounds exist. Thus, the tool's ability to identify the identical set and accordingly adopt the LSO algorithm demonstrates an automated algorithm in the calculation of the RASAR descriptors.

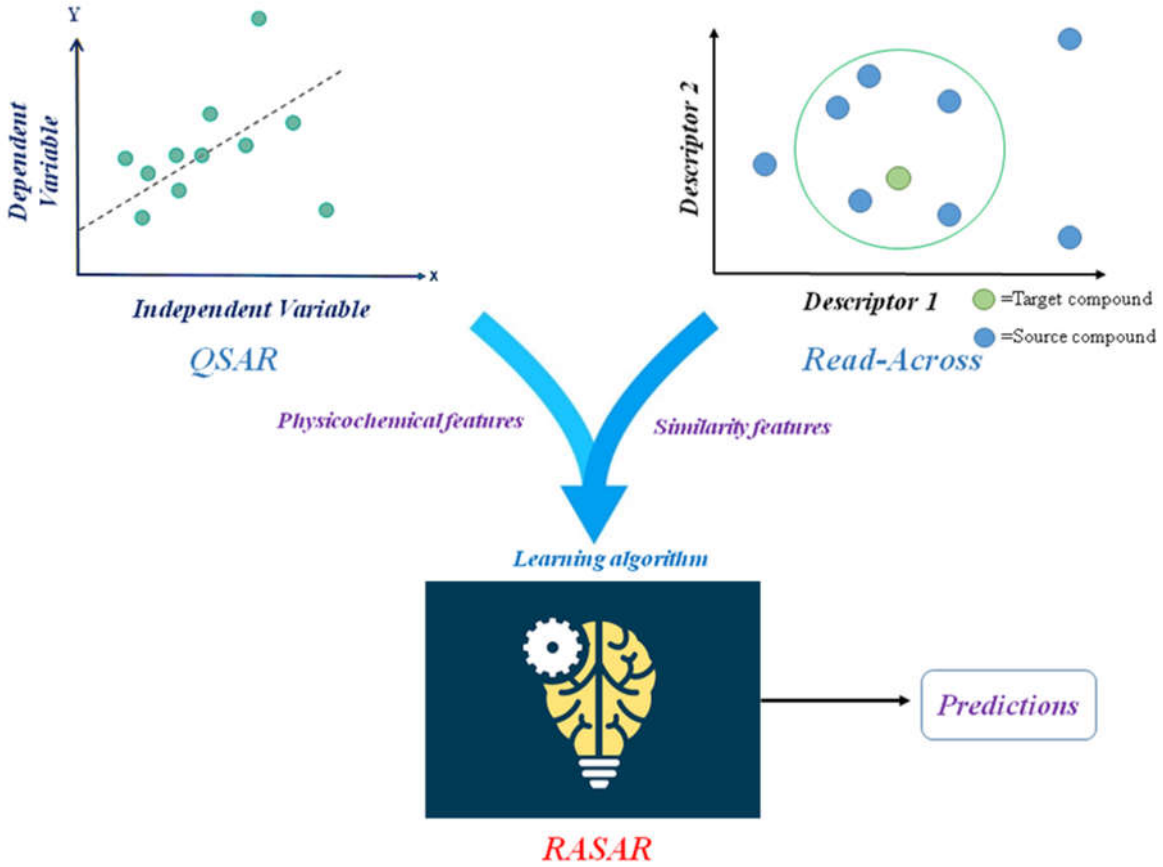


**Figure 1.** Schematic representation of the concept of RASAR.

*2.5. Similarity-based RASAR descriptors (19)*

The tool RASAR-Desc-Calc-v2.0 calculates a total of 15 similarity and error-based descriptors based on any one of the three similarity-based approaches (Euclidean Distance-based, Gaussian Kernel Similarity-based, and Laplacian Kernel Similarity-based) and a given set of close source compounds. The descriptor *RA function* is a Read-Across-derived measure which is a composite function of the selected electronic features and contains information from all these descriptors. The descriptor *SD_Activity* is the weighted standard deviation of the activity values of the close "n" source compounds for a given target compound. *SE* is the weighted standard error of the activity values of the close "n" source compounds for a particular target compound. The descriptor *CVact* is the coefficient of variation of the activity values of the close "n" source compounds for a particular target compound. *CVsim* is the coefficient of variation of the similarity values of the close "n" source compounds for a given target compound. *MaxPos* is the maximum similarity value

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2022 doi:10.20944/preprints202210.0402.v1

of the target compound with a close source compound which has an activity response value higher than the training set mean response. *MaxNeg* is the maximum similarity value of the target compound with a close source compound that has an activity response value lower than the training set mean response. *Abs MaxPos-MaxNeg* or *Abs Diff* is the absolute difference in the values of the *MaxPos* and *MaxNeg* for a particular query compound. The descriptor *Avg. Sim* is the average similarity values of the "n" close source compounds for a particular target compound. *SD_Similarity* is the standard deviation of the similarity values of the close "n" source compounds for a given target compound. The descriptor $g_m$ a.k.a. Banerjee-Roy coefficient is a measure that identifies the probability that the query compound is active or inactive (19). The value of this coefficient ranges from -1 to +1. The descriptors *gm*Avg. Sim* and *gm*SD_Similarity* are the products of the values of $g_m$ with *Avg. Sim* and *SD_Similarity*. The descriptor *Pos.Avg.Sim* indicates the average similarity values among the close "n" source compounds which have response values greater than the training set mean response value, while the descriptor *Neg.Avg.Sim* indicates the average similarity values among the close "n" source compounds which have response values lower than the training set mean response value. The list of training and test set compounds along with important physicochemical and RASAR descriptors are provided in **Supplementary Materials SI-1**.

*2.6. Feature selection and model development*

After computation of the RASAR descriptors for the training and test sets, they were clubbed with the structural descriptors already available previously, and feature selection was performed with the java-based BestSubsetSelection_v2.1 tool available from https://dtclab.webs.com/software-tools. This tool generates all possible combinations of models from a defined number of descriptors as specified by the user. Moreover, it can show all possible models which meet the inter-correlation cut-off for the given number of descriptors. A univariate model was selected when we initially used the two descriptors mentioned above and a bivariate model was selected when we used two additional descriptors apart from the previous ones. The final MLR q-RASAR model with the regression coefficients along with their corresponding validation metric values and the associated Golbraikh and Tropsha's criteria (23) were checked using the java-based tool MLR-PlusValidation1.3 available from https://dtclab.webs.com/software-tools taking the training and test set as inputs. Different validation metrics were computed strictly according to the OECD Guidelines (https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm), and the computed external and internal validation metrics have been listed in **Table 1**. The inter-correlation of the two descriptors in the bivariate model was checked using MINITAB 14 (https://minitab.informer.com/14.1/). The detailed workflow has been demonstrated in **Figure 2**.

**Table 1.** List of the internal and external validation metrics computed for the q-RASAR models (13).

| Metrics | Mathematical Expressions |
|---|---|
| $R^2$ | $1-\frac{\Sigma(Y_{obs}-Y_{calc})^2}{\Sigma(Y_{obs}-\overline{Y}_{training})^2}$ |
| $Q^2_{(LOO)}$ | $1-\frac{\Sigma(Y_{obs(train)}-Y_{pred(train)})^2}{\Sigma(Y_{obs(train)}-\overline{Y}_{training})^2}$ |
| $Q^2_{F1}$ | $1-\frac{\Sigma(Y_{obs(test)}-Y_{pred(test)})^2}{\Sigma(Y_{obs(test)}-\overline{Y}_{training})^2}$ |
| $Q^2_{F2}$ | $1-\frac{\Sigma(Y_{obs(test)}-Y_{pred(test)})^2}{\Sigma(Y_{obs(test)}-\overline{Y}_{test})^2}$ |
| $MAE95\%_{(test)}$ | $\frac{\Sigma\lvert Y_{obs(test)}-Y_{pred(test)}\rvert}{n_{test}}$ |

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2022 doi:10.20944/preprints202210.0402.v1

| | |
|---|---|
| | $n_{test}$= number of test set compounds; the computation is done after omitting 5% high residual data points |

$R^2$=Determination coefficient

$Q^2_{(LOO)}$=cross-validated determination coefficient

$Q^2_{F1}$ & $Q^2_{F2}$ = determination coefficient for the external set

$Y_{obs}$=observed biological activity

$Y_{calc}$=calculated biological activity

$Y_{pred}$=predicted biological activity

$n_{test}$=number of compounds in the test set

*2.7. Random forest predictions*

Random forest (RF), originally suggested by Breiman (24), is one of the most successful machine-learning algorithms for practical prediction purposes. It is an ensemble of unpruned regression (or classification) trees created using bootstrap samples of the training data and random feature selection. Predictions are made by aggregating (averaging in case of regression and majority votes in case of classification) the predictions of the ensemble. In the RF algorithm, the three parameters, N_estimators (the number of trees in the forest), max_features (the number of features to consider when looking for the best split), and min_samples_leaf (the minimum number of samples required to be at a leaf node) affect the model performance the most and need to be determined by users.

In the present study, random forest regression modeling was carried out using only two features selected for the RASAR model 2 in Jupyter Notebook web application (25) in the Anaconda3 navigator version 2022.05 with Python version 3.10.4. The squared error function was used to measure the quality of a split. We have used N_estimators as 100 and the random state as none with all other default options. Fivefold cross-validation (in terms of average cross-validated mean absolute error) was used for the training set to understand the model quality, and the quality of test set predictions was determined by the usual test set quality metrics ($Q^2_{F1}$ and $Q^2_{F2}$).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2022 doi:10.20944/preprints202210.0402.v1

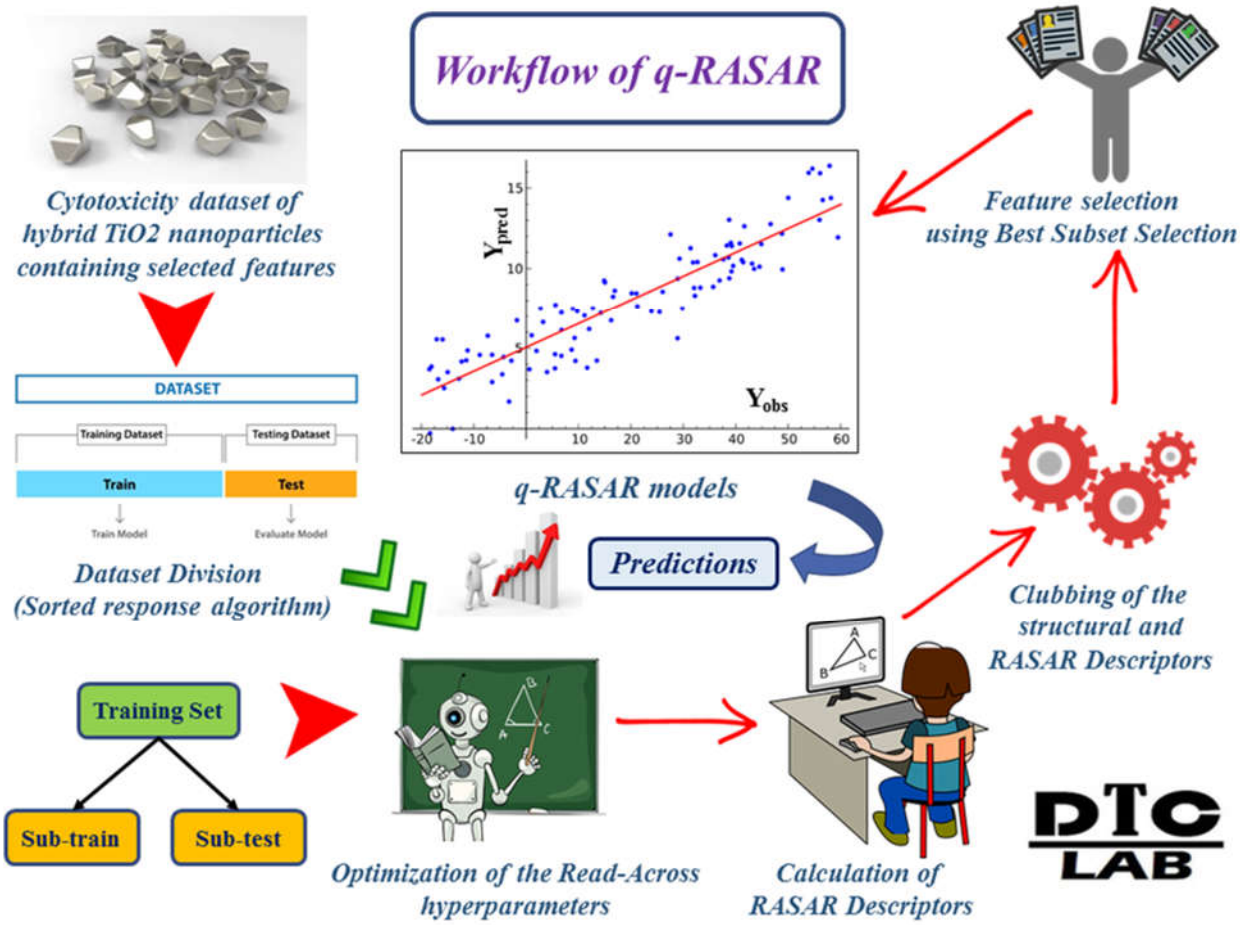


**Figure 2.** Workflow of q-RASAR modeling of hybrid $TiO_2$ nanomaterials.

## 3. Results and Discussion

### *3.1. Optimization of hyperparameters associated with the Read-Across-based predictions*

The Read-Across-based predictions generated from the java-based tool Read-Across-v4.1 utilizes the Euclidean Distance-based similarity, the Gaussian Kernel-based similarity, and the Laplacian Kernel-based similarity approaches. Therefore, it is essential to provide an optimized setting of the hyperparameters to compute the Read-Across predictions. To identify the optimized setting of the hyperparameters in compliance with Machine Learning approaches, the optimization was performed based on the training/source compounds. The training set was divided into sub-training and sub-test sets and was taken as inputs in the java-based tool **Auto_RA_Optimizer-v1.0** available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. This tool computes all possible predictions resulting from different combinations of the hyperparameters, and the results are tabulated in terms of the values of $Q^2_{F1}$ and MAE in the output file **SubTestSetFileName_Metrics.xlsx**. A suitable combination of the hyperparameters ($\sigma$=1, $\gamma$=1, number of close source compounds=2 for RASAR Model 1 and $\sigma$=2, $\gamma$=2, number of close source compounds=5 for RASAR Model 2) was selected based on the values of of $Q^2_{F1}$ and MAE, and this optimized setting was used to compute the final Read-Across-based predictions. The training and test set files were taken as inputs for the tool Read-Across-v4.1, and after providing the optimized hyperparameters as inputs, the Read-Across-based predictions were obtained. For the first case, where we used two electronic descriptors ($\chi_{mix}$ and A), the best predictions were obtained using the Laplacian Kernel-based predictions with the values of $Q^2_{F1}$ and $Q^2_{F2}$ were 0.88 and 0.88 respectively. In the second case, where we have used four physicochemical descriptors ($\chi_{mix}$, A, BET surface area and $\Delta H_f$), the best predictions were obtained using the Gaussian Kernel-based

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2022 doi:10.20944/preprints202210.0402.v1

predictions with the values of $Q^2_{F1}$ and $Q^2_{F2}$ being 0.87 and 0.87 respectively. The output (predictions) of the final Read-Across-based predictions has been provided in **Supplementary Materials SI-1**.

*3.2. Development of q-RASAR models*

Simple, interpretable, reproducible, and transferable MLR q-RASAR models were generated in both cases i.e. originally using two electronic descriptors (MLR RASAR Model 1 with only one final descriptor) and using four electronic descriptors (MLR RASAR Model 2 with only two final descriptors). **Figure 3** represents MLR RASAR Model 1 (a) and MLR RASAR Model 2 (b) in the form of bubble plots which depict the importance of each descriptor in terms of the standardized regression coefficients and their positive or negative contribution towards the toxicity. The MLR RASAR Model 1 is a univariate model (a LR model) consisting of the RASAR descriptor RA function which efficiently predicts the toxicity of $TiO_2$ in terms of $pEC_{50}$. At the same time, the MLR RASAR Model 2 is a bi-variate model consisting of the descriptor $\Delta H_f$ (an electronic descriptor) and SD_Activity (a RASAR descriptor). The internal and external validation of the models was performed adhering strictly to the OECD criteria (https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm), and the corresponding validation metrics have been tabulated in **Table 2**. It is essential to note that both models passed the prescribed limits of the Golbraikh and Tropsha criteria. The inter-correlation among the descriptors in RASAR Model 2 has been checked and is provided in the form of a heat map in **Figure 4**.

Additionally, we have taken the RASAR descriptors from MLR RASAR Model 2 and developed Random Forest regression with five-fold cross-validation using a python-based source code. As in this data set, we have only a limited number of data points, and we have used only selected features for RF regression and default values of different hyperparameters to obtain fully grown trees. The results indicate that the external predictive ability of the newly developed RF RASAR Model was even better than the MLR RASAR Model 2, thus superseding all the previous models in terms of predictivity. The validation metrics of the RF RASAR Model 2 have been reported in **Table 2**. The variable importance plot and sample trees generated in the Random Forest model development have been presented in **Figure 5** and **Figure 6,** respectively. **Figure 5** agrees with **Figure 2(b)** regarding the relative importance of the selected features. **Figure 6** shows the dependence of the toxicity values on different patterns of distribution of the important descriptors $\Delta H_f$ and SD_Activity.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2022 doi:10.20944/preprints202210.0402.v1

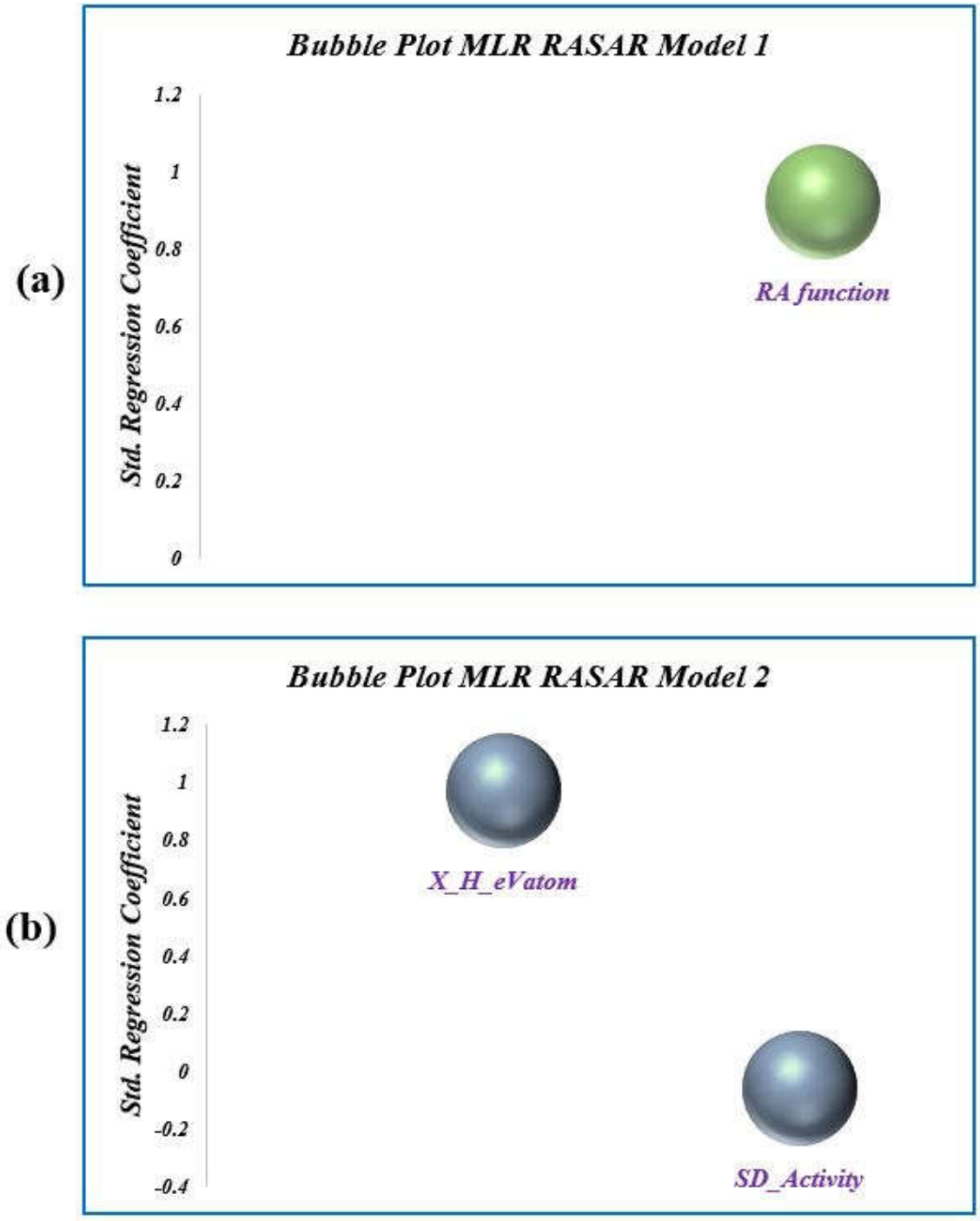


**Figure 3.** Bubble plots representing a) MLR RASAR Model 1 and b) MLR RASAR Model 2 (X_H_eVatom corresponds to $\Delta H_f$).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2022 doi:10.20944/preprints202210.0402.v1

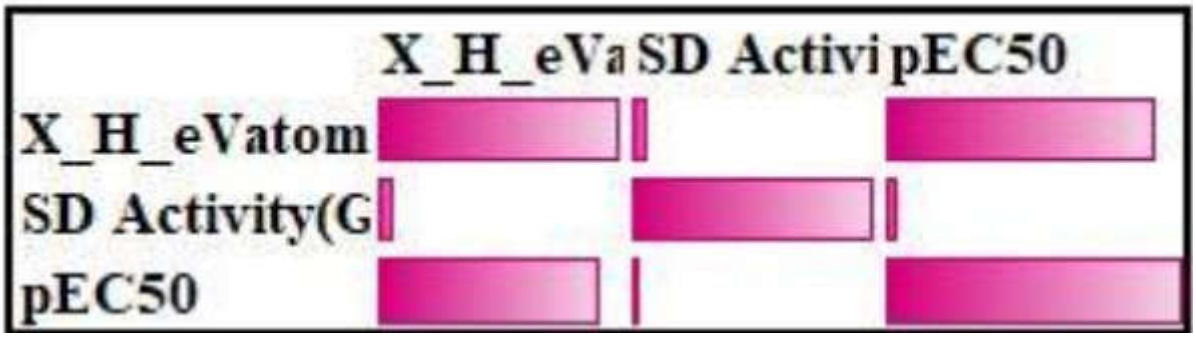


**Figure 4.** Heat map demonstrating the inter-correlation among descriptors in MLR RASAR Model 2 (X_H_eVatom corresponds to $\Delta H_f$).

**Table 2.** Validation metrics of the generated q-RASAR models.

| RASAR Models | $R^2$ | $Q^2_{(LOO)}$ | $Q^2_{F1}$ | $Q^2_{F2}$ |
|---|---|---|---|---|
| MLR RASAR Model 1 | 0.85 | 0.82 | 0.87 | 0.87 |
| MLR RASAR Model 2 | 0.91 | 0.88 | 0.91 | 0.91 |
| RF RASAR | 0.97 | - | 0.93 | 0.93 |

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2022 doi:10.20944/preprints202210.0402.v1

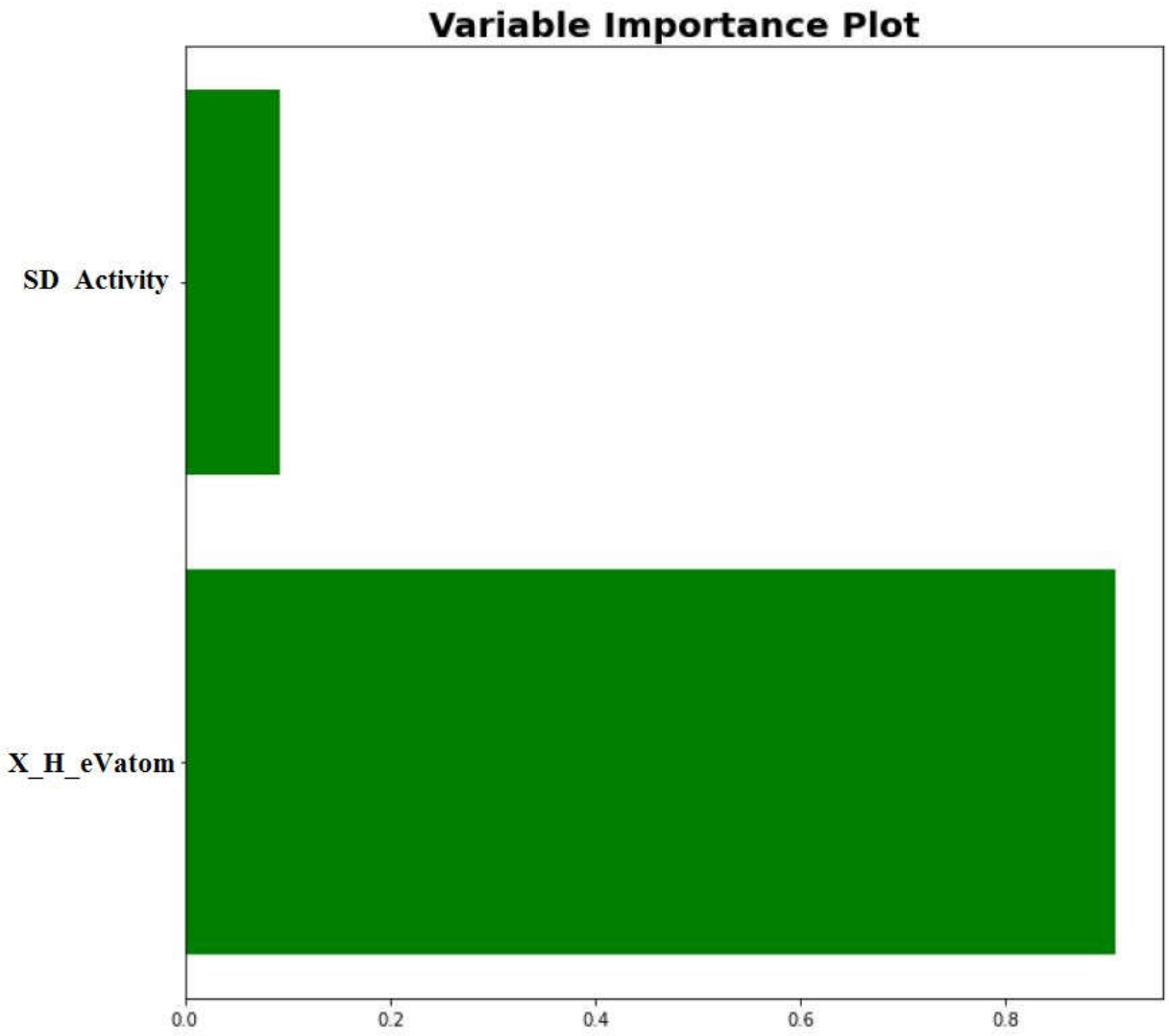


**Figure 5.** Variable Importance Plot for RF RASAR Model (X_H_eVatom corresponds to $\Delta H_f$).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2022 doi:10.20944/preprints202210.0402.v1

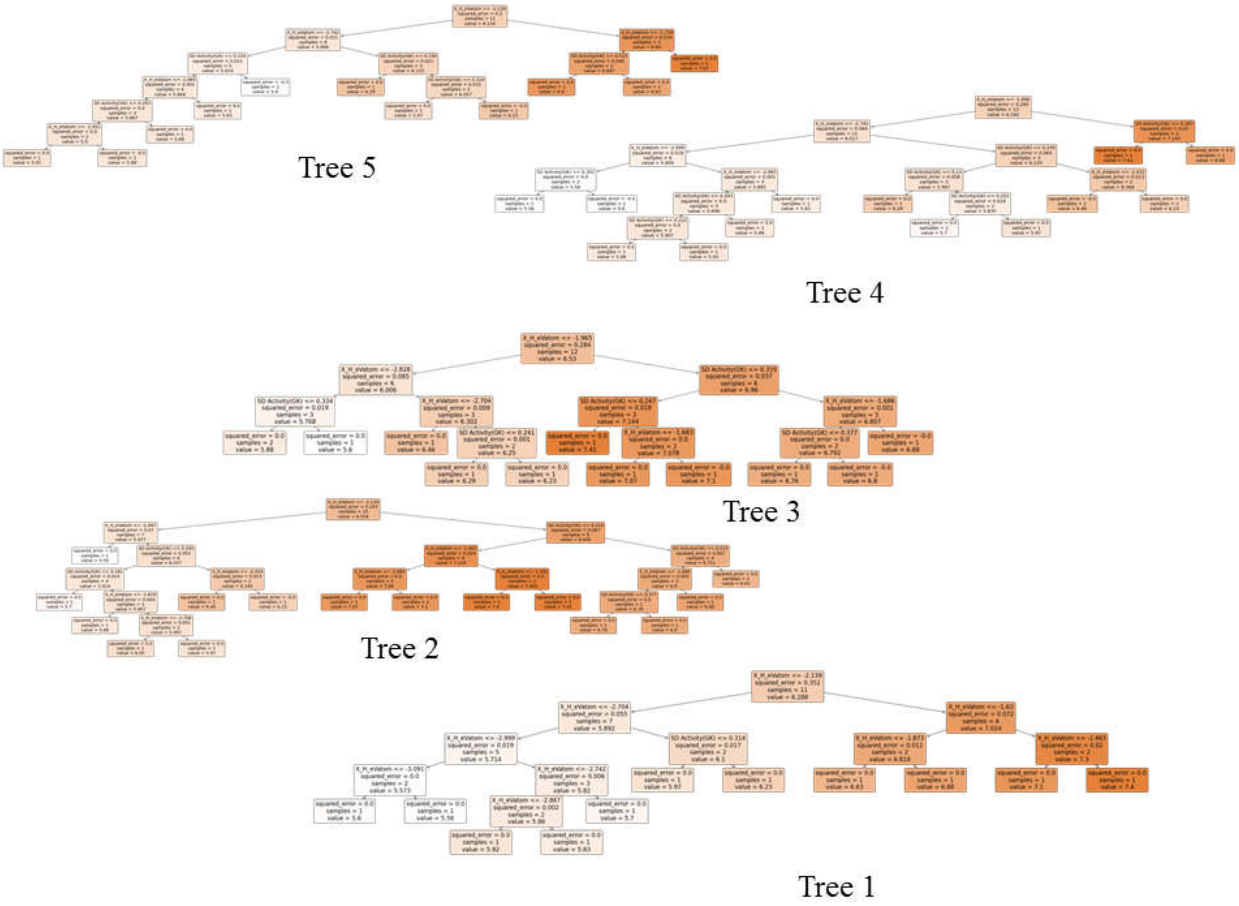


**Figure 6.** Sample Trees generated in the Random Forest model development (X_H_eVatom corresponds to $\Delta H_f$).

*3.3. Interpretation of the descriptors*

As reported in the previous work, the descriptor 'RA function' is a Read-Across-derived RASAR descriptor that encodes properties of all the electronic descriptors (20). Since it is a composite score of all the features responsible for eliciting toxicity, this descriptor contributes positively to the developed model in MLR RASAR Model 1. This can be exemplified by compounds like 4.5Ag@$TiO_2$ (**28**) for which both the RA function and $pEC_{50}$ values are high, whereas compounds like 0.1Ag@$TiO_2$ (**2**) have a low RA function value and thus possess a low $pEC_{50}$ value. The descriptor **$\Delta H_f$** in MLR RASAR Model 2 represents the formation energy concerning the constituent elements of the multicomponent $TiO_2$-based nanomaterials and this contributes positively towards the toxicity of hybrid $TiO_2$ nanomaterials. Higher formation energy implies a greater tendency of the constituent metal ions to exist in their free state (the metal cation). Due to their small size and ability to depolarize the membrane, these metal ions readily cross the biological membranes through the formation of holes and thus induce toxicity (26). This is observed in the case of compounds like 2.5Ag_0.5Pt@$TiO_2$ (**26**) where there is a higher formation energy value and thus increased toxicity as compared to compounds like 0.1Ag_0.5Pd@$TiO_2$ (**4**) where the toxicity value is lower due to the reduced formation energy. The descriptor **SD_Activity** in MLR RASAR Model 2 signifies the weighted standard deviation of the response values of the close source compounds for a particular query compound, which contributes negatively to toxicity. A higher standard deviation of the activity values of the close source compounds indicates dispersion among the values. Moreover, a higher MaxPos value and a higher SD_Activity value signify that the other (positive) close source compounds have a low similarity level with the query compound. This can be exemplified in compounds like 4.5Ag@$TiO_2$ (**28**) where the SD_Activity value is low, but the toxicity

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2022 doi:10.20944/preprints202210.0402.v1

value is high whereas in compounds like 0.1Ag@$TiO_2$ (**2**) the SD_Activity value is higher, while it possesses a much lower toxicity value.

*3.4. Comparison with previous work*

We have developed simple, interpretable, transferable, and reproducible MLR RASAR models which are highly robust and predictive and involve simple electronic descriptors and similarity and error-based RASAR descriptors. In the previous work, Mikolajczyk et al. (11) adopted both genetic algorithm-based multiple linear regression and Decision Tree models to predict cytotoxicity. Although their results are sufficiently good, our simple MLR RASAR Models supersede their models in terms of quality metrics and provide better predictivity and interpretability. Later, the generated RF RASAR model was even better in external predictivity compared to both MLR RASAR models. The complete comparison of the validation metrics has been presented in **Table 3.**

**Table 3.** Comparison of the quality of the models*.

| Models | $R^2_{adj}$ | $Q^2_{bagging}$ | $Q^2_{(LOO)}$ | $Q^2_{EXT}$ | $RMSE_p$ | $MAE_{CV}$ |
|---|---|---|---|---|---|---|
| **Previous work** (Mikolajczyk et al. 2019) (11) | | | | | | |
| MLR-GA | 0.87 | 0.84 | - | 0.80 | 0.19 | - |
| DT | 0.90 | 0.74 | - | 0.90 | 0.16 | - |
| **Our work** | | | | | | |
| MLR RASAR Model 1 | 0.84 | - | 0.82 | 0.87 | 0.19 | - |
| MLR RASAR Model 2 | 0.90 | - | 0.88 | 0.91 | 0.15 | - |
| **RF RASAR** | **0.96** | **-** | **-** | **0.93** | **0.13** | **0.36** |

*The best model is shown in bold.

## 4. Conclusion

This study reports simple, interpretable, transferable, highly robust, and highly predictive MLR q-RASAR models which can efficiently predict the cytotoxicity exerted by hybrid $TiO_2$ nanoparticles. These models explain that electronic properties like additive electronegativity, electron affinity, surface area, and formation energy for the constituent elements of the multicomponent $TiO_2$-based nanomaterials contribute to the cytotoxic properties of the nanomaterials. Apart from these electronic parameters, similarity-based descriptors like SD_Activity and RA function play an important role in estimating cytotoxicity. The optimization of the Read-Across hyperparameters was done using a newly developed Java-based software tool Auto_RA_Optimizer-v1.0 that provides the validation metrics of various combinations of the hyperparameters. Using the optimized setting of the hyperparameters, the RASAR descriptors were calculated based on the Leave-same-out (LSO) algorithm for the training set. When the training set is taken as the query set input file for calculating the RASAR descriptors of the training set, the tool RASAR-Desc-Calc-v2.0 automatically identifies the identical source and target compounds and does not consider that data point while calculating the RASAR descriptors. When the test set is taken as the query set input file, no such elimination of compounds occurs as there are no identical source compounds for the given set of target compounds. This approach reduces overfitting in the overall internal validation of the training set and reflects the true internal validation. The inter-correlation of the descriptors in the bi-variate RASAR model (RASAR Model 2) has been checked, and it was found that there is very low inter-correlation among the descriptors $\Delta H_f$ and SD_Activity which is demonstrated in the form of a heat map. This study also demonstrates how a simple bi-variate MLR q-RASAR model (RASAR Model 2) supersedes the predictive ability of the Decision Tree based machine learning model and MLR-GA model which were reported previously by Mikolajczyk et

 doi:10.20944/preprints202210.0402.v1

al. (11). The univariate q-RASAR model (RASAR Model 1) also supersedes the predictive ability of the previously developed MLR-GA model. Additionally, we have also developed a Random Forest regression model (RF RASAR Model) using the descriptors from the bi-variate MLR RASAR Model 2, and the results of the predictions showed the highest $Q^2_{ext}$ value among all the previously developed models. Moreover, our RF RASAR Model showed the lowest value of RMSEp among all the developed models. Thus, we can infer that the RF RASAR Model is the best suited for predicting the cytotoxicity of the hybrid $TiO_2$ nanoparticles as it is highly predictive. In summary, the q-RASAR approach may be a very potential approach for predicting the toxicity of engineered metal oxide nanoparticles and bridging the toxicity data gaps.

**Supplementary Materials:**The list of training and test set compounds along with important physicochemical and RASAR descriptors and model derived predictions are provided in Supplementary Materials SI-1.

**Author Contributions:** AB (Computation, model development, initial draft); SK (Computation, editing); AGS (Data collection, editing); KR (Conceptualization, supervision, funding, editing).

**Funding:** Life Science Research Board, DRDO, New Delhi, India.

**Acknowledgments:** AB thanks the Life Science Research Board, DRDO, New Delhi for a senior research fellowship. SK wants to thank the administration of Dorothy and George Hennings College of Science, Mathematics and Technology (HCSMT) of Kean University for providing research opportunities and resources. The authors thank Souvik Pore (DTC Lab, JU) for organizing the python codes for the Random Forest implementation in Jupyter Notebook.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1) De, P.; Kar, S.; Roy, K.; Leszczynski, J. Second generation periodic table-based descriptors to encode toxicity of metal oxide nanoparticles to multiple species: QSTR modeling for exploration of toxicity mechanisms. Environ. Sci: Nano. 2018, 5, 2742-2760. https://doi.org/10.1039/C8EN00809D.
2) Chavali, M.S.; Nikolova, M.P. Metal oxide nanoparticles and their applications in nanotechnology. SN Appl. Sci. 2019, 1, 607. https://doi.org/10.1007/s42452-019-0592-3.
3) Roy, J.; Roy, K. Risk assessment and data gap filling of toxicity of metal oxide nanoparticles (MeOx NPs) used in nanomedicines: a mechanistic QSAR approach. Environ. Sci: Nano. 2022, 9, 3456-3470. https://doi.org/10.1039/D2EN00303A.
4) Roy, J.; Roy, K. Assessment of toxicity of metal oxide and hydroxide nanoparticles using the QSAR modeling approach. Environ. Sci: Nano. 2021, 8, 3395-3407. https://doi.org/10.1039/D1EN00733E.
5) Roy, J.; Roy, K. Modeling and mechanistic understanding of cytotoxicity of metal oxide nanoparticles (MeOxNPs) to Escherichia coli: categorization and data gap filling for untested metal oxides. Nanotoxicology. 2022, 16(2), 152-164. https://doi.org/10.1080/17435390.2022.2038299.
6) Gajewicz, A.; Rasulev, B.; Dinadayalane, T.C.; Urbaszek, P.; Puzyn, T.; Leszczynska, D.; Leszczynski, J. Advancing risk assessment of engineered nanomaterials: application of computational approaches. Adv. Drug Deliv. Rev. 2012, 64(15), 1663–1693. https://doi.org/10.1016/j.addr.2012.05.014.
7) Sengul, A.B.; Asmatulu, E. Toxicity of metal and metal oxide nanoparticles: a review. Environ. Chem. Lett. 2020, 18, 1659–1683. https://doi.org/10.1007/s10311-020-01033-6
8) Karlsson, H.L.; Toprak, M.S.; Fadeel, B. Chapter 4 - Toxicity of metal and metal oxide nanoparticles, In: Nordberg, G.F.; Costa, M. (eds), Handbook on the Toxicology of Metals (Fifth Edition), Academic Press, NY, 2022. https://doi.org/10.1016/B978-0-12-823292-7.00002-4.
9) Acharya, S.; Sahoo, S.K. PLGA nanoparticles containing various anticancer agents and tumour delivery by EPR effect. Adv. Drug Deliv. Rev. 2011, 63(3), 170-183. https://doi.org/10.1016/j.addr.2010.10.008.
10) Manuja, A.; Kumar, B.; Kumar, R.; Chhabra, D.; Ghosh, M.; Manuja, M.; Brar, B.; Pal, Y.; Tripathi, B.N.; Prasad, M. Metal/metal oxide nanoparticles: Toxicity concerns associated with their physical state and remediation for biomedical applications. Toxicol. Rep. 2021, 8, 1970-1978. https://doi.org/10.1016/j.toxrep.2021.11.020.
11) Mikolajczyk, A.; Sizochenko, N.; Mulkiewicz, E.; Malankowska, A.; Rasulev, B.; Puzyn, T. A chemoinformatics approach for the characterization of hybrid nanomaterials: safer and efficient design perspective. Nanoscale. 2019, 11, 11808-11818. https://doi.org/10.1039/C9NR01162E.
12) Varnek, A.; Baskin, I. Machine learning methods for property prediction in chemoinformatics: Quo Vadis? J. Chem. Inf. Model. 2012, 52(6), 1413-1437. https://doi.org/10.1021/ci200409x.
13) Roy, K.; Kar, S.; Das, R.N. Understanding The Basics Of QSAR For Applications In Pharmaceutical Sciences And Risk Assessment, Elsevier Inc, NY, 2015, https://doi.org/10.1016/C2014-0-00286-9.

14) García-Fernández, A.J. Ecotoxicological Risk Assessment in the Context of Different EU Regulations. In: Roy, K. (eds) Ecotoxicological QSARs. Methods in Pharmacology and Toxicology. Humana, New York. 2020. https://doi.org/10.1007/978-1-0716-0150-1_1.
15) Kovarich, S.; Ceriani, L.; Gatnik, M.F.; Bassan, A.; Pavan, M. Filling data gaps by read-across: a mini review on its application, developments and challenges. Mol. Inform. 2019, 38(8-9), e1800121. https://doi.org/10.1002/minf.201800121.
16) Chatterjee, M.; Banerjee, A.; De, P.; Gajewicz, A.; Roy, K. A novel quantitative read-across tool designed purposefully to fill the existing gaps in nanosafety data. Environ. Sci.: Nano. 2022, 9, 189-203. https://doi.org/10.1039/D1EN00725D.
17) Banerjee, A.; Chatterjee, M.; De, P.; Roy, K. Quantitative predictions from chemical read-across and their confidence measures. Chemom. Intell. Lab. Syst. 2022, 227, 104613. https://doi.org/10.1016/j.chemolab.2022.104613.
18) Luechtefeld, T.; Marsh, D.; Rowlands, C.; Hartung, T. Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility. Toxicol. Sci. 2018, 165(1), 198-212. https://doi.org/10.1093/toxsci/kfy152.
19) Banerjee, A.; Roy, K. First report of q-RASAR modeling toward an approach of easy interpretability and efficient transferability. Mol. Divers. 2022, 26, 2847-2862. https://doi.org/10.1007/s11030-022-10478-6.
20) Banerjee, A.; De, P.; Kumar, V.; Kar, S.; Roy K. Quick and Efficient Quantitative Predictions of Androgen Receptor Binding Affinity for Screening Endocrine Disruptor Chemicals Using 2D-QSAR and Chemical Read-Across. Chemosphere. 2022, 309(1), 136579. https://doi.org/10.1016/j.chemosphere.2022.136579.
21) Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD), JOM 2013, 65, 1501-1509. https://doi.org/10.1007/s11837-013-0755-4
22) Kirklin, S.; Saal, J.E.; Meredig, B.; Thompson, A.; Doak, J.W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies, npj Computational Materials 2015, 1, 15010. https://doi.org/10.1038/npjcompumats.2015.10
23) Golbraikh, A.; Tropsha, A. Beware of q2!. J. Mol. Graphics Model. 2002, 20(4), 269-276. https://doi.org/10.1016/S1093-3263(01)00123-1.
24) Breiman, L. Random Forests. Machine Learning. 2001, 45, 5–32. https://doi.org/10.1023/A:1010933404324.
25) Kluyver, T.; Ragan-Kelley, B.; Pérez, F.; Granger, B.E.; Bussonnier, M.; Frederic, J.; Kelley, K.; Hamrick, J.B.; Grout, J.; Corlay, S.; Ivanov, P. Jupyter Notebooks - A publishing format for reproducible computational workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas, F. Loizides and B. Schmidt, Eds., IOS Press, 87–90. http://dx.doi.org/10.3233/978-1-61499-649-1-87.
26) Roy, J.; Roy, K. Nano-read-across predictions of toxicity of metal oxide engineered nanoparticles (MeOx ENPS) used in nanopesticides to BEAS-2B and RAW 264.7 cells. Nanotoxicology. 2022, https://doi.org/10.1080/17435390.2022.2132887 .