Preprint version dated 16.11.2022
NOR PEER REVIEWED

A machine learning q-RASPR approach for efficient predictions of the specific surface area of perovskites

Arkaprava Banerjee[a], Agnieszka Gajewicz-Skretna[b], Kunal Roy[a*]

[a]Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India

[b]Laboratory of Environmental Chemoinformatics, Faculty of Chemistry, University of Gdansk, Wita Stwosza 63, 80-308 Gdansk, Poland

***Correspondence to: Kunal Roy (**kunal.roy@jadavpuruniversity.in**)**

Abstract

In this study, the specific surface area of various perovskites was modeled using a novel quantitative read-across structure-property relationship (q-RASPR) approach, which clubs both Read-Across (RA) and quantitative structure-property relationship (QSPR) together. After optimization of the hyper-parameters, certain similarity-based error measures for each query compound were obtained. Clubbing some of these error-based measures with the previously selected features along with the Read-Across prediction function, a number of machine learning models were developed using Partial Least Squares (PLS), ridge regression (RR), linear support vector regression (LSVR), and random forest (RF) regression. Based on the external prediction quality and interpretability, the PLS model was selected as the best predictor which underscored the previously reported results. The finally selected model should efficiently predict specific surface areas of other perovskites for their use in photocatalysis. The new q-RASPR method also appears promising for the prediction of several other property endpoints of interest in materials science.

Keywords: q-RASPR; Machine learning; Perovskites; Specific surface area; Photocatalysis

Introduction

Perovskites are inorganic chemical substances that are structurally similar to $CaTiO_3$, or in other terms; they have a general structural formula of ABO_3, where the B-ions are surrounded by octahedrally arranged O ions[1-6]. In such compounds, the A-site represents a rare-earth element like La^{3+}, whereas the B-site represents transition elements like Fe^{3+}, etc.[7]. Broadly, perovskites can be classified into three main categories: alkaline metal halide perovskites, inorganic oxide perovskites, and organic metal halide perovskites with oxide or halide anions[8]. Although there are various types of perovskites found on the Earth's crust, $FeSiO_3$ and $MgSiO_3$ are known to be the most abundant[5]. The concept of doping is the mixing of other ions to the composite structure of perovskites to achieve certain properties for their applicability as biosensors, catalysts, and conductors[5, 9]. In general, there is a wide array of applications of perovskite materials in biological, chemical, and electronic sciences. An important application of perovskites in chemical industries is their use as catalysts. They provide a solid-state surface and show increased catalytic activity in reactions like reduction and hydrogen/oxygen evolution[5]. Perovskites are also used as sensors like gas sensors, where they are extensively used due to their high thermal stability[5]. They are also used as glucose sensors in clinical and pharmaceutical analysis and have effectively replaced enzymatic substances due to their inherent increased stability. Also, perovskites are used as modified surfaces in neurotransmitter sensors to selectively detect neurotransmitters like Dopamine, even in the presence of interfering substances like Ascorbic acid and Uric acid[5]. Apart from these biological applications, they are also used in fuel cells due to their high electrical and ionic conductivity. Perovskites are also used in solar cells which convert solar energy into electrical energy. The disadvantage of the use of silicon in solar cells is the high price of the electricity generated, and thus with a view of reducing the cost, various organic and inorganic perovskites have been used in solar cells[5].

Researchers are now inclined towards using various machine learning (ML) algorithms for the prediction of activity/property/toxicity of various materials and chemicals. ML algorithms produce fast, reliable, and accurate results and involve limited manpower. One of the most widely used machine learning applications is Quantitative Structure-Property Relationship (QSPR) which predicts certain properties of a set of compounds under a given experimental condition. QSPR results are accepted by regulatory bodies like EU-REACH[10]

(https://echa.europa.eu/regulations/reach/understanding-reach) for data gap filling. The basic algorithm involved in a QSPR methodology is the development of models consisting of one or more dependent variables (properties) and one or more independent variables (features/descriptors) which contribute to the dependent variable(s) and are expressed in numerical terms[11]. These models are derived after training a known set of compounds and are then used to predict an external set of query chemicals and thus, follow a supervised machine learning algorithm. The QSPR models may involve simple regression models like multiple linear regression (MLR) and Partial Least Squares (PLS) regression models or classification models like linear discriminant analysis (LDA). However, more sophisticated machine learning approaches like Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Random Forest (RF) have gained popularity in recent times for predicting the response values of query chemicals[12]. Another recent trend is to adhere to "similarity-based" approaches like Chemical Read-Across[13]. It is performed by interpolation or extrapolation of the properties of a set of source compounds to one or more target compounds based on the similarity values between the source and the target compounds[13, 14] that demonstrates an unsupervised machine learning algorithm. Read-Across-based predictions are based on (chemical or biological) similarity to close congeners, and these are not model-derived predictions. Both the QSPR and Read-Across approaches are extensively used for data gap filling (predicting activity/property/toxicity values of compounds devoid of experimentally derived endpoint values). Although sufficiently predictive, many of the machine learning approaches lack interpretability and act like a black box. Recently, Luechtefeld et al.[15] introduced the concept of classification-based Read-Across Structure-Activity Relationship (RASAR) by combining the concepts of Read-Across and QSAR using machine learning algorithms. More recently, Banerjee and Roy[16] combined Chemical Read-Across and regression-based 2D-QSAR and named it Quantitative Read-Across Structure-Activity Relationship (q-RASAR). In this technique, a wide array of similarity and error-based measures are calculated for each query compound, and these measures are used as descriptors to develop simple and interpretable q-RASAR models.

In the recent past, perovskites have been one of the centers of attraction for researchers working in diverse scientific fields. Shi et al.[7] worked on the prediction of the specific surface area of perovskites to demonstrate their catalytic activity. They have adopted three different machine learning algorithms namely Partial Least Squares (PLS), Artificial Neural Network (ANN) and

Support Vector Regression (SVR) to identify the features associated with the increase or decrease of the specific surface area of perovskites and used them to predict an external set of compounds. Zhang and Xu[17] worked on predicting the oxygen ionic conductivities of perovskites of type ABO3 using a Gaussian process regression model. Wang et al.[18] utilized Random Forest (RF) to predict the photon energies associated with quasi-2D perovskite materials and their precursors. Kim et al.[19] combined density functional perturbation theory and machine learning to estimate the dielectric constant of perovskites. Lyu et al.[20] adopted ML approaches on lead iodide-based perovskites to predict their dimensionality. Yan et al.[21] also used ML approaches to predict five unexplored perovskites with low bandgap, short circuit current density, and open circuit voltage for the design of highly efficient perovskite solar cells.

The catalytic activity of perovskites is directly proportional to the specific surface area presented by it. In this work, we have predicted the specific surface area of perovskites using a novel Quantitative Read-Across Structure-Property Relationship (q-RASPR) algorithm, which combines the advantages of QSPR and Read-Across approaches. This novel method has been derived from the works of Banerjee and Roy[16] on the development of regression-based q-RASAR models for the first time. After dataset division, feature selection, and hyperparameter optimization, the training and test sets were subjected to Chemical Read-Across-based predictions. The optimized setting of the Read-Across-based predictions was used to calculate the RASPR descriptors. Using the similarity and error-based RASPR descriptors obtained in an ML-based approach, further feature selection was employed, and a final Partial Least Squares (PLS) model was generated which is robust, predictive, transferable, and reproducible. The performance of the developed PLS model was also compared with several other machine learning regression approaches (*vide infra*).

Materials and Methods

Collection of the specific surface area data for perovskites

The dataset containing specific surface areas of 50 data points was obtained from Shi et al.[7] and is provided in an excel sheet of **Supplementary Material SI-1**. This dataset contains a list of compounds, their specific surface area data, and a set of 24 features (descriptors). Out of the 24 descriptors, 21 depict structural information while 3 are experimental process variables.

Dataset Division and feature selection

In the absence of a true external set, the general practice in any predictive modeling is to divide the available dataset into training and test sets to explore the predictive ability of the algorithm. The current dataset was rationally divided into training and test sets of 38 and 12 data points, respectively, based on the sorted response-based division.

For the feature selection, the mean values of all the descriptors for the 10 compounds with the highest property values and 10 compounds with the least property values were calculated from the scaled (between 0 and 1) training set descriptor matrix. Additionally, the absolute differences between their mean values were computed. The descriptors with an absolute mean difference value > 0.15 were considered as the selected features.

Machine Learning-based Read-Across predictions

Read-Across is a similarity-based prediction approach that does not require, in its original form, the involvement of a statistically reliable model. The main limitation of a conventional QSPR model is that the model becomes unreliable when there is a limited number of data points with limited degrees of freedom for statistical fitting[13]. Since Read-Across, in its original form, is not a statistical approach, rather it is a similarity-based approach, it tends to yield better prediction results even for small datasets, and thus, can be a very useful tool for data gap filling. Read-Across-v4.1 available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home is a java-based software tool that takes the training and test set files and a few hyperparameters as inputs and quickly computes the Read-Across-based predictions based on similarity considerations with the Euclidean Distance-based approach, the Gaussian Kernel Similarity-based approach and the Laplacian Kernel Similarity-based approach. The tool also computes the corresponding validation metrics for predictions along with the compound-specific similarity and error-based measures for the confidence of predictions[13, 22]. The pre-requisite to perform the Read-Across-based predictions is to identify the optimized setting of the hyperparameters. The training set consisting of the selected features is further divided into sub-training and sub-test sets and these

are used to obtain the optimized number of close source compounds. The number of close source compounds yielding the maximum Q^2_{F1}, Q^2_{F2} values for the sub-test (or validation) sets for the Euclidean Distance-based predictions can be considered as the optimized number of close source compounds as shown in **Figure 1**. Using this information, the original training and test set files with the selected features are used as inputs in the Read-Across-v4.1 tool and a set of Read-Across-based predictions are obtained.

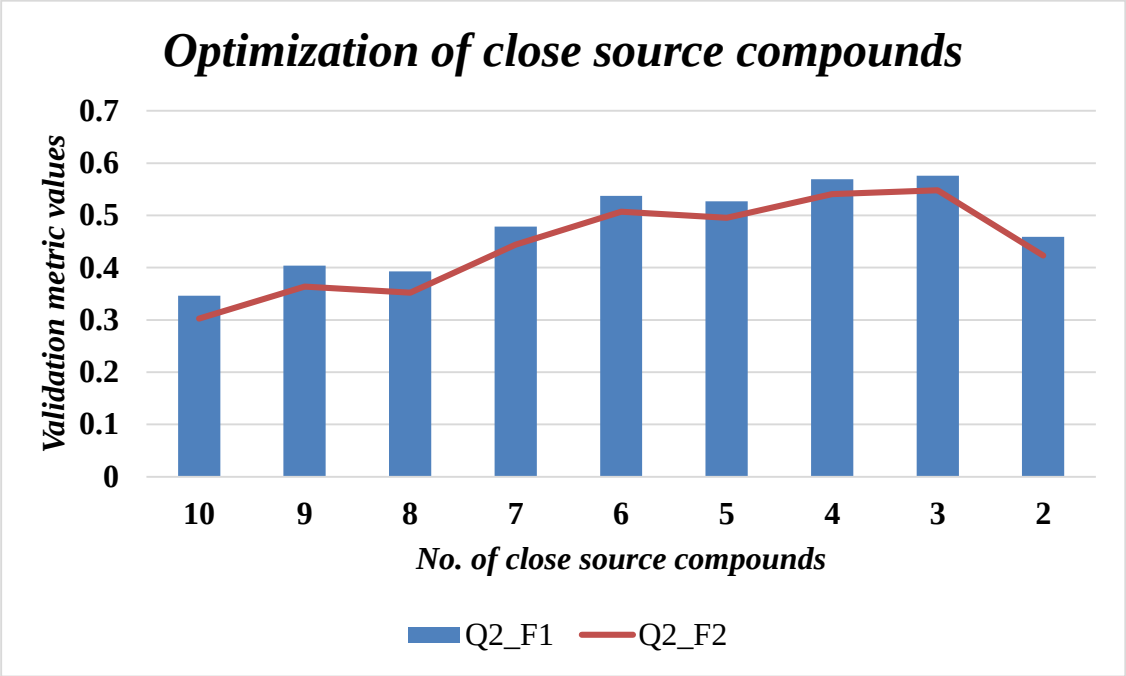

Figure 1. Demonstration of the validation metric values (validation sets) for different numbers of close source compounds

Development of novel q-RASPR models using Machine Learning algorithms

Read-Across Structure-Activity Relationship (RASAR) is an integrated approach that combines the concepts of Read-Across and QSAR and generates simple and transferable models. This approach was initially introduced by Luechtefeld et al.[15] who performed classification-based RASAR while Banerjee and Roy[16] performed it for quantitative predictions and named it q-RASAR. Since the present study deals with the properties of compounds, the approach has been

renamed q-RASPR, as it now combines Read-Across and QSPR. Utilizing the optimized settings for the hyperparameters and the similarity-based approach used for the final Read-Across-based predictions, we have computed the RASPR descriptors using the tool RASAR-Desc-Calc-v2.0[16] available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. This tool asks for the training set file and the query set file. The RASPR descriptors for the test set were calculated when the test set file was taken as input for the query set file, and the RASPR descriptors for the training set were calculated when the query set file corresponds to the training set itself. It is essential to note that while computing the training set RASPR descriptors, a Leave-Same-Out (LSO) strategy was adopted. The program automatically identifies the identical compounds in the close "n" source compounds using an in-built supervised machine learning algorithm and does not take the information of that compound while computing the RASPR descriptors in order to avoid bias. Once the RASPR descriptors were calculated for both the training and test sets, they were then clubbed with the previously selected structural descriptors and process variables to obtain a complete descriptor pool. This descriptor pool was further subjected to feature selection using the java-based tool BestSubsetSelection_v2.1 available from https://dtclab.webs.com/software-tools. This tool generates MLR models from all possible combinations of descriptors. The model with the best internal and external validation metrics, along with simple and interpretable descriptors was chosen. Finally, a PLS model which aims to nullify the inter-correlation among descriptors was generated using 3 latent variables and evaluated based on the computation of various internationally accepted internal and external validation metrics[23, 24].

In order to compare the performance of the developed PLS model with other regression-based machine learning algorithms, we also developed ridge regression (RR), linear support vector regression (LSVR) and random forest (RF) regression models using the same descriptor combinations as appeared in the final PLS RASPR model. Ridge regression (RR) estimates the coefficients of multiple-regression models by the application of L2 regularization[25]. Support vector regression (SVR) maps training examples to points in space in order to maximize the distance between the two classes[26]. Random forest (RF) is an ensemble of decision tree predictors used for classification or quantitative prediction purpose[27]. The optimization of hyperparameters (number of components in case of PLS regression, alpha in case of ridge regression, 'C' in case of linear LSVR, and the number of estimators with minimum number of samples for splitting in case of random forest regression) was done using the grid search option by the five-fold cross-validation

technique as shown in **Figure 2**. In the present study, machine learning modeling was carried out in Jupyter Notebook web application[28] in the Anaconda3 navigator version 2022.05 with Python version 3.10.4.

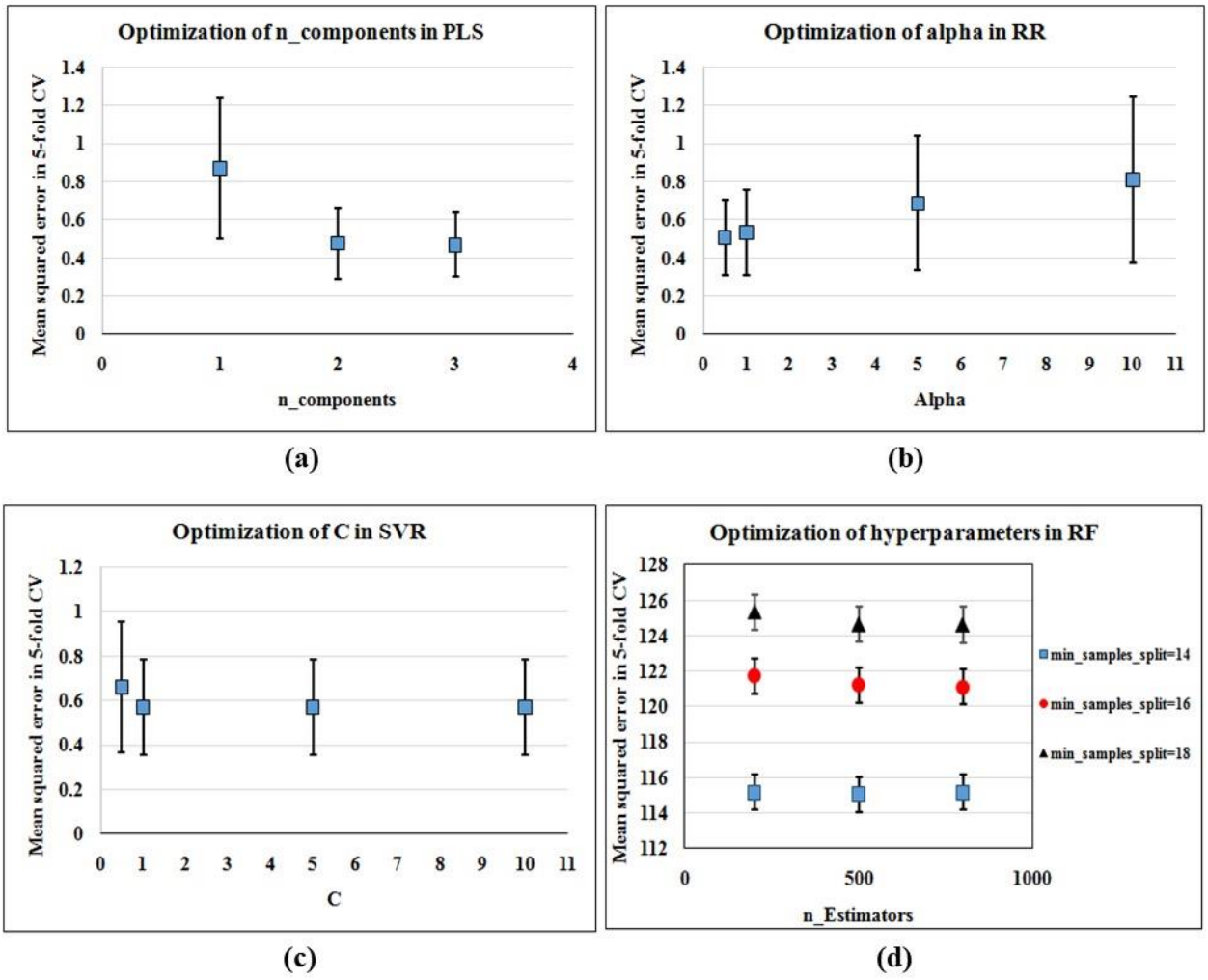

Figure 2. Optimization of hyperparameters by the grid search option for (a) PLS regression (b) Ridge Regression (c) SVR (d) Random Forest regression

PLS Plots and Y-Randomization test

The PLS plots were generated using Simca-P v10.0 available from https://www.sartorius.com/ and is made available in **Supplementary Material SI-2**. The **Score Plot**, whose axes represent the first two LVs, shows the allocation of the compounds in the LV space. Compounds lying inside the ellipse and are located closer to each other can be considered "similar", whereas compounds

lying outside the ellipse represent outliers. We have also checked the AD using the **DModX approach (Figure S1)** represented in the form of bar graphs. The **Regression Coefficient Plot (Figure S2)** depicts the contribution of different descriptors to the response value in terms of magnitude and direction (positive or negative). The **Loading Plot** establishes the relation between the dependent variable (Y) and the independent variables (X) of a PLS model. A longer distance of the independent variable from the origin is directly proportional to the importance of the descriptor[29]. The **Variable Importance Plot (Figure S3)** reflects the importance of the descriptors in the form of a bar graph, where the X-axis represents the descriptors and the Y-axis represents the importance[29]. The **Y-Randomization Plot (Figure S4)** tests the fact whether the generated PLS model is obtained by chance or not. Although both the X and Y variables can be permuted, we have performed here a permutation of the Y-column entries for 100 combinations. A Scatter Plot **(Figure 9)** of the observed v/s the predicted specific surface area (SSA) values was drawn to study the deviation of the predicted SSA values from the observed SSA values.

Prediction for a designed data set

We have generated a dataset of 450 data points to check the predictions of our developed q-RASPR model. We initially took 5 perovskite compounds with altered atomic fractions at the B-site as a result of doping and then calculated the atomic descriptors from the dataset already provided. The descriptors denoting the experimental conditions were kept constant for these 5 perovskite materials. The subsequent data points have been generated by altering the combination of these experimental descriptors to yield a designed set of 450 data points. The required RASPR descriptors for this query set of 450 data points have been generated by the java-based tool RASAR-Desc-Calc-v2.0 (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home). The generated q-RASPR model was used to predict the SSA values of these data points using the Prediction Reliability Indicator tool (PLS Version)[30]. The predicted values and their AD analysis have been provided in an excel sheet of **Supplementary Materials SI-1**.

The entire workflow of nano q-RASPR has been demonstrated pictorially in **Figure 3**.

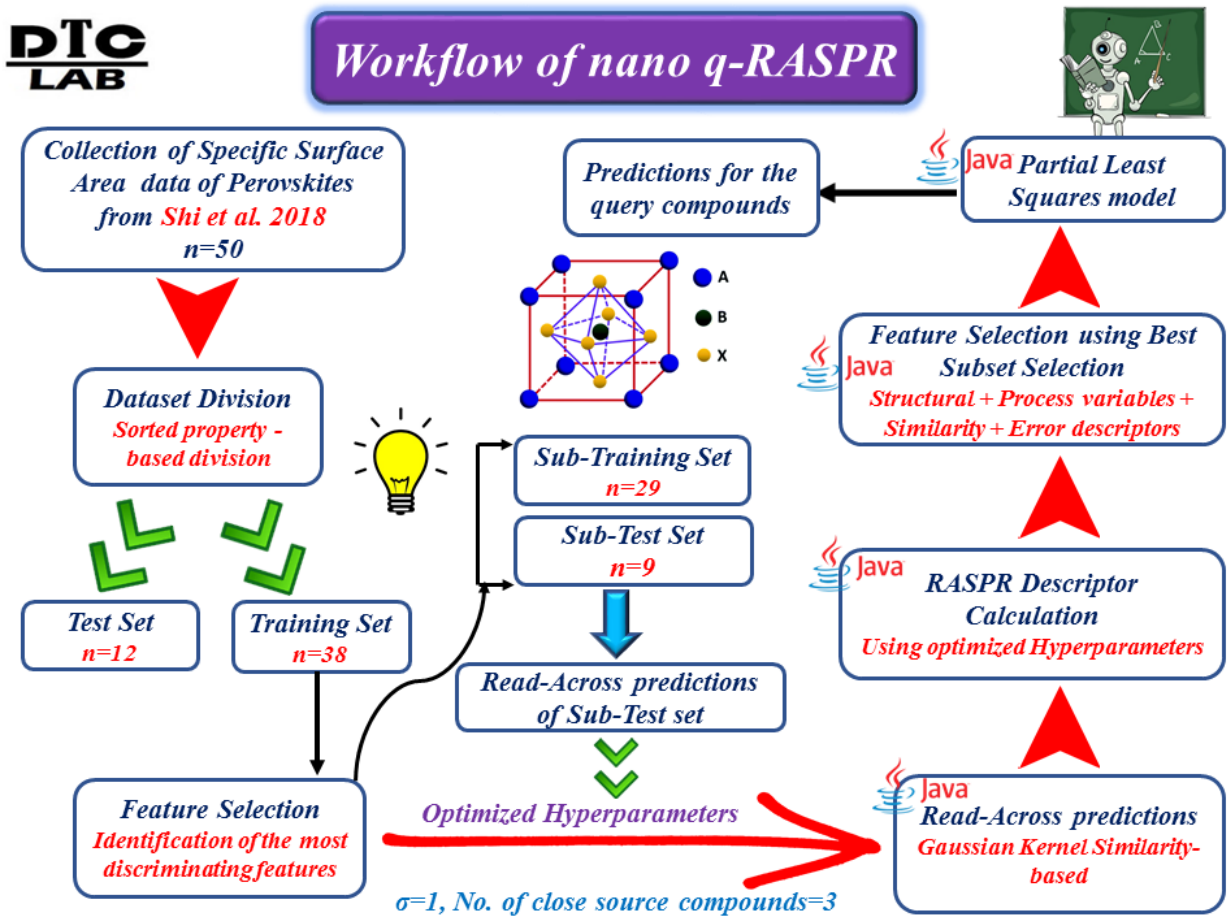

Figure 3. Workflow using the nano q-RASPR algorithm

Results and Discussion

Optimization of the number of close source compounds and identification of the best similarity-based approach for read-across predictions

It is very essential to identify the optimized setting of the hyperparameters in the well-defined machine-learning expressions for Read-Across and RASPR-based predictions. In cases where there is a structural diversity in the data set, a higher number of close source compounds generally improves the quality of predictions and vice versa. According to the principles of QSPR predictions, the optimization of the number of close source compounds should be done based on the training set only (supervised learning) without the involvement of a test/query set. Thus, the sub-training and sub-test sets were used to optimize the number of close source compounds based on the external validation metrics for the Euclidean Distance-based predictions. The predictions

were best when the number of close source compounds was 3. Using this optimized information, Read-Across-based predictions were obtained at σ=1 and γ=1 (default values) and it was found that the best prediction results ($Q^2_{F1} = 0.74$ and $Q^2_{F2} = 0.73$) were generated using the Gaussian Kernel similarity-based approach, derived from the concepts of Support Vector Machines (SVM). The RASPR descriptors were also calculated using the Gaussian Kernel Similarity-based approach at σ = 1 and number of close source compounds = 3.

q-RASPR Analysis

We have used a training set containing 38 data points for the development of the q-RASPR model while 12 data points were used for the prediction purpose. The generated PLS q-RASPR model has been represented graphically in **Figure 4** (standardized coefficients in the bubble plot with the size of the bubble proportional to VIP values) and the corresponding equation is shown in **Supplementary Material SI-2**. The descriptors that we have considered after feature selection for deriving the *RA function* (*vide infra*) in the q-RASPR model development have been reported in **Table 1**.

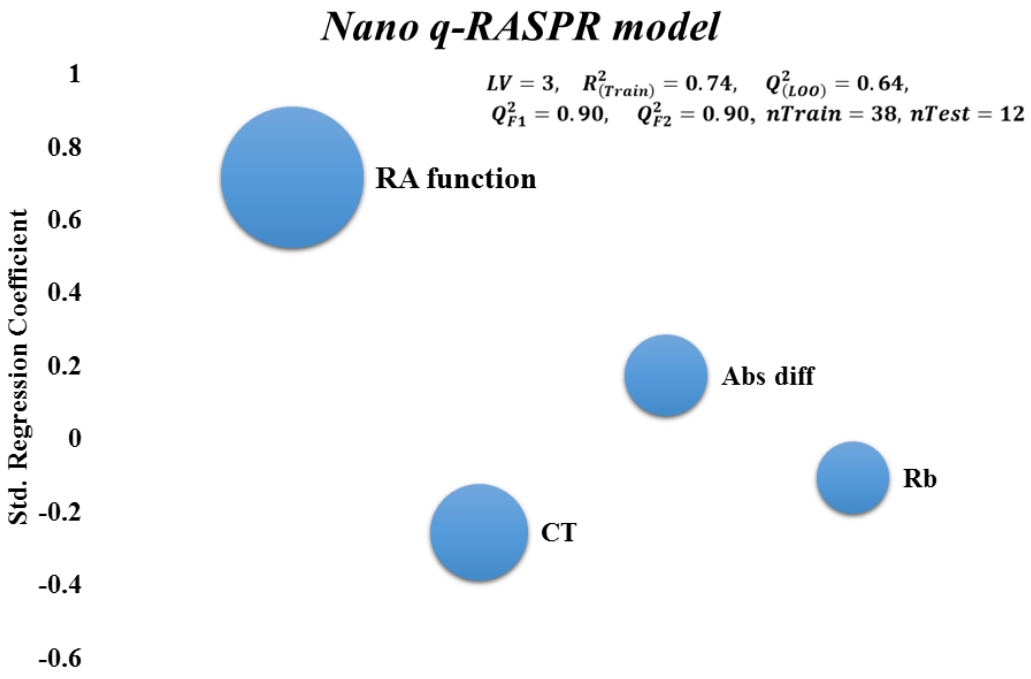

Figure 4. Graphical representation of the nano q-RASPR model (PLS) (size of the bubble proportional to VIP values)

Table 1: List of structural descriptors and process variables selected for q-RASPR model development

Descriptors	Significance
Rb	The radius of the atom at the B position
Ea	Electronegativity of the atom at the A position
aO3	Unit cell lattice edge
Mass	Molecular mass
A-Tm	Melting point of the atom at the A position
D-A	Density of the A position
CT	Calcination Temperature
AH	Calcination Time
DT	Drying Temperature

The developed q-RASPR model possesses an acceptable quality in terms of the internal validation metrics considering the heterogeneity of the dataset; however, the predictive ability of the model is very good as reflected in the values of the external validation metrics.

RA function – the Read-Across derived RASPR descriptor

The Read-Across function (*RA function*) is a descriptor derived using a machine learning algorithm by the tool RASAR-Desc-Calc-v2.0. Information from each of the structural descriptors and process variables is condensed in this single descriptor, and it acts similarly to a composite variable. This descriptor relates how close the query data point is to the mean response value of the close source data points. Since it uses all the selected structural features and process variables, it is expected that this descriptor should possess the highest importance as evident from the **Variable Importance Plot** (**Figure S3**). However, although it can have the same value for two different query data points, the individual distance/similarity of the close source data points might be different which explains why two or more query data points can have the same *RA function* value but possess different response values. This is the reason why other descriptors are also used while developing a q-RASPR model. **Figure 5** represents the algorithm involved in the *RA function* descriptor.

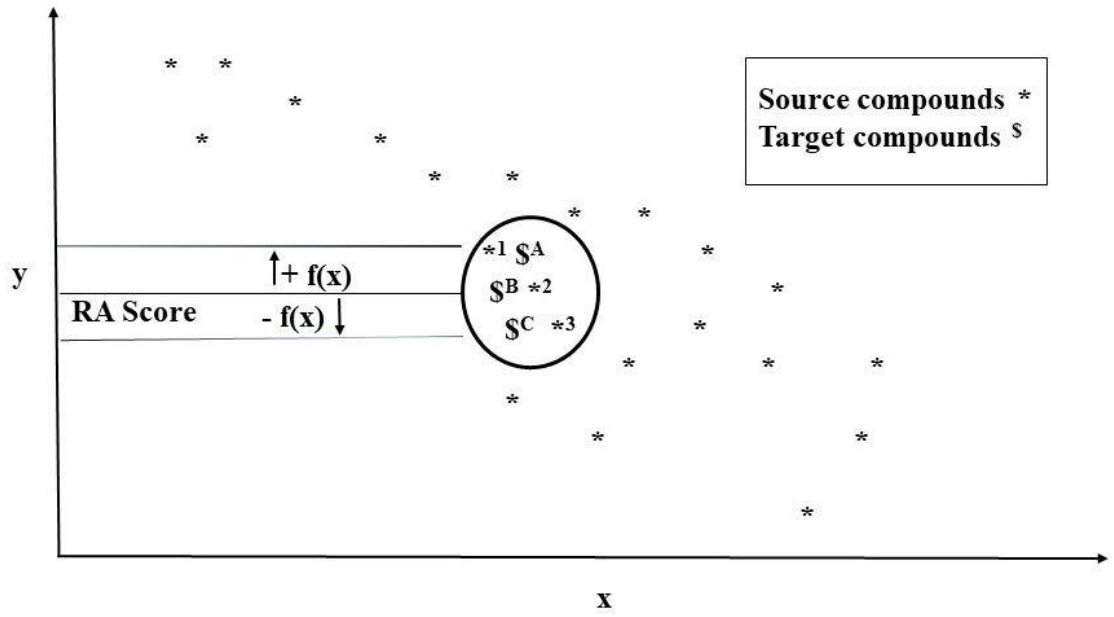

Figure 5. Pictorial representation of the RA function/RA Score

Interpretation of the descriptors

The descriptor *RA function*, as already discussed above, is a Read-Across derived descriptor which encodes information from all the individual structural descriptors and process variables (**Figure 6**). This descriptor contributes positively to the developed PLS model as it expresses the composite contribution of all selected descriptors. This can be observed in data points like $LaFe_{0.1}Co_{0.9}O_3$ (**27**) which has a high RA function and the corresponding specific surface area, while data points like $LaNiO_3$ (**40**) have a comparatively lower *RA function* value and consequently have a lower value of the specific surface area. The process variable *CT* represents the calcination temperature, and this descriptor contributes negatively to the developed model. An increase in the calcination temperature decreases the volume of reactants adsorbed thus leading to the reduced specific surface area[31]. This can be observed in $La_{0.5}Bi_{0.2}Ba_{0.2}Mn_{0.1}FeO_3$ (**38**) which has a high *CT* value but a very low specific surface area value while in data points like $LaMg_{0.4}Cr_{0.6}O_3$ (**22**), the *CT* value is lower but the value of the specific surface area is much higher. Another RASPR descriptor *Abs diff* (|*MaxPos-MaxNeg*|) signifies the absolute difference in the *MaxPos* and *MaxNeg* values[16,

[22]. Since a greater number of compounds in our training set has a higher *MaxPos* value than the *MaxNeg* value, the descriptor *Abs diff* expresses a positive contribution to the response. This can be exemplified using compounds like $La_{0.01}Sr_{0.99}TiO_3$ (**32**) where the *Abs diff* value as well as the specific surface area value is higher as compared to $LaNiO_3$ (**42**) which has a much lower *Abs diff* and specific surface area value. The descriptor *Rb* represents the atomic radius of the atom at the B position, and this contributes negatively to the developed model. As the specific surface area is defined as the total surface area of a substance per unit mass, an increase in the radius of the atom at the B-position increases the atomic mass, thereby reducing the specific surface area value. The compound $LaMg_{0.6}Cr_{0.4}O_3$ (**23**) has a higher *Rb* value and a lower specific surface area value than $La_{0.5}Bi_{0.2}Ba_{0.2}Mn_{0.1}FeO_3$ (**35**) which has a lower *Rb* value and a higher specific surface area value.

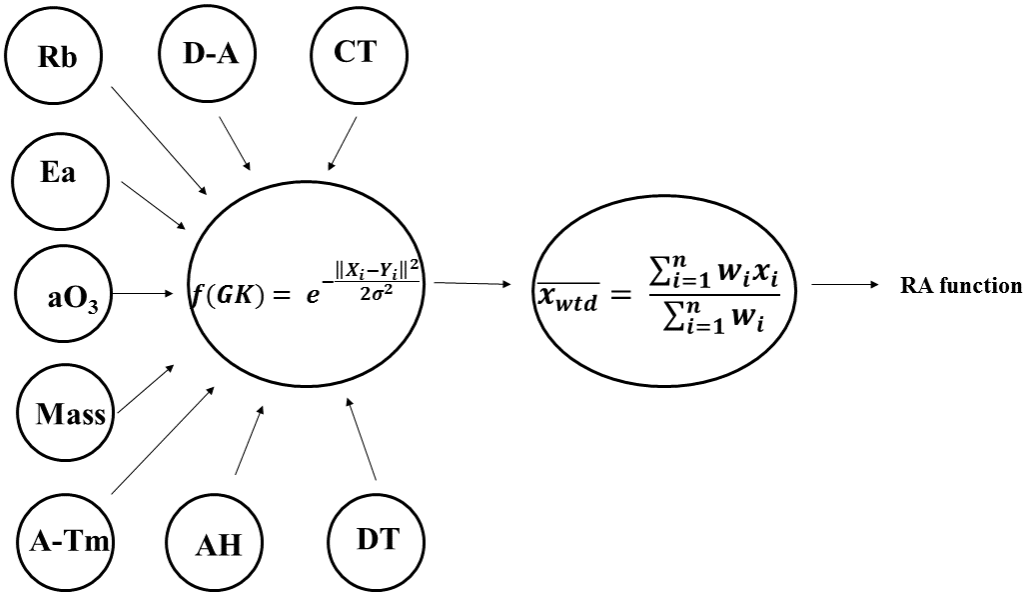

Figure 6. RA function as a composite of the atomic/structural descriptors and process variables

Interpretation of the PLS Plots

The information from the PLS plots reflects the quality, applicability domain, contribution of the descriptors towards the response and the chance factors involved in the developed PLS model. The **Score Plot** (**Figure 7**) depicts that only one data point (**27**) is outside the applicability domain and lies outside the ellipse while most of the other compounds have average properties and lies inside

the domain. Clusters are observed in the range of (-1, 1) and (1, 1) depicting the close similarity among the compounds. The **DModX Applicability Domain** plots (**Figure S1**) shows that all the DModX values are within the critical limit which signifies that all the compounds lie inside the applicability domain at LV=3. The **Regression Coefficient Plot** (**Figure S2**) signifies that the descriptors *RA function(GK)* and *Abs diff* (|*MaxPos-MaxNeg*|) contribute positively to the specific surface area values while the other descriptors Rb and CT contribute negatively to the specific surface area values. The **Loading Plot** (**Figure 8**) signifies that the descriptor *RA function(GK)* has the highest importance as it is located farthest from the origin. The **Variable Importance Plot** (**Figure S3**) depicts that the descriptor RA function(GK) has the highest importance and contributes the most to the prediction of specific surface area, while the descriptor *Rb* is the least important among the four descriptors and is depicted by the smallest bar. The **Y-Randomization Plot** (**Figure S4**) shows that our model is not obtained by chance. This is because the values of R^2_Y and Q^2_Y are well within the limits of 0.3 and 0.05 respectively[32]. The **Scatter Plot** (**Figure S5**) shows us that there is not much difference in the observed and predicted SSA values of the data points reflecting good-quality predictions for the test set.

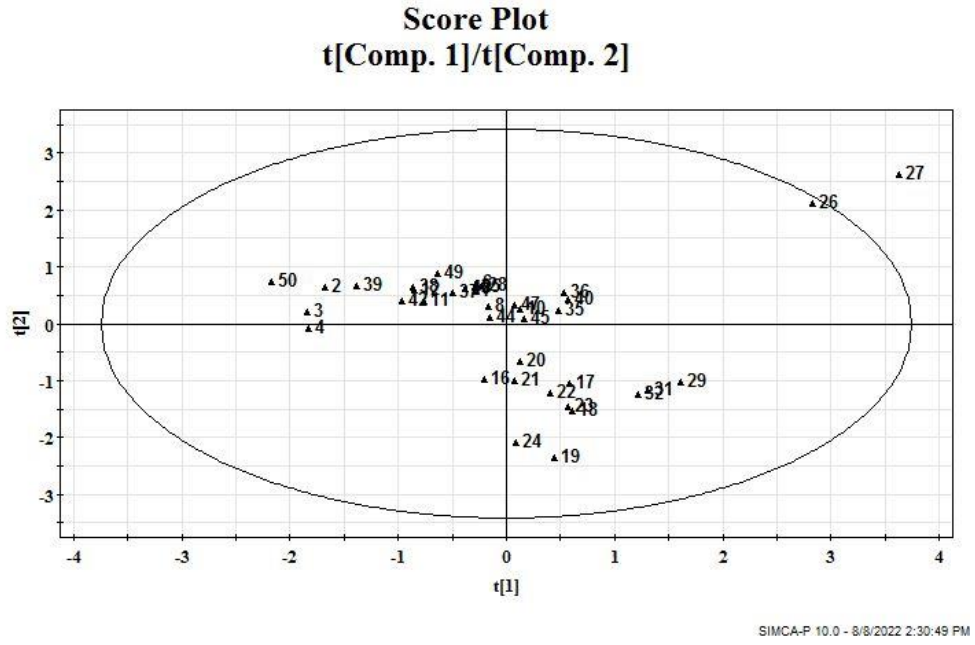

Figure 7. Score Plot representing the Applicability Domain

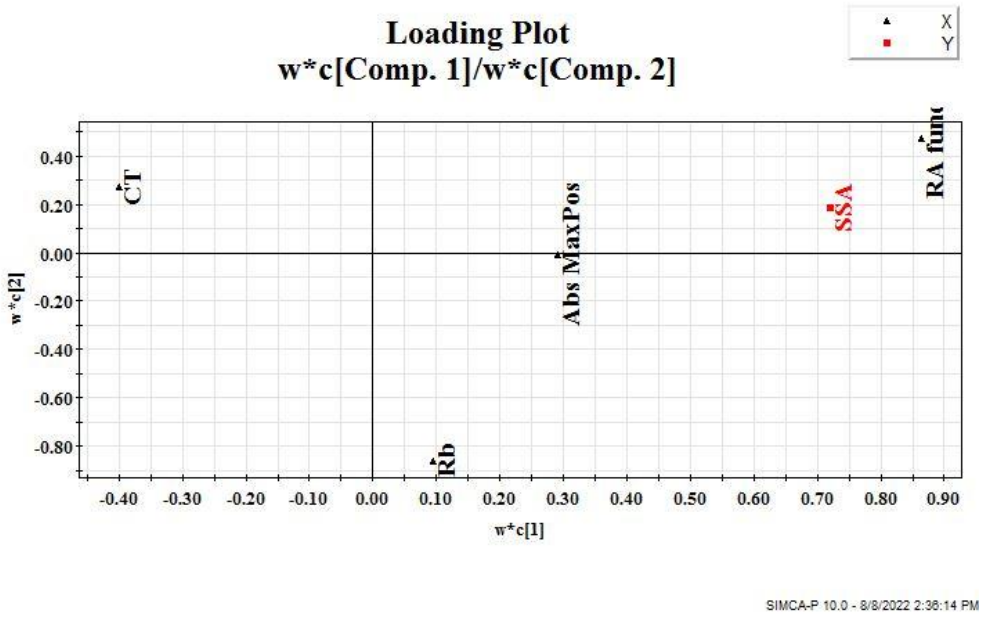

Figure 8. Loading Plot representing the contribution of the descriptors

Comparison of the prediction quality of the PLS model with other machine learning regression models

The prediction quality of all ML regression models for the test set compounds was almost comparable confirming the importance of the selected features including RASPR descriptors (**Table 2**). However, the Q^2_{F1} metric values of the RF regression and Linear SVR models were a little lower while those of the PLS and RR models were almost the same (0.901). We also considered a non-linear SVR model (with radial basis function) for this data set; however, the external prediction quality was poor and hence not reported here. Considering the interpretability of the final model in terms of different diagnostic plots like loading plot, score plot, etc., and also considering the results of five-fold cross-validation, we have considered the PLS model as the best predictor for the current data set.

Table 2. Comparison of the prediction quality of PLS model with other ML regression models

Model Type	R^2_{Train}	R^2_{Test}	$MAE_{(LOO)}$	$MSE_{(LOO)}$	MAE (5-fold CV)	MSE (5-fold CV)	Q^2_{F1}	Optimum hyperparameters
PLS*	**0.737**	**0.898**	**5.1**	**40.7**	**6**	**54**	**0.901**	**n_Components=3**
RR	0.737	0.898	5.1	41	6.1	56.4	0.901	α=0.5
LSVR	0.719	0.894	5.1	38.8	7.1	69.2	0.897	C=15, n_iterations=10000
RF	0.782	0.878	5.3	48.6	8.9	115.1	0.881	n_estimators=500, min_samples_split=14

*Best model based on five-fold CV error and Q^2_{F1}

Prediction of the designed set data

The prediction of the designed data set was done using the Prediction Reliability Indicator Tool (PLS Version) available from https://dtclab.webs.com/software-tools[30]. The training, test, and designed set data were taken as inputs and the result of the predictions have been provided in an excel sheet of **Supplementary Material SI-1**. The SSA values of the designed data set are mainly dependent on the *RA function* and *CT* values (positive and negative contributions, respectively). A close analysis of the designed set shows the data points with higher SSA values are associated with a higher calcination time (*AH*) and a lower drying temperature (*DT*).

Comparison with the previous work

Shi et al.[7] in 2018 worked on the prediction of the specific surface area of perovskites using various machine learning algorithms. The results obtained from their PLS model and the artificial neural network (ANN) approach have very low values of correlation coefficients (*R*) while the correlation coefficient obtained in their SVR approach was good. However, they did not report the quality of the external validation metrics. The PLS q-RASPR model reported in the current paper is simple, interpretable, transferable, and does not require exhaustive system resources yet follows a well-defined machine learning algorithm and includes non-linear *RA function*. We have reported both

the $R^2_{(Train)} = 0.74$ and $Q^2_{(LOO)} = 0.64$ which suggests that the model is highly robust and has sufficient internal validation quality. Moreover, our q-RASPR model delivers a very high predictivity as evident from the values of $Q^2_{F1} = 0.90$ and $Q^2_{F2} = 0.90$. The quality of external predictions of the PLS q-RASPR model is also evident from the scatter plot (Fig. 9). We have designed a set that consists of 450 data points, predicted their specific surface area, checked the prediction quality, and finally the status of AD. Shi et al. also reported that the descriptor *CT* exerts a negative contribution towards the model, but contrastingly, they have used higher values of *CT* for their external (designed) data points and claimed their higher values of predicted SSA. In our designed set, we have used both lower and higher values of *CT* for the data points. While most of the data points were inside AD, there were 38 data points outside the AD. It was observed that the *CT* values of these 38 data points were all very high, which supports the findings from our developed model. Thus, our PLS q-RASPR model appears much superior to the PLS, ANN, and SVR models reported by Shi et al.[7]

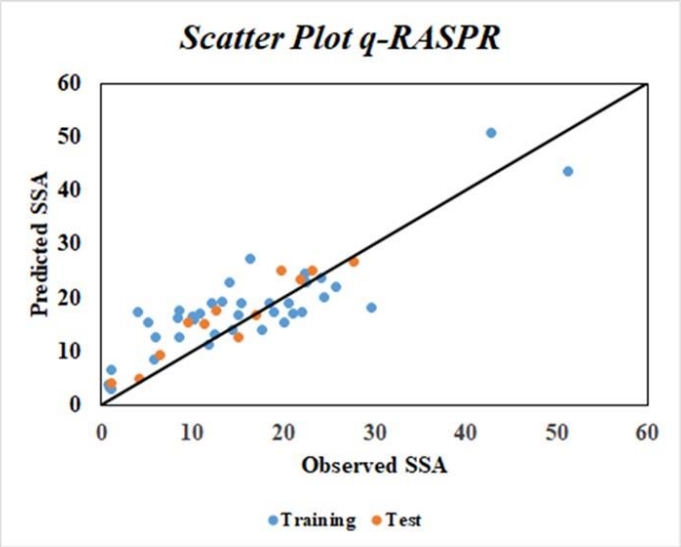

Figure 9. Scatter Plot representing the prediction quality of the PLS q-RASPR model.

Conclusion

This study aims at estimating the catalytic activity of perovskites by modeling their specific surface area. As a general notion, an increase in the surface area of catalysts provides a greater surface for the adsorption of the reactant molecules thus enhancing their reactivity and lowering of the activation energy. We have used the novel q-RASPR approach to predict the specific surface area of perovskite molecules. In the absence of a true external set, the dataset was divided into training and test sets, and the most discriminating descriptors were identified. Using this set of features, the training set was subdivided into sub-training and sub-test sets, and Read-Across-based predictions were generated by changing the number of close source compounds. The close source compound which yields the maximum values of Q_{F1}^2 and Q_{F2}^2 for the sub-test set was identified as the optimized number of close source compounds, and the original test set Read-Across-based predictions were performed using the same setting. This setting was also used to calculate the RASPR descriptors for both the training and test sets. After the computation of the RASPR descriptors, they were then clubbed with the originally selected atomic information descriptors and process variables, and a further feature selection algorithm was employed. As a result, four descriptors were identified to be important, and a Partial Least Squares (PLS) model was generated. This model explains that the composite feature *RA function* is essential as it incorporates information on all the most discriminating descriptors selected initially. It also explains that the calcination temperature (*CT*) is also another important process variable that is responsible for the change in the specific surface area. The descriptor *Abs diff* shows its importance in terms of the similarity aspects as there are more training compounds having a higher *MaxPos* value as compared to the *MaxNeg* value. Also, the radius of the atom at B position (*Rb*) is also an important feature as it is directly related to the mass of the molecule which in turn is related to the specific surface area. This generated PLS model was also used to predict the specific surface area of a designed external dataset of 450 compounds. To compare the predictive ability of the generated PLS model, we have also applied various other Machine Learning approaches like Ridge Regression, Linear Support Vector Regression, and Random Forest. It was found that these ML models do not supersede the predictive ability of the generated PLS model. Thus, we may infer that the PLS q-RASPR model is best suited for the prediction of specific surface area of perovskites, and also it provides simplicity, reproducibility, and transferability.

- **Data and Software availability**

The DTC Laboratory tools used in this study are available free of charge from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home and http://teqip.jdvu.ac.in/QSAR_Tools/ .

- **Supporting Information**

Supplementary Materials SI-1 contains raw data files in Excel format.

Supplementary Material SI-2 contains the model equation, applicability domain plots, Regression coefficient plot, Variable importance plot and Randomization plots of the PLS model.

- **Conflict of interest**

Declared none.

- **Author Information**

Corresponding author.

Kunal Roy, Drug Theoretics and Cheminformatics Laboratory**,** Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India. Email: kunal.roy@jadavpuruniversity.in . ORCID: https://orcid.org/0000-0003-4486-8074

First author.

Arkaprava Banerjee, Drug Theoretics and Cheminformatics Laboratory**,** Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India. ORCID: https://orcid.org/0000-0001-8468-0784

Second author.

Agnieszka Gajewicz-Skretna, Laboratory of Environmental Chemoinformatics, Faculty of Chemistry, University of Gdansk, Wita Stwosza 63, 80-308 Gdansk, Poland. ORCID: https://orcid.org/0000-0001-7702-210X

- **Author Contributions**

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

- **Funding**

This research is funded by Science and Engineering Research Board (SERB), New Delhi under the MATRICS scheme (MTR/2019/000008). AB thanks the Life Science Research Board, DRDO, New Delhi for a senior research fellowship

- **Notes**

A preprint version of this manuscript has been deposited to *ChemRxiv* https://chemrxiv.org/engage/chemrxiv/public-dashboard .

- **Abbreviations**

QSAR (Quantitative structure-activity relationship); QSPR (Quantitative structure-property relationship); RA (Read-across); RASAR (Read-Across Structure-Activity Relationship); PLS (Partial least squares); RF (Random Forest); LSVR (Linear Support Vector Regression); RR (Ridge Regression); ANN (Artificial Neural Networks); LDA (Linear Discriminant Analysis); RASPR (Read-Across Structure-Property Relationship); MAE (Mean Absolute Error); MSE (Mean Squared Error)

References

(1) Quan, L. N.; Rand, B. P.; Friend, R. H.; Mhaisalkar, S. G.; Lee, T. W.; Sargent, E.H. Perovskites for next-generation optical sources. *Chem. Rev.* 2019, 119, 7444-7477.

(2) Wolfram, T.; Ellialtioğlu, S. Electronic and optical properties of d-band perovskites. 1st ed. Cambridge University Press: New York; 2006.

(3) Galasso, F. S. Structure, Properties and preparation of perovskite-type compounds. In: Smoluchowski R, Kurti N, editors. 1st ed. Pergamon Press: New York; 1969. 3– 49. Chapter 2.

(4) Johnsson, M.; Lemmens, P. Crystallography and chemistry of perovskites. *ArXiv.* 2005.

(5) Assirey, E. A. R. Perovskite synthesis, properties and their related biochemical and industrial application. *Saudi Pharm. J.* 2019, 27, 817-829

(6) Kim, J. Y.; Lee, J. W.; Jung, H. S.; Shin, H.; Park, N. G. High-Efficiency perovskite solar cells. *Chem. Rev.* 2020, 120, 7867-7918.

(7) Shi, L.; Chang, D.; Ji. X.; Lu, W. Using data mining to search for perovskite materials with higher specific surface area. *J. Chem. Inf. Model.* 2018, 58, 2420-2427.

(8) Shellaiah, M.; Sun, K. W. Review on sensing applications of perovskite nanomaterials. *Chemosensors.* 2020, 8(55), 1-35.

(9) Moradi, Z.; Fallah, H.; Hajimahmoodzadeh, M. Nanocomposite perovskite based optical sensor with broadband absorption spectrum. *Sens. Actuators A.* 2018, 280, 47–51.

(10) 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.

(11) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Sos. Rev.* 1995, 24(4), 279-287

(12) Varnek, A.; Baskin, I. Machine learning methods for property prediction in chemoinformatics: Quo Vadis. *J. Chem. Inf. Model*. 2012, 52, 1413-1437.

(13) Chatterjee, M.; Banerjee, A.; De, P.; Gajewicz-Skretna, 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.

(14) Manganelli, S.; Benfenati, E. Use of Read-Across tools. In: Benfenati E. (Ed.), In Silico Methods for Predicting Drug Toxicity, Humana Press, New York; 2016, 305-322.

(15) 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.

(16) Banerjee, A.; Roy, K. First report of q-RASAR modeling toward an approach of easy interpretability and efficient transferability. *Mol. Divers.* 2022, 26(5), 2847-2862.

(17) Zhang, Y.; Xu, X. Modeling oxygen ionic conductivities of ABO_3 Perovskites through machine learning. *Chem. Phy*. 2022, 558, 111511.

(18) Wang, W.; Li, Y.; Zou, A.; Shi, H.; Huang, X.; Li, Y.; Wei, D.; Qiao, B.; Zhao, S.; Xu, Z.; Song, D. Predicting the photon energy of quasi-2D lead halide perovskites from the precursor composition through machine learning. *Nanoscale Adv*. 2022, 4, 1632-1638.

(19) Kim, E.; Kim, J.; Min, K. Prediction of dielectric constants of ABO_3-type perovskites using machine learning and first-principles calculations. *Phys. Chem. Chem. Phys*. 2022, 24, 7050-7059.

(20) Lyu, R.; Moore, C. E.; Liu, T.; Yu, Y.; Wu, Y. Predictive design model for low-dimensional organic−inorganic halide perovskites assisted by machine learning. *J. Am. Chem. Soc*. 2021, 143, 12766-12776.

(21) Yan, W.; Liu, Y.; Zang, Y.; Cheng, J.; Wang, Y.; Chu, L.; Tan, X.; Liu, L.; Zhou, P.; Li, W.; Zhong, Z. Machine learning enabled development of unexplored perovskite solar cells with high efficiency. *Nano Energy*. 2022, 99, 107394.

(22) 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.

(23) Roy, K.; Mitra, I. On various metrics used for validation of predictive qsar models with applications in virtual screening and focused library design. *Comb. Chem. High Throughput Screen*. 2011, 14, 450-474.

(24) Roy, K.; Das, R. N.; Ambure, P.; Aher, R. B. Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemom. Intell. Lab. Syst*. 2016, 152, 18-33.

(25) Hoerl, A.E.; Kennard, R.W. Ridge Regression: Biased estimation for nonorthogonal problems. *Technometrics*. 1970, 12(1), 55-67.

(26) Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* 1995, 20, 273–297.

(27) Breiman, L. Random Forests. Mach. Learn. 2001, 45, 5–32.

(28) 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.

(29) Seth, A.; Roy, K. QSAR modelling of algal low level toxicity values of different phenol and aniline derivatives using 2D descriptors. *Aquat. Toxicol*. 2020, 228, 1-11.

(30) Roy, K.; Ambure, P.; Kar, S. How precise are our quantitative structure-activity relationship derived predictions for new query chemicals?. *ACS Omega*. 2018, 3, 11392-11406.

(31) Gaber, A.; Abdel-Rahim, M. A.; Abdel-Latief, A. Y.; Abdel-Salam M. N. Influence of calcination temperature on the structure and porosity of nanocrystalline SnO_2 synthesized by a conventional precipitation method. *Int. J. Electrochem. Sci*. 2014, 9, 81-95.

(32) Wold, S.; Sjostrom, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* 2001, 58(2), 109-130.