First report of q-RASAR modeling towards an approach of easy interpretability and efficient transferability

Arkaprava Banerjee *and* *Kunal Roy**

Drug Theoretics and Cheminformatics Laboratory,

Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India.

URL: https://sites.google.com/site/kunalroyindia

*Corresponding author

E-mail: kunal.roy@jadavpuruniversity.in

Abstract

Quantitative structure-activity relationship (QSAR) and read-across techniques have recently been merged into a new emerging field of Read-across Structure-Activity Relationship (RASAR) that uses the chemical similarity concepts of read-across (an unsupervised step) and finally develops a supervised learning model (like QSAR). The RASAR method has so far been used only in case of graded predictions or classification modeling. In this work, we attempt, for the first time, to apply RASAR for quantitative predictions (q-RASAR) using a case study of androgen receptor binding affinity data. We have computed a number of error-based and similarity-based measures such as weighted standard deviation of the predicted values, coefficient of variation of the computed predictions, average similarity level of close training compounds for each query molecule, standard deviation and coefficient of variation of similarity levels, maximum similarity levels to positive and negative close training compounds, a concordance measure indicating similarity to positive, negative or both classes of close training compounds, etc. We have clubbed these additional measures along with the selected chemical descriptors from the previously developed QSAR model and redeveloped new partial least squares (PLS) models from the training set, and predicted the endpoint using the query data set. Interestingly, these new models outperform the internal and external validation quality of the original QSAR model. In this study, we have also introduced a new similarity-based concordance measure that can significantly contribute to the model quality. A q-RASAR model also has the advantage over read-across predictions in providing easy interpretation and indicating quantitative contributions of important chemical features. The strategy described here should be applicable to other biological/toxicological/property data modeling for enhanced quality of predictions, easy interpretability, and efficient transferability.

Keywords: q-RASAR, Read-across, QSAR, Similarity, Prediction

1. Introduction

Most of the activity/toxicity prediction methods and screening tools rely on the molecular similarity principles which suggest that compounds with similar structural features will exhibit similar activity/property/toxicity.[1, 2] The conventionally used methods like quantitative structure-activity/property/toxicity relationships (QSAR/QSPR/QSTR), pharmacophore mapping, ligand-based virtual screening etc. are based on this basic assumption. While we aim to develop a set of training or source compounds of maximum structural diversity, we compare the similarity of the test or query compounds to some of the source compounds. The structure-based methods also are indeed based on similar principles when the similarity consideration is extended to ligand – receptor interactions in general. Molecular similarity provides a popular method for virtual screening with application of clustering methods on chemical databases. On the other hand, molecular diversity analysis explores the way of molecules to cover a determined structural space. Molecular similarity, in general, has three components: descriptors (or fingerprints), coefficients or indices, and a suitable weighting scheme.[3] Quantitative molecular similarity is expressed in terms of different indices or metrics which may eventually lead to clustering or grouping of compounds. One of the similarity – based methods, very popular in computational toxicology, is read-across which can be applied for both quantitative and qualitative predictions[4, 5] even when only limited amount of experimental data is available. Recently, a tie between read-across and QSAR methods has been implemented leading to a new field Read-across structure-activity relationship (RASAR) which appears to be of much promise in predictive toxicology.[6]

Animal – based toxicity studies are resource intensive and involve ethical considerations. Most of the experimental toxicity studies are related to evaluation of lethal potential of chemicals and do not address sufficiently the mechanistic aspects. To address this issue, the scientific community has now inclined to adopt a mechanistic approach of Adverse Outcome Pathway (AOP) consisting of key events, molecular initiating events and adverse outcome.[7] Starting from a conceptual framework, AOPs have rapidly evolved into a formalized framework for organizing biological and toxicological knowledge as per a set of principles and guidelines that are generally accepted by the scientific and regulatory communities like Organization for Economic Co-operation and Development (OECD).[6, 7] To avoid the problems associated with animal experimentation, scientists now rely on computational toxicity prediction methods to bridge the data gaps,[8] as it is practically impossible to gather information of toxicity of thousands of chemicals against hundreds of different endpoints. In this sense, computational prediction methods that also involve mechanistic explanation like consideration of AOPs appear to be very much promising.[9] Recently, machine learning based analysis of big data has led to the generation of RASAR models which have been claimed to outperform reproducibility of animal experimentations.[5] RASAR can be applied simultaneously to a large number of endpoints and can be related to multiple toxicological targets addressing MIEs in the adverse outcome pathway of several toxicity endpoints.[10]

Considering these aspects, RASAR appears to be a promising tool for developing expert systems of predictions of toxicity and ecotoxicity of drugs and organic chemicals in general.

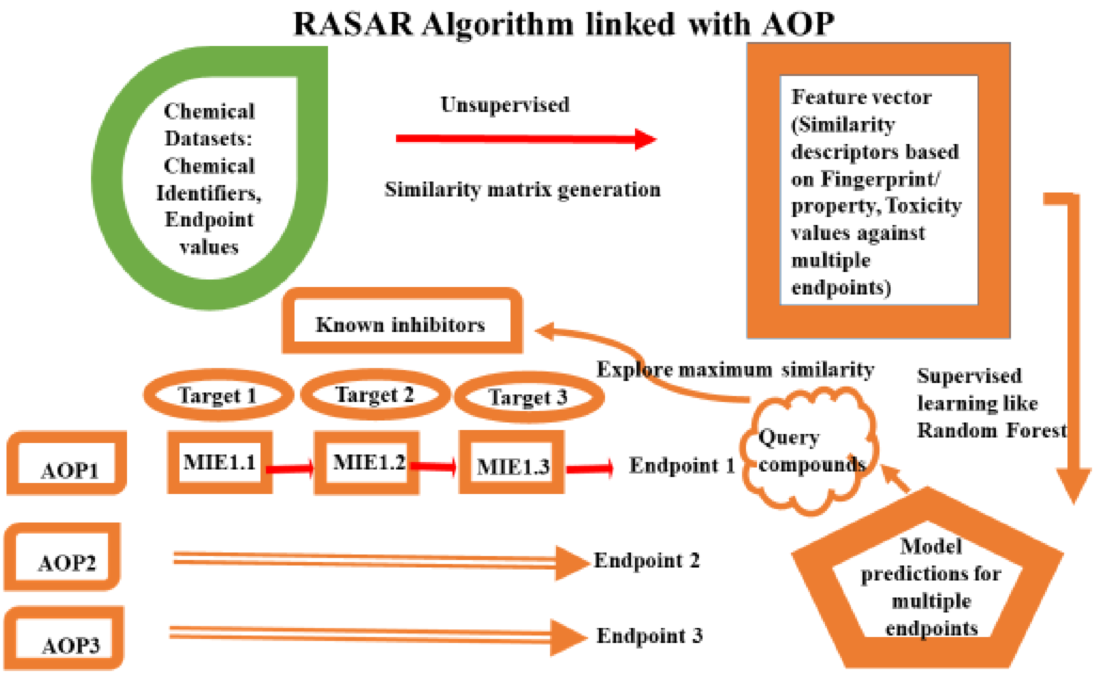
RASAR Algorithm linked with AOP
Chemical Datasets: Chemical Identifiers, Endpoint values
Unsupervised
Similarity matrix generation
Feature vector (Similarity descriptors based on Fingerprint/ property, Toxicity values against multiple endpoints)
Known inhibitors
Explore maximum similarity
Supervised learning like Random Forest
Target 1
Target 2
Target 3
AOP1
MIE1.1
MIE1.2
MIE1.3
Endpoint 1
Query compounds
AOP2
Endpoint 2
AOP3
Endpoint 3
Model predictions for multiple endpoints

Figure 1. Generalised RASAR algorithm linked with AOP and MIEs using chemical similarity-based approaches

The RASAR method has so far been used only in case of graded predictions or classification modeling. In this work, we attempt, for the first time, to apply RASAR for quantitative predictions (q-RASAR) using a case study of androgen receptor binding affinity data. Although data fusion RASAR allows the application of multiple endpoints simultaneously in an approach to relate them to different biological targets involving appropriate molecular initiating events in different adverse outcome pathways[9] (**Figure 1**), we use here only the simple RASAR approach to model a particular endpoint of androgen receptor binding of endocrine disruptor chemicals. However, the strategy mentioned here should be extendable to more complex problems involving multiple endpoints.

2. Materials and Methods

For this report, we have used a data set androgen receptor binding affinity (RBA) recently used by us for QSAR model development and chemical read-across predictions.[11] The RBA data were originally collected from the Endocrine Disruptor Knowledge Base (EDKB) database (https://www.fda.gov/science-research/bioinformatics-tools/endocrine-disruptor-knowledge-base) and chemical curation of the compounds was performed by the application of a KNIME workflow (https://sites.google.com/site/dtclabdc/) taking the single .sdf file as input. Further details of the SMILES notation of the compounds and observed RBA values along with other raw data are available in **Supplementary Material SI-1.** The current work uses chemical read-across predictions using the tool Read-Across ver. 4.0 (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) as originally reported in a previous publication.[12] The workflow of chemical read-across predictions is presented in **Figure 2**. We have finally used the descriptors selected in the previous QSAR model as the important physicochemical measures of the compounds in addition to different similarity measures as described below for the q-RASAR analysis.

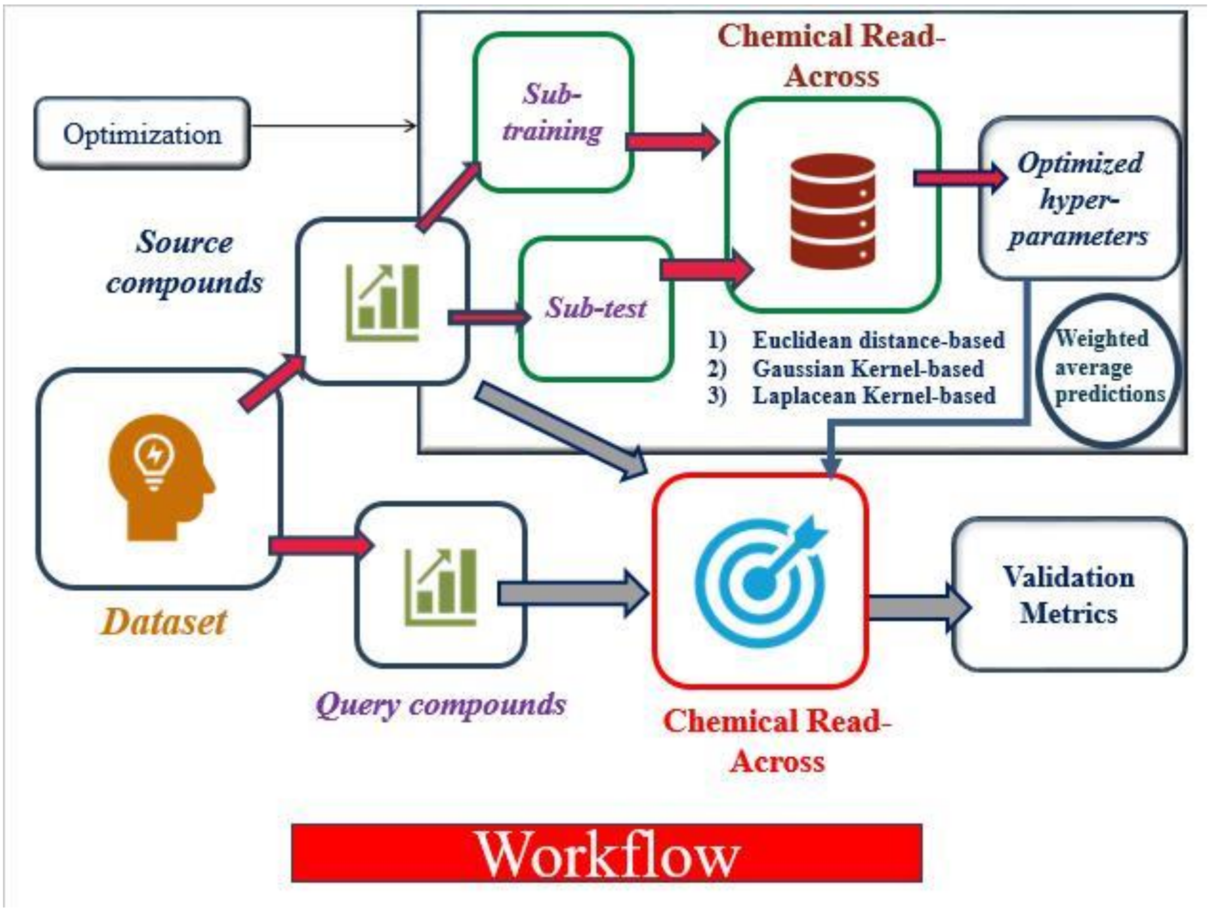

Figure 2. Workflow of the Chemical Read-Across predictions

2.1. Computation of similarity measures. We have used here the same division of training (source) and test (query) sets as used in our previous analysis.[11] However, one compound (no. **187**) in the training set was an outlier and structurally significantly dissimilar from rest of the training compounds. Thus, while computing similarity metrics, its close congeners could not be found and hence this particular compound was not considered for q-RASAR model development. This may be considered as an advantage of the q-RASAR methodology to identify outlier compounds as influential observations enabling further refinement of the models to derive statistically more meaningful relationships. We have derived read-across predictions for the query compounds (test set) as reported in the previous work. Our read-across tool generates in addition to read-across predictions various similarity and error measures such as standard deviation and coefficient of variation of the activity of similar training compounds for each query compound, average and standard deviation of similarity levels and their coefficient of variation of similar training compounds (up to 10 in number) to each query compound, maximum similarity level to positive and negative compounds (based on the "training set" response mean), a concordance measure indicating similarity to positive, negative or both classes of close training compounds,[13] etc. as detailed in **Table 1**. We have used these measures along with the selected structural and/or physicochemical features (**Table 2**) as input for the q-RASAR analysis. Please note that the read-across tool reports the similarity measures for the query set compounds. In order to prepare the similarity descriptor matrix for the source compounds, the training test itself was used as a test set, and the derived descriptors were used for subsequent q-RASAR model development. Here, the development of the similarity measures represents the unsupervised step (without using the response values) followed by application of the supervised learning step (statistical modeling) like QSAR. Although Read-Across v 4.0 computes similarity based on Euclidean distance, Gaussian kernel and Laplacian kernel - based functions, we have used here different measures computed from Gaussian kernel (GK) - based function only, as this approach gave the best read-across predictions in the previous analysis on this data set.[11]

2.2. q-RASAR model development. In this work, we have used simple linear model building for easy interpretation of the selected features and easy transferability. The pooled set of descriptors was subjected to best subset selection using the tool available from http://teqip.jdvu.ac.in/QSAR_Tools/ and a number of MLR models were selected based on their performance in the internal validation and external validation and maintaining diversity with respect to the selected descriptors in the final models. The descriptors selected in these models were then separately subjected to Partial Least Squares (PLS) regression. PLS being a generalized and more robust version of multiple linear regression, we have relied on this approach for the final model development.[14] PLS is able to handle a higher number of descriptors in the final model without compromising with the degree of freedom by using latent variables (LVs), and it can handle noisy data in a better way than MLR. The number of LVs was selected based on the leave-one-out (LOO) cross-validation.[15] PLS results can be presented in an MLR like equation with the advantage that quantitative contribution of each appearing descriptors can be explicitly depicted which helps the user to make a clear diagnosis of underlying mechanism for the response being modelled. We have also avoided here using any machine learning technique to maintain simplicity and transferability of the developed models for its wide usability.

The flowchart of the present work is shown in **Figure 3**.

The quality of the q-RASAR models has been evaluated based on the traditional quality and validation metrics of QSAR models like determination coefficient (R^2), mean absolute error (MAE), leave-one-out cross-validated determination coefficient (Q^2_{LOO}), external predicted variance with different variants (R^2_{pred} or $Q^2_{ext\_F1}$, $Q^2_{ext\_F2}$, $Q^2_{ext\_F3}$), etc.[16] We have also applied MAE-based criteria for external predictions as described previously.[17] The PLS models have been explained and interpreted using various plots like score plot (allocation of the compounds in the LV space showing their distribution and similarity/diversity among the compounds), loading plot (the loadings of individual descriptors into the first two latent variables are plotted and the distance of the X-variables from the origin correlates with the importance of the descriptor), randomization plot (a plot of R^2 and Q^2 values of the random models (Y-axis) vs. the correlation coefficient between the original Y-values and the permuted Y values (X-axis)), scatter plot (a plot of predicted response values (Y-axis) vs. observed response values (X-axis)), applicability domain (AD) plot using distance to model in X space (DModX) approach, etc.[14]

After development of individual q-RASAR models, we have then attempted to pool the descriptors appearing in the best models to develop new models in an attempt to further enhance the quality. Attempt has also been made to derive intelligent consensus predictions from the individual models as these have previously proved to enhance the quality of external validation.[18]

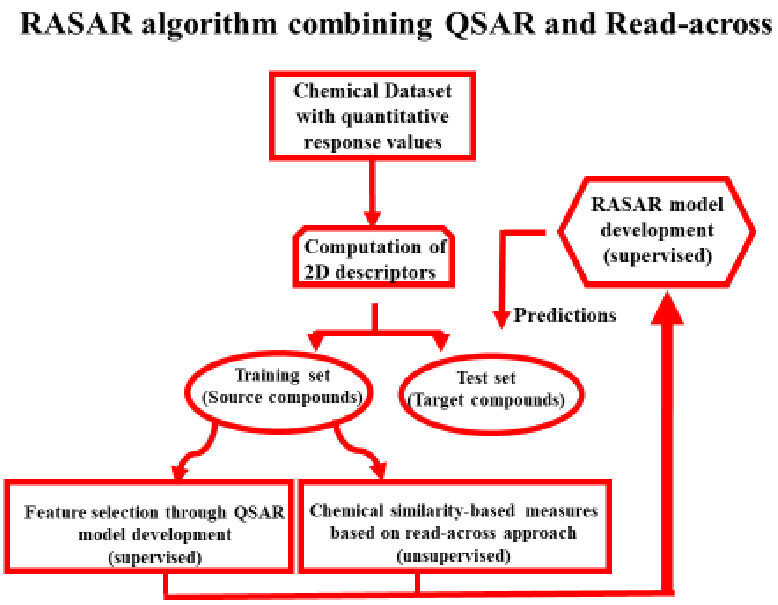
RASAR algorithm combining QSAR and Read-across
Chemical Dataset with quantitative response values
Computation of 2D descriptors
RASAR model development (supervised)
Predictions
Training set (Source compounds)
Test set (Target compounds)
Feature selection through QSAR model development (supervised)
Chemical similarity-based measures based on read-across approach (unsupervised)

Figure 3. The generalised q-RASAR algorithm

Table 1. List of similarity and various error measures generated for each query compound during read-across predictions

Measure	Definition
Dispersion measures	
SD_activity	Standard deviation of the (observed) activity values of the selected close source compounds for each query compound

CV_activity	Coefficient of variation of the response
Similarity measures	
Average similarity	Mean similarity to the close source compounds for each query compound
SD_similarity	Standard deviation of the similarity values of the selected close source compounds for each query compound
MaxPos	Maximum Similarity level to the Positive close source compounds (based on source set observed mean)
MaxNeg	Maximum Similarity level to the Negative close source set compounds (based on source set observed mean)
AbsDiff	Absolute difference between MaxPos and MaxNeg
Concordance measure	
g	$g = 1 - 2 \times \|PosFrac - 0.5\|$, where $PosFrac$ is the fraction of the close source compounds belonging to the Positive Class based on the source set response mean as the threshold.[13]

Table 2. List of physicochemical features selected from the previously reported QSAR model[11]

Measure	Description	Comment
SsssCH	Sum of E-state value of tertiary carbon atoms of type >CH-	E-state index
MaxaaCH	Maximum E-state value of the carbon atom of type aaCH	E-state index

nCconj	Number of non-aromatic conjugated carbons (sp^2)	Constitutional descriptor
LOGP99	Wildmann-Crippen octanol-water partition coefficient	Hydrophobicity measure
F10[C-O]	Frequency of C and O at the topological distance 10	Atom pair index
minsOH	Minimum Estate of the -OH hydroxyl group	E-state index
N%	The percentage of nitrogen present in the molecular structure	Constitutional descriptor
F08[O-F]	The frequency of O and F atoms at the topological distance of 8	Atom pair index

2.3. Software used in the q-RASAR modeling study

The computation of different error and similarity-based measures was done using the tool Read-Across v4.0 available freely from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home . The best subset selection, PLS regression, and intelligent consensus predictions were done using MLR Best Subset Selection, Partial Least Squares and Intelligent Consensus Predictor tools available freely from http://teqip.jdvu.ac.in/QSAR_Tools/ . The PLS plots were done using SIMCA-P v10.0 software (https://www.sartorius.com/) and the bubble plots were done using SigmaPlot v11 (http://www.sigmaplot.co.uk/).

3. Results and Discussion

Four individual PLS models were finally derived from the chemical feature - based descriptors along with similarity measures, and these are presented in **Table 3**. The statistical quality and validation measures of the models are shown in **Table 4**. The regression coefficient bubble plots (bubble size is proportional to individual VIPs[19]) are shown in **Figure 4**. It is evident from **Figure 4** that similarity – based measures show higher importance in the models in most of the cases while chemical descriptors such as N%, F10[C-O] and minsOH are less important descriptors as per the VIP values. This is also evident from the loading plots (**Figure 5**) where in most of the cases MaxPos resides near logRBA (and away from the origin) while MaxNeg appears in the opposite side of the X-axis. The chemical descriptors F10[C-O], N% and minsOH are located near the origin suggesting their less significant contributions to the models. The score plots show that most of the compounds are within the applicability domain (**Figure S1 in Supplementary Materials SI-2**) as also evident from the AD plot based on the distance to model in X space (DModX) approach (**Figure S2 in Supplementary Materials SI-2**). Model randomization plots with low $R^2_{\text{intercept}}$ and $Q^2_{\text{inetrcept}}$ also reveal that the models are not derived by chance **(Figure S3 in Supplementary Materials SI-2**). The scatter plots show that there is good concordance between the observed and predicted values (**Figure 6**).

3.1. Comparison of the quality of q-RASAR models. We have developed four individual PLS q-RASAR models (**M1**, **M2**, **M3** and **M4**) which are robust and predictive superseding the quality of our previous predictions using QSAR and Read-Across methodologies[11] in terms of the quality of external validation metrics **(Table 4)**. However, in terms of the internal validation, the R^2 and $Q^2_{(LOO)}$ values are comparable to the previous QSAR model. It is also important to note that the $MAE_{(TEST)}$ values of these q-RASAR models are lower than our previous 2D-QSAR model; however, the $MAE_{(TEST)}$ value of the previous Read-Across predictions was lower than the present individual q-RASAR models. We may note here that the Read-Across methodology does not involve the development of any model and thus interpretation of quantitative contributions of various contributing factors is not possible in Read-Across, but this can be done in case of q-RASAR models like any QSAR model. In search of a more improved model, we have pooled the descriptors and developed three different pooled descriptor PLS models (**P1**, **P2**, **P3**), and we found that the pooled descriptor PLS model **P2** had better internal validation metric values as compared to the previous QSAR model, and there was an overall improvement observed in the internal validation of the pooled PLS models as compared to the individual PLS models **(Table 4)**. Also, the predictivity of model **P1** was even better than all of the individual PLS q-RASAR models as well as the previous 2D-QSAR and Read-Across approaches. The $MAE_{(TEST)}$ values of the pooled descriptor PLS models were also significantly lower than the individual PLS models and the previous QSAR model.[11] We have then applied an intelligent consensus prediction method[18] in order to check the predictivity of the individual PLS models. This method selects the best model for a particular query compound using different consensus-based prediction methods with enhanced predictivity and reduced $MAE_{(TEST)}$ values. The consensus-based predictions have the advantage of utilization of a greater number of features as compared to a single PLS model thus handling the complexity of a data set in a much more efficient way, which ultimately increases the applicability domain in terms of chemical space and leads to enhanced predictivity with reduced prediction errors. In addition, the shortcoming of an individual model may be overcome by the use of another model. Only those individual models qualify for a particular query compound in case at least three compounds from the training set have their Euclidean Distances with respect to the query compound within the threshold. This threshold value is derived from the training set compounds, and it corresponds to the mean Euclidean Distance + k × SD, where k = 3. Consensus Model 1 (CM1) is developed using the mean of predictions from all qualified individual models. Consensus Model 2 (CM2) is derived from the weighted average predictions (WAPs) from all qualified individual models. Consensus Model 3 (CM3) is involved in the compound-wise best selection of predictions from individual models.[18] Among the ICP models thus generated, the consensus model 3 (**ICP3**) shows a significant reduction in the $MAE_{(TEST)}$

which is better than the individual PLS models, pooled PLS models, previous QSAR approach and even the Read-Across predictions.

Table 3. List of q-RASAR models

Model No.	**Equation**
Individual q-RASAR models	
M1	$logRBA = -1.33 + 2.27MaxPos(GK) - 3.57Avg.Sim(GK) - 1.02g(GK) + 0.04minsOH - 0.14N\% - 0.06F10[C - O]$
M2	$logRBA = -2.38 - 1.66MaxNeg(GK) + 0.78MaxPos(GK) + 4.32SD\ Similarity(GK) + 0.06minsOH - 0.09N\% - 0.05F10[C - O]$
M3	$logRBA = -1.97 + 0.35SsssCH + 1.55MaxPos(GK) - 0.34MaxaaCH - 1.31Avg.Sim(GK) + 0.01minsOH - 0.04F10[C - O]$
M4	$logRBA = -2.93 - 1.25MaxNeg(GK) + 1.22MaxPos(GK) + 0.73SD\ Activity(GK) + 0.05nCconj + 2.47SD\ Similarity(GK) + 0.03minsOH$
Pooled descriptor q-RASAR models	
P1 (M1 + M2)	$logRBA = -1.71 - 1.47MaxNeg(GK) + 1.06MaxPos(GK) + 2.88SD\ Similarity(GK) - 0.86Avg.Sim(GK) + 0.05minsOH - 0.41g(GK) - 0.10N\% - 0.05F10[C - O]$

P2 (M1+M2+M3)	$logRBA = -1.76 - 1.00MaxNeg(GK) + 0.29SsssCH + 0.91MaxPos(GK) - 0.24MaxaaCH - 0.40Avg.Sim + 1.32SD\ Similarity(GK) + 0.03minsOH - 0.04F10[C-O] - 0.05N\% + 0.17g(GK)$
P3 (M1+M2+M4)	$logRBA = -2.55 - 1.13MaxNeg(GK) + 1.10MaxPos(GK) + 0.72SD\ Activity(GK) + 0.08nCconj - 0.48Avg.Sim(GK) + 1.81SD\ Similarity(GK) + 0.03minsOH - 0.05F10[C-O] - 0.06N\% + 0.13g(GK)$

Table 4. Statistical quality and validation metric values of various q-RASAR models ($n_{Training}$ =102, n_{Test}=44) and comparison with previous QSAR and read-across predictions ($n_{Training}$ =103, n_{Test}=44)*

PLS Model(s)	**LVs**	R^2	$Q^2_{(LOO)}$	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	$MAE - Fitted_{(Train)}$	$MAE - LOO_{(Train)}$	$MAE_{(Test)}$
Individual Models									
M1	4	0.672	0.620	0.665	0.665	0.678	0.513	0.550	0.518
M2	3	0.713	0.660	0.655	0.655	0.669	0.478	0.516	0.494
M3	3	0.703	0.655	0.621	0.621	0.636	0.459	0.494	0.524
M4	2	0.721	0.653	0.613	0.612	0.628	0.451	0.486	0.504
Pooled descriptor models									
P1 (M1+M2)	3	0.718	0.666	**0.671**	**0.670**	**0.683**	0.479	0.517	0.478
P2 (M1+M2+M3)	2	**0.754**	**0.718**	0.630	0.629	0.644	**0.441**	**0.470**	0.504
P3 (M1+M2+M4)	2	0.720	0.646	0.638	0.638	0.652	0.441	0.478	0.480
Intelligent Consensus Models									
ICP1 (M1+M2+M3) (CM3)	-	-	-	0.657	0.657	0.670	-	-	0.484

ICP2 (M1+M2+M4) (CM3)	-	-	-	0.621	0.621	0.636	-	-	0.496
ICP3 (M1+M2+M3+M4) (CM3)	-	-	-	0.652	0.652	0.665	-	-	**0.463**
Previous 2D-QSAR model and Read-Across predictions (*Banerjee et al.* 2022)[11]									
2D-QSAR	3	0.737	0.680	0.582	0.582	0.606	0.456	0.497	0.539
Quantitative Read-Across (*Gaussian Kernel Similarity-based*)	-	-	-	0.635	0.635	0.656	-	-	0.468

***The best values of individual metrics are shown in bold.**

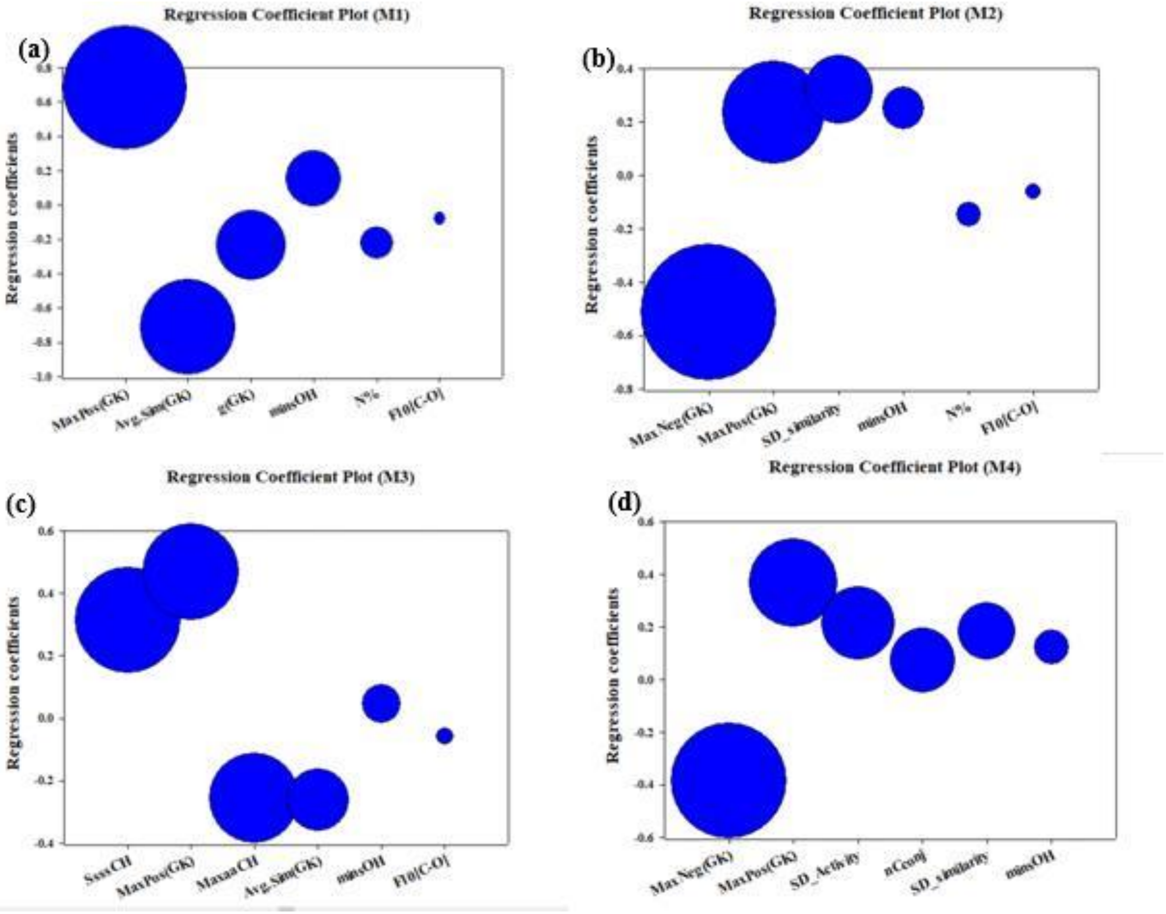

Figure 4. Bubble plot for regression coefficients of Models (a) M1, (b) M2, (c) M3 and (d) M4 [bubble size is proportional to individual VIP]

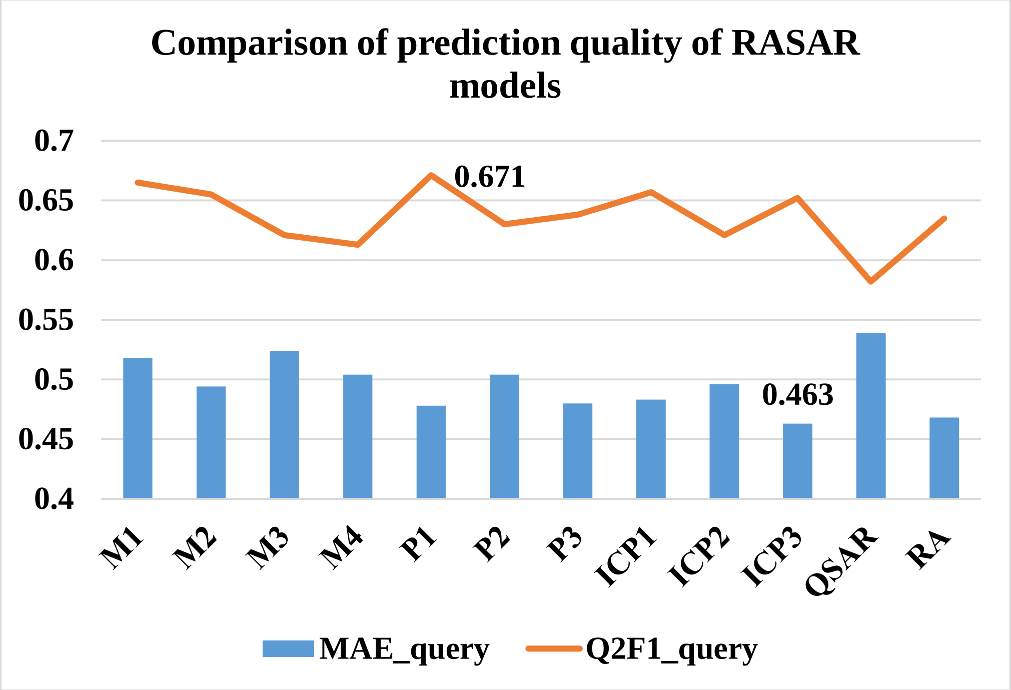

Figure 5. Comparison of prediction quality of q-RASAR models

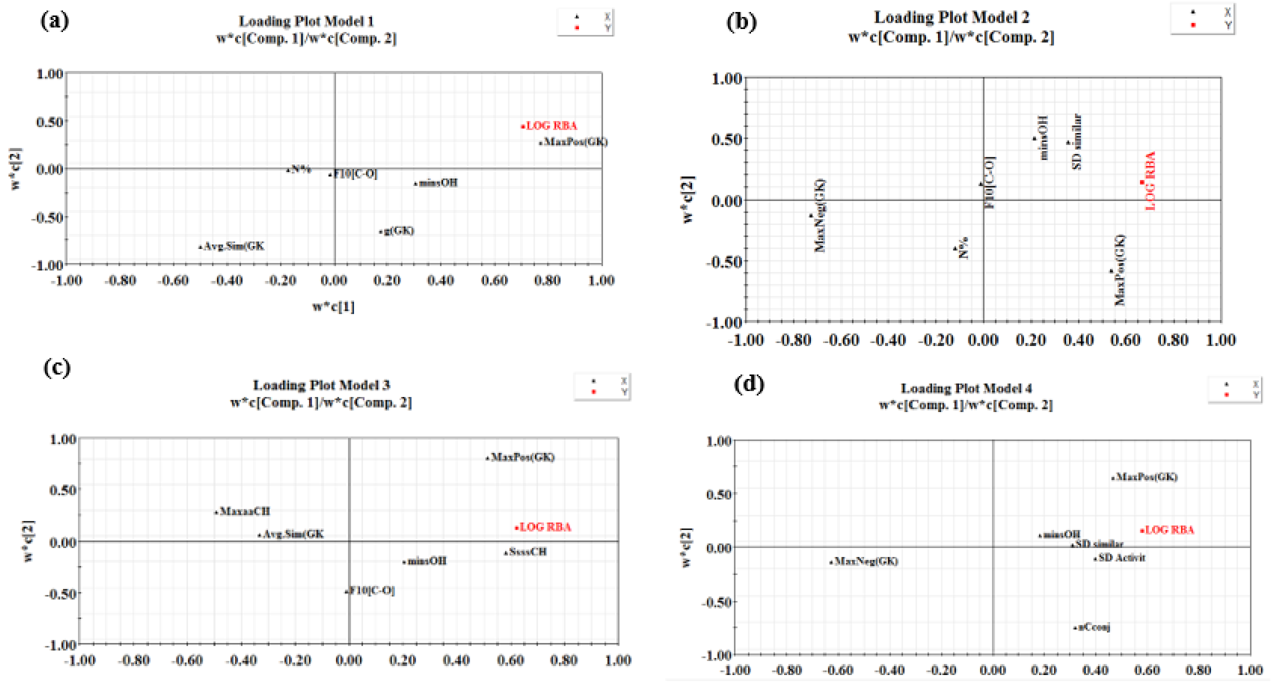

Figure 6. Loading plots of models (a) M1, (b) M2, (c) M3, (d) M4

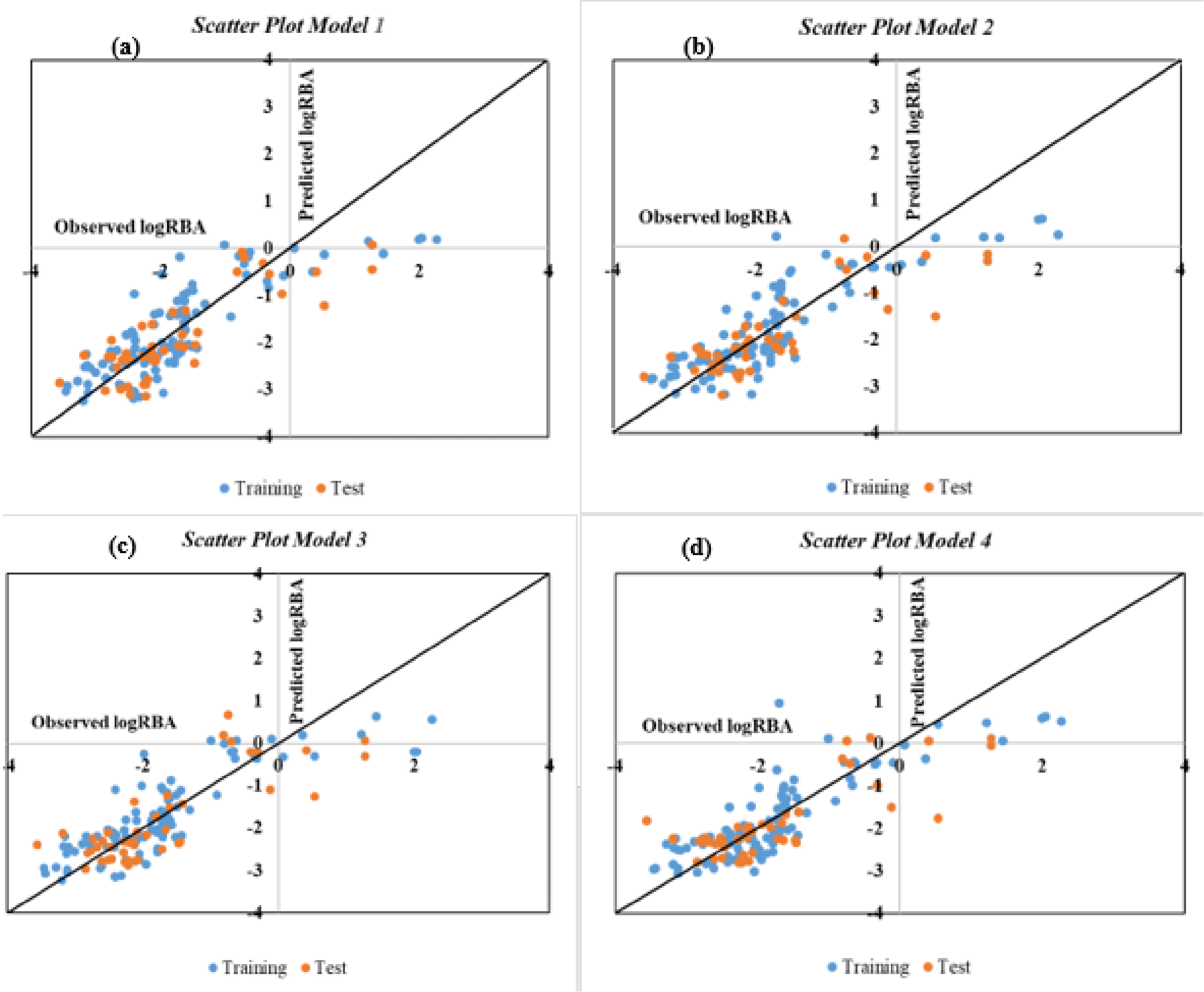

Figure 7. Scatter plots of models (a) M1, (b) M2, (c) M3, (d) M4

3.2. Interpretation of q-RASAR models. The descriptor **minsOH** signifies minimum atom-type E-State for the hydroxyl group, and it contributes positively to the developed models. The descriptor **F10[C-O]** stands for the frequency of Carbon and Oxygen atoms at the topological distance of 10, and this descriptor contributes negatively to the androgen receptor binding affinity of endocrine disruptors. **nCconj** is a functional group count descriptor which stands for the number of non-aromatic conjugated carbon atoms which are sp2 hybridized, and it also contributes positively to the receptor binding affinity in the developed models. The descriptor **SsssCH** stands for sum of sssCH E-state indices, and it encodes the presence of steroidal structures in the dataset. Due to its hydrophobic nature, SsssCH contributes positively to the response values. **MaxaaCH** descriptor denotes the maximum aaCH E-state indices, and it signifies the presence of CH groups in an aromatic ring. Since aromaticity in a molecule decreases its hydrophobicity due to the formation of induced dipoles, this descriptor contributes negatively to the receptor binding affinity. **N%** denotes the percentage of Nitrogen in a particular compound, and this is shown to have a negative effect on the receptor binding affinity. In the previous study[11] using the same dataset, N% contributed positively to the androgen receptor binding affinity. This can be explained from the Variable Importance Plot of Model 1 that the descriptor **MaxPos(GK)** (similarity value of the closest positive source compound to the target compound) has the highest importance while N% has a smaller importance, and compounds like Hydroxylinuron (**192**), apart from possessing the highest MaxPos(GK) value, also possesses Nitrogen in their structures, whose small but significant importance is already included in the MaxPos(GK) descriptor itself, and thus it results in positive contribution towards the receptor binding affinity. Hence, in order to balance the importance, N% shows a negative contribution in the developed RASAR models. The descriptor **MaxNeg(GK)** demonstrates the similarity value of the closest negative source compound to the target compound. A higher value of MaxNeg(GK) indicates that the target compound is very similar to such a source compound whose activity value is less than the threshold value. This indicates that there is a high probability that the target compound has low receptor binding affinity. Thus, the descriptor MaxNeg(GK) contributes negatively to the receptor binding affinity. The compound Bis(n-octyl) phthalate (**114**) has the highest MaxNeg(GK) value while its MaxPos(GK) value is very low, and thus this compound has a very low receptor binding affinity. On the other hand, the compound Trenbolone (**157**) has the highest MaxPos(GK) value but its MaxNeg(GK) value is much lower, and so this compound possesses a higher receptor binding affinity. In compounds like 5,6-Didehydroisoandrosterone (**139**), the MaxNeg(GK) value is only slightly greater than MaxPos(GK) value and thus the observed response value is only slightly lower than the threshold or it can be termed as a marginally inactive compound. The descriptor

Avg.Sim(GK) indicates the average similarity value of the close “n” source compounds with respect to a particular target compound where n is a positive integer value which is given as a user input in Read-Across-v4.0, and this descriptor contributes negatively to the developed models. This can be explained by careful observation of the data structure where there is a greater number of negative compounds as compared to the positive ones (with respect to the biological activity threshold) among the close “n” source compounds. A higher overall average similarity value of the close “n” source compounds indicates that there is a greater chance that the target compound is also negative. The compound 3-Chlorophenol (**52**) has a high Avg.Sim(GK) value, and thus it possesses a very low receptor binding affinity. If we study the first ten close source compounds for 3-Chlorophenol, we find that nine of them have their activity values below the threshold with their similarity levels ranging from 1 to 0.69. Thus, we may conclude that 3-Chlorophenol, which has very high similarity values with negative compounds, is less active, which is also evident from its experimental receptor binding affinity data. The descriptor **SD_Similarity(GK)** denotes the standard deviation of the similarity values of the close “n” source compounds, and this descriptor contributes positively to the receptor binding affinity values. The larger deviation of the similarity values indicates that some of the similarity values are very low which justifies that the target compound has properties to be positive. The compound R1881 (**207**) has a high SD_Similarity value, and this is evident from its MaxPos(GK) value which is the highest in its class and the value for MaxNeg(GK) is very low, and this compound possesses a high receptor binding affinity. In contrast, compounds like $4^{/}$-Chloroacetanilide (**87**) has a high SD_Similarity value and a high MaxNeg(GK) value but low MaxPos(GK) value exhibiting a much lower receptor binding affinity. The descriptor **SD_Activity(GK)** stands for the standard deviation of the biological activity values of the close “n” source compounds, and this contributes positively to the response value. The compounds having lower SD_Activity(GK) values tend to have their activities close to the average response value (threshold) as in p-p$^{/}$-DDE (**203**), and such compounds cannot be conclusively classified as active or inactive while compounds having a higher SD_Activity(GK) value along with higher MaxPos(GK) and lower MaxNeg(GK) values as in 11-keto testosterone (**108**) are likely to have a higher receptor binding affinity. In case of compounds like Bis(n-octyl) phthalate (**114**) where the SD_Activity(GK) is high but MaxPos(GK) is low and MaxNeg(GK) is high, we may suggest that the descriptor SD_Activity(GK) balances the contribution of MaxPos(GK) and MaxNeg(GK). The descriptor ***g*(GK)** is a concordance measure whose value ranges from 0 to 1. Compounds having low values of *g*(GK) implies that the close source compounds and the corresponding target compound are either active or inactive. In the case of Milbolerone (**193**), where the values of *g*(GK) is low, MaxNeg(GK) value is either low or insignificant but the value of MaxPos(GK) is high, and the compound shows a higher receptor binding affinity which explains the negative contribution of *g*(GK) in two of the developed models (**M1** and **P1**). Similarly, in the case of Flavanone (**182**) where the values of *g*(GK) and MaxPos(GK) are low but the value of MaxNeg(GK) value is maximum, the compound exhibits a much lower receptor binding affinity. This probably explains how *g*(GK) has positive regression coefficients in the other two developed models (pooled descriptor models **P2** and **P3**).

3.3. A new concordance measure g_m. The one drawback that arises in the use of *g* is that when the value of *g* is zero, it is unable to differentiate whether the close source compounds are positive or negative. This is because when all the close source compounds are positive, i.e. when the positive fraction corresponds to 1, the value of *g* becomes 0. Similarly, when all the close source compounds are negative, the positive fraction corresponds to 0 and the value of *g* also becomes 0. In addition, *g* attains the same value at a particular level of either PosFrac or NegFrac (**Figure 8**). Thus, to obviate this ambiguous nature of *g*, we have introduced a novel modified expression of *g* (let us call it g_m or *Banerjee-Roy coefficient*), which is able to differentiate the maximum similarity to either positive or negative compounds, and the corresponding equation for calculating g_m is:

$$g_m = (-1)^n 2|PosFrac - 0.5|$$

Here, n is a positive integer value which is either 1, when MaxPos < MaxNeg, or 2, when MaxPos >= MaxNeg. Note that here the directionality is applied with respect to the MaxPos or MaxNeg values and not with respect to the PosFrac value. This is because read-across predictions are made using the Read-Across ver 4.0 tool[12] as the weighted average predictions, and the compounds with higher level of similarity have higher weightage. Thus, a query compound with the highest level of similarity to a positive compound may have lower levels of similarity to several negative compounds. However, higher contributions to the read-across predictions are made by compounds with higher level of similarity in the weighted average prediction formula.[12]

The measure g_m(GK) is also a concordance measure with values ranging from -1 to +1, and the novelty here is that the g_m(GK) value of a particular target compound is positive only when the most similar source compound is positive with respect to the activity threshold and the value becomes negative when the most similar source compound is negative. Thus, compound **176** with the PosFrac value of 0.7 has a g_m value of -0.4 (as MaxNeg > MaxPos) while compound **126** with the NegFrac value of 0.7 has a g_m value of +0.4 (as MaxPos > MaxNeg) while both of them have the g value of 0.6 **(Figure 8)**. Model **P1a** is developed by modification of the g value based on the directionality applied with respect to MaxPos and MaxNeg, and we have denoted this modified g(GK) value expressed as $\boldsymbol{g_m}$**(GK)** in the equation.

$$logRBA = -1.21 - 1.31MaxNeg(GK) + 0.58g_m(GK) + 0.21MaxPos(GK) + 2.23SD\ Similarity(GK) - 0.67Avg.Sim(GK) + 0.06minsOH - 0.10N\% - 0.13F10[C-O] \qquad \textbf{(P1a)}$$

$$n_{Training} = 102\ \ n_{Test} = 44\ \ LV = 4$$

$$R^2 = 0.753 \quad Q^2_{(LOO)} = 0.698 \;\; Q^2_{F1} = 0.674 \;\; Q^2_{F2} = 0.674 \;\; MAE_{(TEST)} = 0.461$$

We find that the values of the internal validation metrics are better than the previous QSAR and all of the developed q-RASAR models except **P2,** while the external validation metric values were the best among all the developed q-RASAR models as well as the previous QSAR and Read-Across predictions **(Table 4)**. Another important aspect is that the corresponding $MAE_{(TEST)}$ value of model **P1a** is the least among all of the individual and pooled descriptor RASAR models, the consensus-based predicted models, the previous QSAR and Read-Across predictions. Evidently, this particular descriptor (g_m) contributes positively to the receptor binding affinity of a particular target compound. It is also essential to note that the importance of g_m(GK) is much higher as compared to the g(GK) values of all the other models and this is reflected in the Variable Importance Plot of **P1a** (**Figure S4 in Supplementary Materials SI-2**). In contrast, the importance of the previous g(GK) descriptor is much lower as evident from the Variable Importance Plots of **M1 (Figure 4)**, **P1**, **P2** and **P3** (Figures not shown). This indicates the impact of this newly developed descriptor g_m(GK) in influencing the receptor binding affinity of a particular compound. It is expected that g_m will also make a major contribution while modeling other endpoints using the RASAR technique.

It is also evident that the contributions of the measures SD_similarity, SD_activity and g are dependent on whether the MaxPos value is higher than MaxNeg or vice versa. Thus, contributions of these measures should be explained in a relative sense along with MaxPos and MaxNeg values. A heat map of the scaled (0 -1) values of MaxPos, MaxNeg, gm and logRBA (observed) values of the query compounds is shown in **Figure 9**.

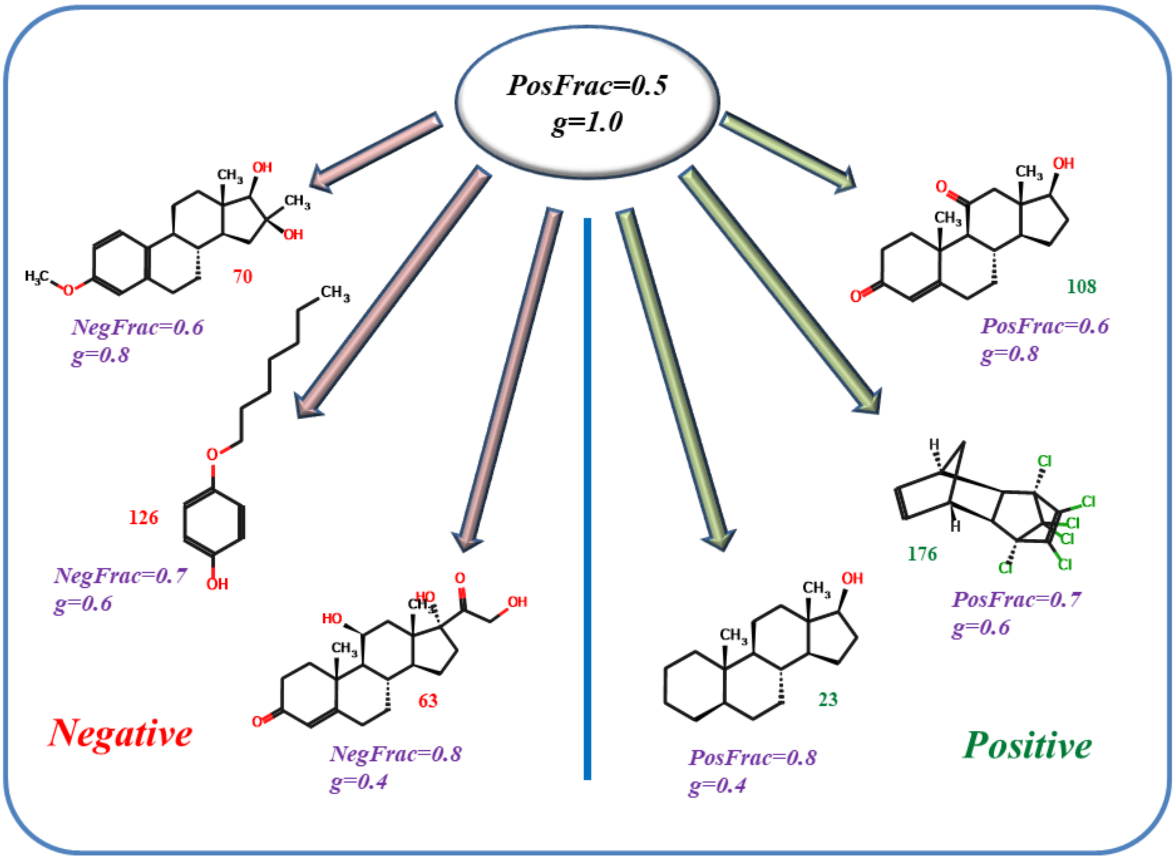

Figure 8. Ambiguous values of *g*.

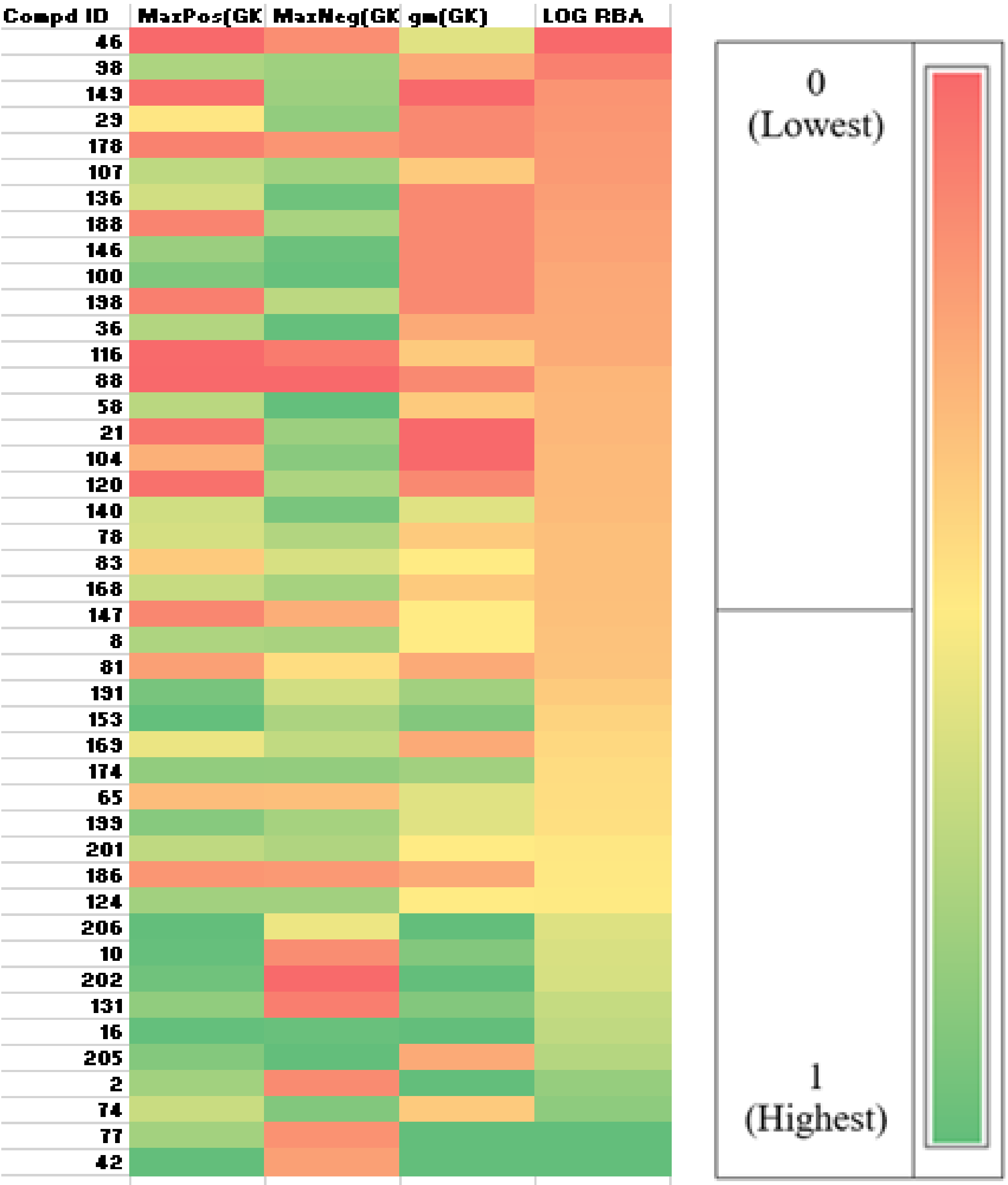

Figure 9. Heat map of the scaled (0 -1) values of MaxPos, MaxNeg, g_m and logRBA (observed) values of the query compounds.

4. Conclusion

The importance of QSAR in designing bioactive chemicals and drugs has been well established to the biological chemistry community though classical examples Hansch analysis, Free-Wilson model, Fujita-Ban modification, topological descriptors, etc. followed by gradual evolution of high dimensional QSARs.[20, 21] The concept of read-across has mainly been applied in regulatory toxicology for data gap filling. RASAR modeling has evolved by combining the concept of read-across and QSAR leading to the generation of statistical relationships (QSAR like) with similarity – based measures (like read-across) along with physicochemical descriptors. The previously reported RASAR models have mainly been applied for classification – based problems, and they have provision to consider multiple endpoints simultaneously with the application of machine learning techniques (data fusion RASAR) which may eventually be related to relevant AOPs with underlying MIEs. In the present communication, we have developed, for the first time, quantitative RASAR (q-RASAR) models taking a single data set of androgen receptor binding affinity. These models could supersede the previously developed QSAR model in the quality of external validation. q-RASAR models have the advantage over the read-across approach in their ability to make convincing conclusion about the quantitative contributions of different features towards the quantitative response values. The workflow can also identify potential outliers in the training set having less similar congeners in the data set thus helping in developing more robust models. It is also possible to apply the concept of consensus modeling as has been done in the present study demonstrating further enhancement of external validation quality even over read-across predictions. We have also developed in the current study a novel concordance measure (g_m) which showed a significant importance in the q-RASAR models. The model generated using this measure outperformed all other models (other q-RASAR models, previous QSAR and read-across predictions) in the external validation metric values. It appears that this concordance measure should further be studied in a greater detail for its potential application in q-RASAR modeling involving other endpoints of pharmaceutical, property, toxicological and other diverse applications. The q-RASAR strategy described in this study should be extendable to more complex problems involving multiple endpoints possibly with the application of more sophisticated machine learning techniques.

- **Data and Software availability**

The data set was originally collected from the Endocrine Disruptor Knowledge Base (EDKB) database (https://www.fda.gov/science-research/bioinformatics-tools/endocrine-disruptor-knowledge-base) and is available in Supplementary Information in the Excel format. 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 the Excel format.

Supplementary Material SI-2 contains score plots, applicability domain plots and randomization plots of models M1 to M4.

- **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

- **Author Contributions**

The manuscript was written through contributions of both authors. Both 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 Jadavpur University, Kolkata for a scholarship.

- **Abbreviations**

QSAR (Quantitative structure-activity relationship); RA (Read-across); RASAR (Read-across structure activity relationship); PLS (Partial least squares); ICP (Intelligent consensus predictions); CM (Consensus model); SD (Standard deviation); CV (Coefficient of variation)

References

(1) Bowes, J.; Brown, A. J.; Hamon, J; Jarolimek, W.; Sridhar, A.; Waldron, G.; Whitebread, G. Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nature Reviews Drug Discovery*, **2012**, *11*, 909.

(2) Knapen, D.; Angrish, M. M.; Fortin, M. C.; Katsiadaki, I.; Leonard, M.; Margiotta-Casaluci, L.; Munn, S.; O'Brien, J. M.; Pollesch, N.; Smith, L. C.; Zhang, X.; Villeneuvei, D. L., Adverse outcome pathway networks I: Development and applications. *Environmental Toxicol. Chem.*, **2018**, *37(6)*, 1723–1733.

(3) Maldonado, A, G,; Doucet, J. P.; Petitjean M.; Fan, B.T., Molecular similarity and diversity in chemoinformatics: From theory to applications. *Mol. Divers.* **2006**, *10*, 39–79.

(4) Schultz, T. W.; Amcoff, P.; Berggren, E.; Gautier, F.; Klaric, M.; Knight, D. J.; Mahony, C.; Schwarz, M.; White, A.; Cronin, M.T.D., A strategy for structuring and reporting a read-across prediction of toxicity. *Regul. Toxicol. Pharmacol.*, **2015**, *72*, 586–601.

(5) Luechtefeld, T.; Maertens, A.; Russo, D. P.; Rovida, C.; Zhu, H.; Hartung, T., Analysis of publically available skin sensitization data from REACH registrations 2008-2014. *ALTEX* **2016**, *33(2)*, http://dx.doi.org/10.14573/altex.1510055

(6) 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.

(7) Hemmerich, J.; Ecker, G. F., In silico toxicology: From structure–activity relationships towards deep learning and adverse outcome pathways. *WIREs Comput Mol Sci.* **2020**, 10: e1475. https://doi.org/10.1002/wcms.1475

(8) Diwan, M.; AbdulHameed, M.; Liu, R.; Schyman, P.; Sachs, D.; Xu, Z.; Desai, V.; Wallqvist, A., ToxProfiler: Toxicity-target profiler based on chemical similarity. *Comput. Toxicol.* **2021**, *18*, 100162, https://doi.org/10.1016/j.comtox.2021.100162

(9) Knapen, D.; Angrish, M. M.; Fortin, M. C.; Katsiadaki, I.; Leonard, M.; Margiotta-Casaluci, L.; Munn, S.; O'Brien, J. M.; Pollesch, N.; Smith, L. C.; Zhang, X.; Villeneuve, D. L. Adverse outcome pathway networks I: Development and applications. *Environ. Toxicol. Chem.*, **2018**, *37(6)*, 1723–1733, 2018.

(10) Wu, J.; D'Ambrosi, S.; Ammann, L.; Stadnicka-Michalak, J.; Schirmer, K.; Baity-Jesi, M., Predicting chemical hazard across taxa through machine learning, *Environ. Int.*, **2022**, *163*, 107184. doi:10.1016/j.envint.2022.107184

(11) Banerjee, A.; De, P.; Roy, K. Assessment of androgen receptor binding affinity of endocrine disruptors: A 2D-QSAR approach, in: Int. Conf. Emerg. Trends Drug Discov. Dev., 2022. https://conferences.vensel.org/event/2/contributions/22/.

(12) 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.

(13) Wu, J.; D'Ambrosi, S.; Ammann, L.; Stadnicka-Michalak, J.; Schirmer, K.; Baity-Jesi, M., Predicting chemical hazard across taxa through machine learning, Environ. Int., 2022, 163,107184. doi:10.1016/j.envint.2022.107184

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

(15) Roy, K.; Kar, S.; Das, R.N., Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment, Academic Press, NY, 2015.

(16) 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 Sys*, **2016**, *152*, 18-33.

(17) Gramatica, P.; Sangion, A., A historical excursus on the statistical validation parameters for QSAR Models: A clarification concerning metrics and terminology. *J. Chem. Inf. Model.* **2016**, *56*, 1127–1131.

(18) Roy, K.; Ambure, P.; Kar, S.; Ojha, P.K., Is it possible to improve the quality of predictions from an "intelligent" use of multiple QSAR/QSPR/QSTR models? *J. Chemom.* **2018**, *32*, e2992, http://dx.doi.org/10.1002/cem.2992

(19) Eriksson, L.; Byrne, T.; Johansson, E.; Trygg, J.; Vikström, C., Multi- and Megavariate Data Analysis Basic Principles and Applications, Umeå: Umetrics Academy, 2013.

(20) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A., QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.*, **2014**, *57*, 4977-5010.

(21) Fujita, T.; Winkler, D. A., Understanding the roles of the "Two QSARs", *J. Chem. Inf. Model*., **2016**, *56(2)*, 269-274.