Quantitative Predictions from Chemical Read-Across and Their Confidence Measures

Arkaprava Banerjee, Mainak Chatterjee, Priyanka De 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

In silico modeling new approach methodologies (NAMs) are viewed as a promising starting point for filling the existing gaps in safety and ecosafety data. Read-across is one of the most widely used alternative tools for hazard assessment, aimed at filling data gaps. However, there are no systematic studies or recommendations on the measures to identify the quality of read-across predictions for the data points without any experimental response data. Recently, we have reported a new similarity-based read-across algorithm for the prediction of toxicity (biological activity in general) of untested compounds from structural analogues (the tool available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home). Three similarity estimation techniques such as, Euclidean distance-based similarity, Gaussian kernel function similarity, and Laplacian kernel function similarity are used in this algorithm. As the confidence of predictions for untested compounds is an important information, we have addressed this issue here by consideration of several similarity and error – based criteria. The role of these measures in discriminating high and low residual query compounds is studied in three different approaches: (a) comparison of means of a measure for high and low residual groups; (b) development of classification models for absolute residuals to identify the contributing measures; (c) application of the sum of ranking differences (SRD) approach to identify the measures closer to the reference rank defined by the absolute residuals. Finally, the frequency of occurrences of different measures in the three approaches is compared. The results from three data sets with 10 divisions of source and target compounds in each case indicate that weighted standard deviation of the predicted response values appear to be the most deterministic feature for the reliability of predictions followed by different similarity-based features. The derived reliability measures will provide a greater confidence to the quality of quantitative predictions from the chemical read-across tool for new query compounds.

Keywords: Read-across; Similarity; Prediction; Residual; Discriminant function

Introduction

Computational prediction tools are designed and developed as an alternative to experimental biological activity/toxicity tests in order to potentially minimize the need for animal testing, reduce the associated cost and time required for such experimental studies, and improve the quality and availability of data from activity/toxicity prediction and risk/safety assessment [1, 2]. More importantly, *in silico* tools can estimate activity/toxicity of virtual compounds even before their synthesis thus minimizing the cost involved in the synthesis and testing of less potential or less prioritized chemicals. This can help design industrial chemicals/drug candidates with better toxicity/pharmacokinetic profile and prioritize them for experimental testing. Computational methods of toxicity predictions are accepted as tools to bridge data gaps by regulatory agencies like Organization of Economic Cooperation and Development (OECD), European Chemicals Agency (ECHA), Food and Drug Administration (FDA), etc [3-6].

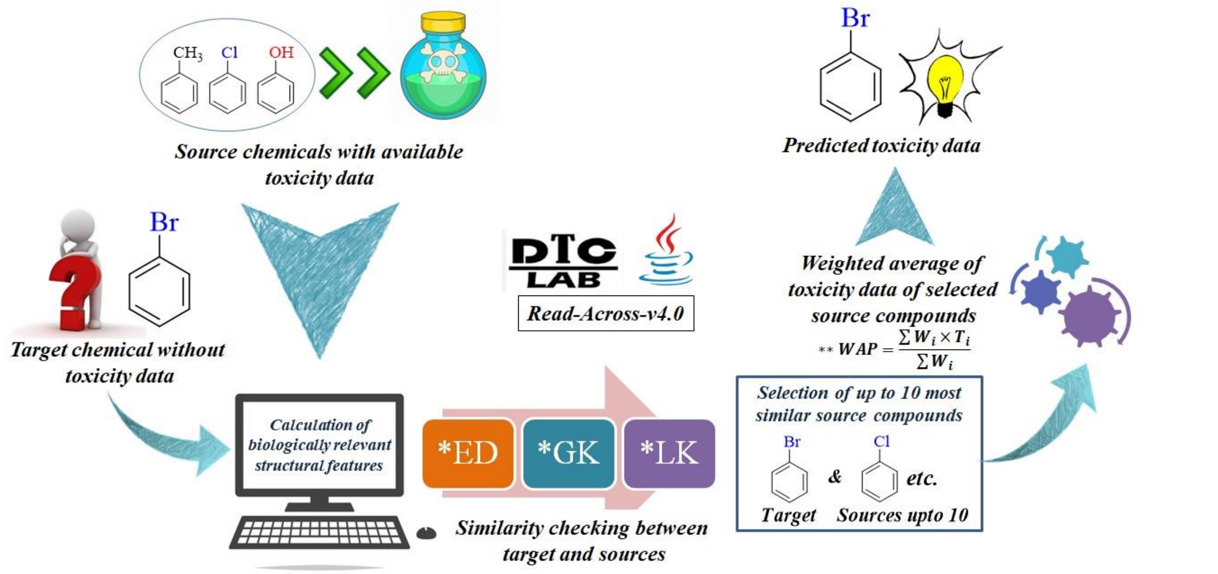
Among various *in silico* techniques for data gap filling, quantitative structure-activity relationship (QSAR) modeling is a popular method [7]. QSAR is a statistical model building process requiring sufficient number of data points for meaningful model development. In addition, in most of the cases, the data points available are required to be split into training and test sets for validation purpose in order to comply with the requirements as recommended by the OECD (https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm). Thus, a portion of the available experimental data cannot be used for model building and are kept aside for model validation. In case of small data sets, such waste may lead to statistically less reliable model development. Read-across, a chemical similarity-based grouping technique [8], can better address the situation as it does not rely on statistical model development. It is a non-animal alternative data gap filling method that provides information for biological activity/toxicological risks of *target* compounds derived from known activity/toxicity data of *source* compound(s) with a *similar* property or chemical profile. It is one of the most important contemporary *in silico* approaches which is majorly applied in the ecotoxicological data generation, data gap filling, and regulatory decision making. The qualitative read-across approach is most popular and widely used by the regulatory authorities, although the use of quantitative read-across methods has also been seen in the recent past. The query chemicals are mostly termed as the target chemicals whereas the chemical analogues with known toxicity data are called source chemicals. In common practice, read-across predictions are obtained by analogue and category approaches. The analogue approach essentially takes a single source chemical for the prediction, whereas more than one source chemicals are used in the category approach; thus it is more robust and reliable one. Easy algebraic calculations are used in the quantitative read-across algorithm which makes it a computationally less exhaustive process. Apart from that, this method is also an effective approach for the prediction of toxicity of small datasets due to the use of simple calculation (independent of statistical operations). The weighted average of toxicity data (**equation 1**) of chemical analogues is a way for the prediction of untested chemicals.

$$Weighted\ average = \frac{\sum W_i \times X_i}{\sum W_i} \qquad (1)$$

where, W_i is the weightage of i^{th} source compounds which is calculated based on the similarity with the target compound; X_i is the toxicity of the corresponding source compound *i*.

For a successful read-across operation, the identification of chemical category and the associated uncertainty of this identified category is very important to claim the reliability of predictions. The major objective of the read-across technique is to provide prediction data that is thought to be (more or less) equivalent to the omitted standard experimental assay, and hence this has been applied mainly for toxicity/ecotoxicity data gap filling of chemicals in a regulatory context, However, these new approach methods (NAMs) are finding applications in several other regulatory frameworks, including in the assessment of impurities and degradation products of pharmaceuticals, assessment of plant protection product metabolites, extractables from personal protective and medical devices, food-contact substances, and cosmetics [9]. Structural similarity and similar properties, fate and/or activities between the source and target chemicals provide a convenient means of identifying likely analogues and are thus used as a basis for justifying read-across [10]. Apart from only the structural similarity consideration, one should additionally consider physico-chemical properties, reactivity and metabolism, and mechanistic similarity for the precision of predictions [11]. In this direction, the researchers may perform the grouping based on changes in structural aspects and physico-chemical properties and possible fates, degradation and/or the mode of metabolism. Furthermore, identification of the experimental data gaps in physico-chemical characterization, exposure and hazard assessments within the defined groups/categories should also be done [10, 11]. For regulatory acceptance, a read-across prediction should be robust, reliable and easily explicable. Two important aspects of any read-across predictions are the degree of similarity between target(s) and source substance(s) and defining the uncertainties in the read-across predictions [12]. It is generally accepted that the reliability of a read-across prediction depends on the aspects of the defined similarity and the type and degree of uncertainty associated with the particular read-across. Therefore, addressing these two elements in an unambiguous manner is of utmost necessity. Although there are several reports on read-across predictions for different toxicity and ecotoxicity endpoints [13-16], there are no systematic studies and recommendations on the measures of reliability of quantitative read-across predictions for new query compounds. The current manuscript addresses this gap and explores the important measures that may be used to identify the quality of quantitative read-across predictions in absence of experimental data.

Development of novel algorithms for read-across predictions is a topic of contemporary research in regulatory toxicology. Extensive research has not yet been done for developing algorithms of quantitative read-across predictions. This is especially interesting when limited experimental data for the endpoint of interest is available. Recently, we have developed a read-across tool that predicts the endpoint data of query chemicals based on chemical similarity to the available source compounds using the Euclidean, Gaussian kernel or Laplacian kernel-based similarity functions **(Figure 1)** [17]. We have also applied this tool for prediction of nanotoxicity data of three different data sets showing better quality of predictions than the previously reported read-across and QSAR predictions [17]. However, it is indeed important to know the reliability of read-across predictions for new query compounds without having experimental response values thus not allowing a comparison of predictions with the observed responses. There must be some measures and features that would provide us with confidence or uncertainty of predictions in such cases. The present communication tries to explore the factors governing the reliability of predictions for new query chemicals using the read-across prediction tool.


CH3
Cl
OH
Source chemicals with available toxicity data
Br
Predicted toxicity data
Br
Target chemical without toxicity data
DTC LAB
Read-Across-v4.0
Weighted average of toxicity data of selected source compounds
** WAP = ΣWi×Ti / ΣWi
Calculation of biologically relevant structural features
*ED
*GK
*LK
Similarity checking between target and sources
Selection of up to 10 most similar source compounds
Br
Cl
&
etc.
Target
Sources upto 10

**ED: Euclidean distance-based similarity; GK: Gaussian kernel function similarity; LK: Laplacian kernel function similarity*
***WAP: Weighted average predictions; W_i: weightage of i^{th} source compound (based on similarity); T_i: toxicity of i^{th} source compound*

Figure 1. General workflow of Read-across predictions

Materials and Methods

The read-across predictions have been done using the Read-Across-v4.0 tool available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. In this tool, for each query compound, up to 10 close source compounds are selected based on the similarity measure (Euclidean, Gaussian kernel and Laplacian kernel–based similarity), and read-across predictions are made using a weighted average approach [17]. In the output, weighted average prediction ($\overline{x_{wtd}}$) values along with weighted standard deviation ($s_{weighted}$) and weighted standard error ($(s_{\bar{x}})_{weighted}$) are reported [18]. The details are available in Supplementary Materials SI-1. The user can choose the number of close source compounds to be used by the tool, optimize the hyper-parameters for Gaussian and Laplacian kernel functions and provide with the distance and similarity threshold values.

We have used in the current study three different data sets recently used by us for QSAR modeling: (1) acute contact toxicity of plant protection products against honey bees [19], (2) Bobwhite Quail ecotoxicity data [20], and androgen receptor binding affinity [21]. We have used the same physicochemical features as reported in the original QSAR reports for the present study. For each data set, we have used the original division pattern (training and test sets) in addition to nine additional new divisions made using a variety of approaches like sorted response, Kennard-Stone, k-medoids, and random division [22] maintaining the similar training-test size ratio and ensuring diversity in composition. We have used here the term "training set" for the whole set of source compounds and the term "test set" for the whole set of query compounds. For the original division of each data set, optimization of hyper-parameters and distance and similarity threshold settings were done based on a sub-training set and a validation set. The optimized settings were used for the original division and nine additional divisions for read-across predictions. As our objective of this study is not to obtain the best predictions from a given data set, and it is rather to explore the features indicating the quality of quantitative predictions, we have not done optimization of the settings separately for each division.

The 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 source compounds for each query compound, average and standard deviation of similarity levels and their coefficient of variation of similar close compounds to each query compound, maximum similarity level to positive and negative compounds (based on the whole "training set" response mean), a concordance measure indicating similarity to positive, negative or both classes of close source compounds [23], etc. as detailed in **Table 1**.

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

Measure	Description	Comment	Formula
SD_activity ($s_{weighted}$)	Standard deviation of the (observed) response values of the selected close source compounds for each query compound	Dispersion measure	$s_{weighted} = \sqrt{\frac{\sum_{i=1}^{n} w_i (x_i - \overline{x_{wtd}})^2}{\sum_{i=1}^{n} w_i} \times \frac{n}{n-1}}$ where, $\overline{x_{wtd}} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$, w_i is the respective weight for the response x_i, n is the number of data points used in computation of the average.
CV_activity	Coefficient of variation of the response	Relative Error measure	$CV_{activity} = \frac{s_{weighted}}{\overline{x_{wtd}}}$
Average similarity	Mean similarity to the close source	Similarity measure	$Similarity_{average} = \frac{\sum_{i=1}^{n} Similarity_i}{n}$

	compounds for each query compound		
SD_similarity	Standard deviation of the similarity values of the selected close source compounds for each query compound	Dispersion measure	$s_{similarity=}\sqrt{\frac{\sum_{i=1}^{n}(Similarity_i - \overline{S\imath m\imath lar\imath ty})^2}{n-1}}$ where $\overline{S\imath m\imath lar\imath ty} = \ Similarity_{average}$
MaxPos	Maximum Similarity level to Positive close source set compounds (based on the "training set" observed mean)	Similarity measure	
MaxNeg	Maximum Similarity level to Negative close source set compounds (based on the "training set" observed mean)	Similarity measure	
AbsDiff *or* Abs(MaxPos-MaxNeg)	Absolute difference between MaxPos and MaxNeg	Similarity measure	$AbsDiff = \lvert MaxPos - MaxNeg \rvert$
g	This is a concordance measure	Similarity measure	$g = 1 - 2 \times \lvert PosFrac - 1/2 \rvert$ where *PosFrac* is the fraction of the close source compounds belonging to the Positive Class based on the "training set" response mean as the threshold [23].

We have attempted to understand the role of the above measures in determining the quality of quantitative read-across predictions for new query compounds. We have done this analysis using three different strategies by studying:

1. The frequency of measures showing statistically significant differences between the corresponding means of high residual and low residual target set compounds

2. The frequency of measures appearing important in the developed classification models for high and low residual target compounds

3. The frequency of measures found important to rank target compounds based on their absolute residuals in the Sum of Ranking Difference approach.

The above three strategies were applied to ten different divisions (source and target compounds) of three different data sets (Figure 2).

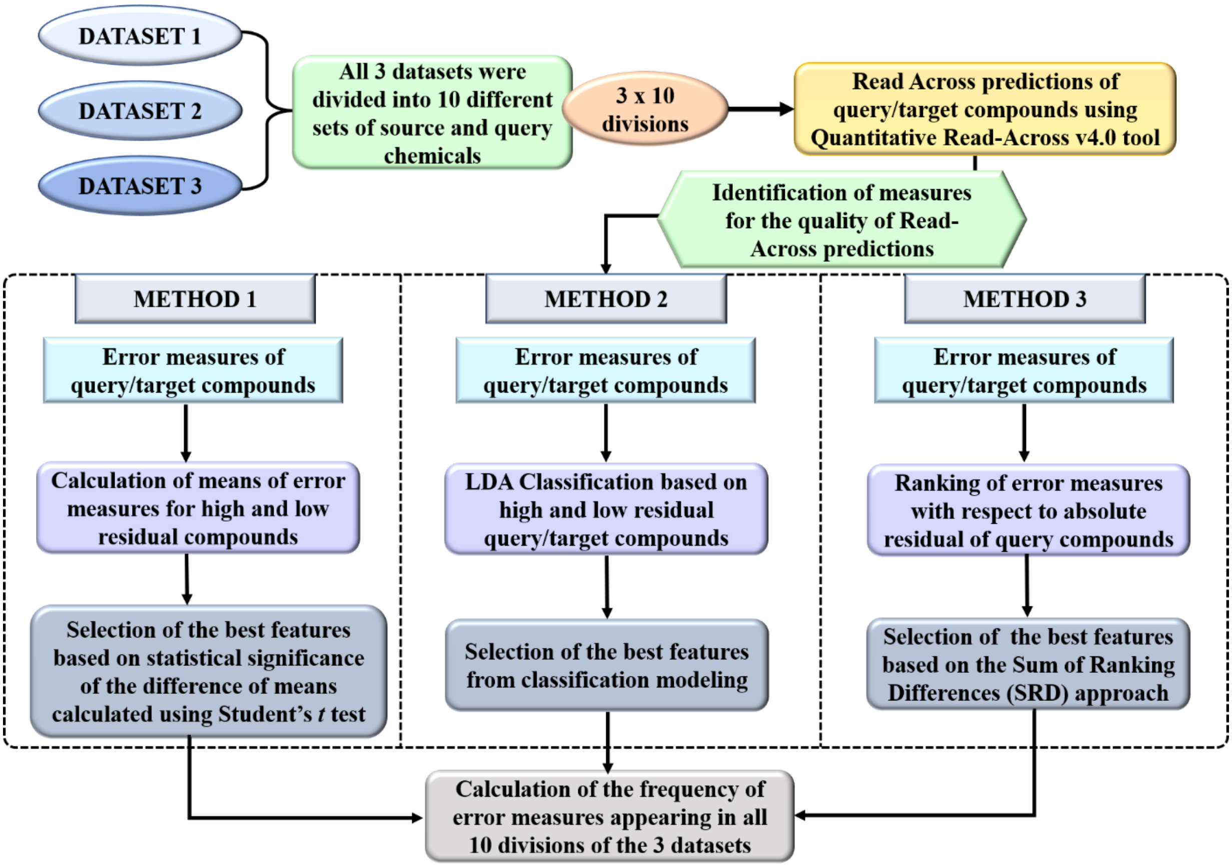

DATASET 1
DATASET 2
DATASET 3
All 3 datasets were divided into 10 different sets of source and query chemicals
3 x 10 divisions
Read Across predictions of query/target compounds using Quantitative Read-Across v4.0 tool
Identification of measures for the quality of Read-Across predictions
METHOD 1
Error measures of query/target compounds
Calculation of means of error measures for high and low residual compounds
Selection of the best features based on statistical significance of the difference of means calculated using Student's *t* test
METHOD 2
Error measures of query/target compounds
LDA Classification based on high and low residual query/target compounds
Selection of the best features from classification modeling
METHOD 3
Error measures of query/target compounds
Ranking of error measures with respect to absolute residual of query compounds
Selection of the best features based on the Sum of Ranking Differences (SRD) approach
Calculation of the frequency of error measures appearing in all 10 divisions of the 3 datasets

Figure 2. General workflow of the current study.

1. **Study 1. Comparison of means of high residual and low residual groups:** For each set of division of each data set, we have compiled the read-across predictions of the query set compounds along with various similarity and error measures (as in Table 1) in addition to the observed and predicted response values, then ranked the query compounds in the descending order based on predicted residuals, and finally, identified two sets of 10 compounds with the highest and lowest predicted residual values. We have then compared the means of residuals and different similarity and error measures of the two sets of compounds (high residual and low residual compounds) to identify the important similarity and error measures showing significant difference between the two groups of compounds. The *t* test for comparison of means of two groups [24] was used for this purpose (Supplementary Materials SI-1) with the Gaussian distribution assumption of both the classes.

2. **Study 2. Linear discriminant analysis of graded residuals using error and similarity measures:** We used the compiled data of residual values along with different similarity and error measures as described under Study 1 above and graded the data points as positive (1) or negative (0) based on the mean residual value of the corresponding “training set”. Then we used the graded response as the dependent variable (Y) and different similarity and error measures and the predicted activity as the independent variables (X) for developing linear discriminant analysis (LDA) models using stepwise variable selection with the F-to-enter 4 and F-to-remove 3.9 setting (in most of the cases) using SPSS statistics software [25]. The LDA tries to maximize the variance between the classes while minimizing the within-class variance, using a linear discriminant function [26]. This also assumes that data in every class are described by a Gaussian probability density function with the same covariance. A linear discriminant function, which is a linear combination of the independent (X) variables, divides the feature space by a hyperplane decision surface. Although we understand that it is overoptimistic to model predicted residuals or errors in this approach and hence, we do not aim at obtaining a perfect classifier, this exercise will definitely help identifying important measures and throwing a light on the reliability of predictions.

3. **Study 3. Application of the sum of ranking differences (SRD) to identify the important measures for ranking the query compounds based on their quality of predictions**

   The sum of ranking differences [27] is a useful way to compare metrics, methods, models, methods, analytical techniques, etc. in a general manner. We have used this method to compare the performance of various similarity metrics to understand the quality of read-across predictions. Here, the cases (query compounds) to be ranked are placed in the rows and the metrics in the columns of an input matrix. Then, the results of each metric for each case are ranked in the order of increasing magnitude. The difference between the rank of the metric results and the rank of the known or standard results (here absolute residuals) is then computed. This is followed by the calculation of the sum of absolute values of the differences for all metrics. A lower value of SRD (close to 0) indicates a better metric. The closeness of SRD values indicates the similarity of the metrics, whereas large variation indicates dissimilarity. A permutation test is used for the validation of the SRD method which uses a recursive algorithm for the computation of the discrete distribution for a small number of objects (n<14) or the normal distribution if the number of objects is large. The theoretical distribution is visualized for random numbers and it can be used to identify SRD values for metrics that are far from being random. A random resampling with sevenfold cross-validations is also applied to validate the obtained results, and the results are presented with a Box-Whiskers plot of the cross-validated SRD values [28]. The SRD runs were made using the program available from http://aki.ttk.hu/srd/.

The results obtained from Studies 1 to 3 above are compared and discussed to conclude on the features responsible for the reliability of quantitative read-across predictions. Please note that the objective of the present analysis is not making new predictions for the data sets being considered or comparing them to the previously reported analysis. We try to explore here various features that may be useful in determining the uncertainty of quantitative predictions from the read-across tool for new query compounds.

Results and Discussion

We present here the results obtained from the two strategies of our analysis. The frequency of occurrences of different measures for discriminating high and low residual compounds at *p* <=

0.05 is shown in **Figure 3** while that for developed LDA models for predicting the class of high or low residuals is shown in **Figure 4** (in addition to **Supplementary Materials SI-1**). The frequency of occurrences of different measures for correctly ranking high and low residual compounds as per the SRD analysis is shown in **Figure 5** (also see **Supplementary Materials SI-1).** The details of the results and raw data are available in **Supplementary Materials SI-2** and on request from the authors.

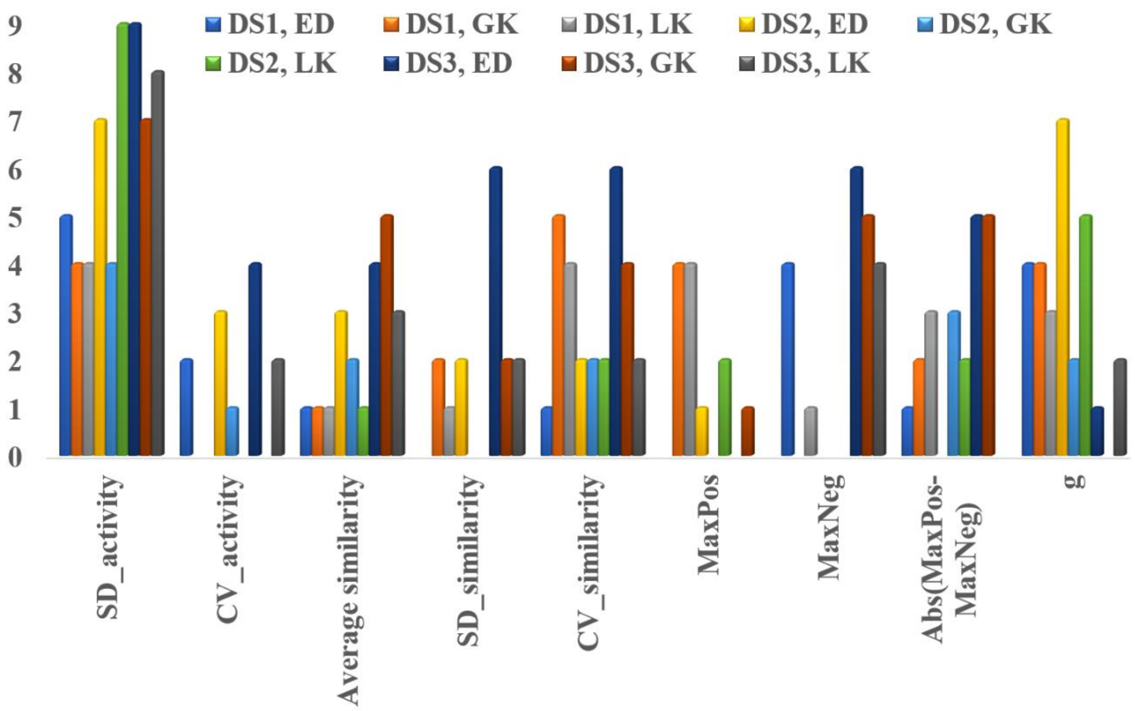


Figure 3. Frequency of occurrences of different dispersion and similarity measures in differentiating high and low residual compounds (out of 10 trials for each division of each data set) (DS1 = data set 1, DS2 = Data set 2, DS3 = Data set 3, ED = Euclidean distance, GK = Gaussian kernel, LK = Laplacian kernel).

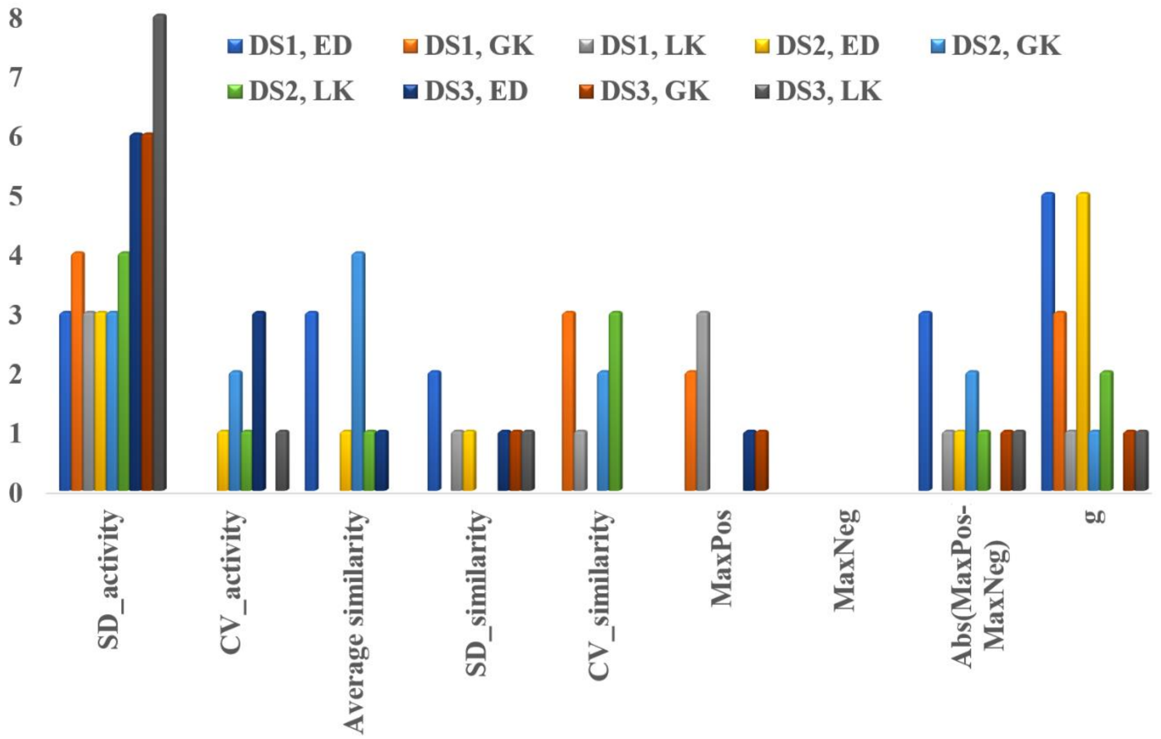


Figure 4. Frequency of occurrences of different dispersion and similarity measures in the developed LDA models for predicting the class of high or low residuals (out of 10 trials for each division of each data set) (DS1 = data set 1, DS2 = Data set 2, DS3 = Data set 3, ED = Euclidean distance, GK = Gaussian kernel, LK = Laplacian kernel).

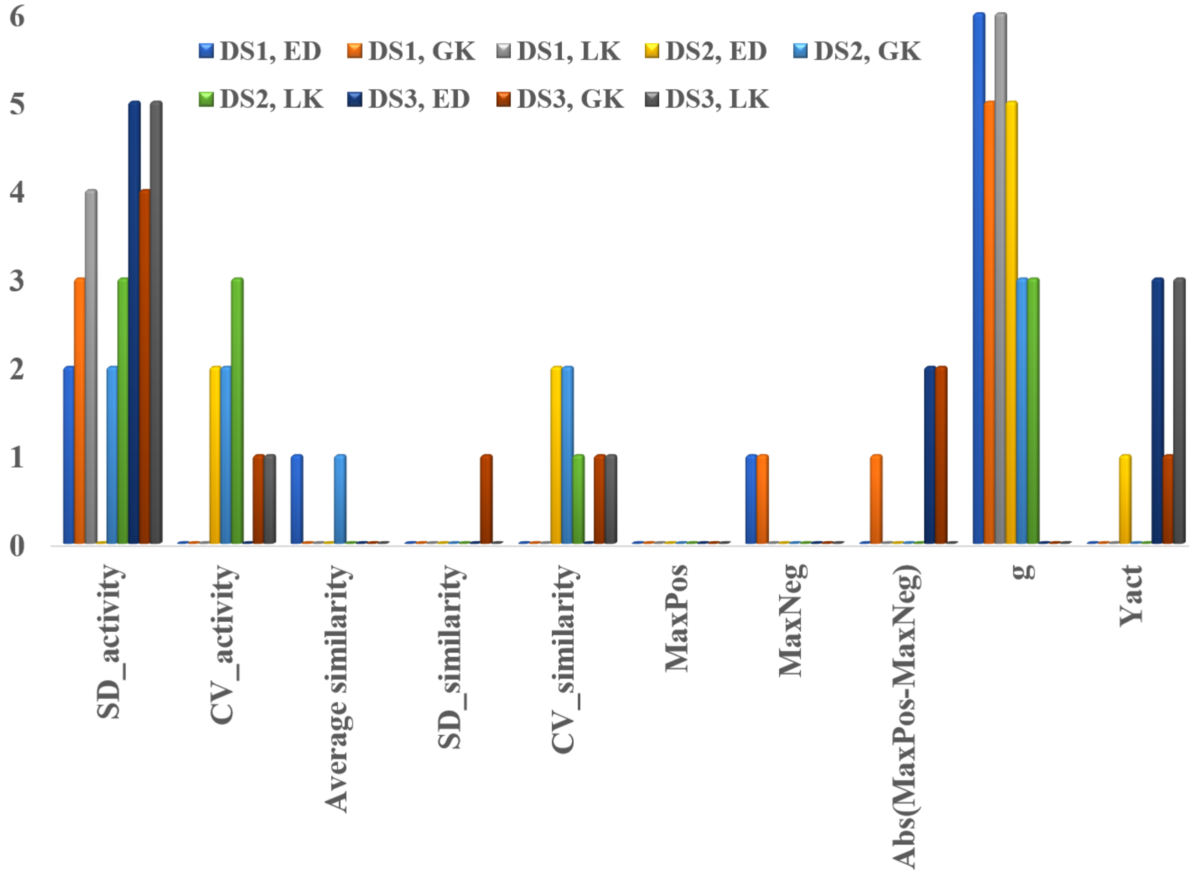


Figure 5. Frequency of occurrences of different dispersion and similarity measures for correctly ranking high and low residual compounds as per the SRD analysis ((DS1 = data set 1, DS2 = Data set 2, DS3 = Data set 3, ED = Euclidean distance, GK = Gaussian kernel, LK = Laplacian kernel).

Study 1.

Figure 1 shows the difference in arithmetic means of different similarity and error measures along with the predicted residuals between the two groups of compounds (high residual and low residual compounds) in the query sets of different divisions of the three data sets as has been found significant at p<0.05 based on the *t* test for comparison of means. It is obvious from the results (**Supplementary Materials SI-2**) that the residual means show statistically significant difference between the two groups in all cases. Among the other measures, we have presented here only those which show statistically significantly difference at *p* <=0.05 **(Figure 3).** For Data set 1, in case of the Euclidean distance - based read-across predictions, SD_activity, MaxNeg and *g* occur 4 times or more (out of 10 trials); in case of the Gaussian kernel-based predictions, SD_activity, CV_similarity, MaxPos and *g* occur 4 times or more (out of 10 trials); in case of Laplacian kernel - based predictions, SD_activity, CV_similarity and MaxPos occur 4 time or more (out of 10 trials). It is to be noted that SD_activity occurs as the most influential feature considering all similarity-based read-across prediction methods while CV_similarity and MaxPos occur in case of two similarity-based prediction methods. For Data set 2, SD_activity and *g* emerge as the most significant discriminating features: in case of Euclidean distance - based predictions, both of them occur 7 times out of 10 trials; in case of Gaussian kernel- based predictions, SD_activity occurs 4 times, in case of Laplacian kernel-based predictions, SD_activity occurs 9 times while *g* occurs 5 times out of 10 trials. For Data set 3, in case of Euclidean distance- based predictions, SD_activity appears nine times followed by predicted response (8 times), observed response, CV_similarity, SD_similarity and MaxNeg (6 times each), absolute difference between MaxPos and MaxNeg (5 times), CV_activity and average similarity (4 times each); in case of Gaussian kernel - based predictions, the most frequently appearing measure is SD_activity (7 times) followed by observed and predicted responses, MaxNeg, average similarity and absolute value of difference between MaxPos and MaxNeg (5 times each) and CV_similarity (4 times); in case of Laplacian kernel- based predictions, the most frequently appearing measures are SD_activity and predicted response (8 times each), observed response (6 times), MaxNeg (4 times). Interestingly, observed and predicted response also show statistically significantly different means in considerable number of trials for Data set 3.

A close analysis of the results from three data sets **(Figure 6a)** shows that SD_activity is the most frequently appearing feature for all three data sets. SD_activity corresponds to the dispersion of the observed responses of close source compounds from which a target compound is predicted. If this dispersion is higher, the reliability of predictions for the query compound will also be lower as obvious from the high residual values in such cases. The next important measures are *g* and CV_similarity, each of which occurs for 28 times for the three data sets. While *g* appears to be important for Datasets 1 and 2, it is not so for Data set 3 where CV_similarity is more important. It appears that either *g* (concordance measure) or CV_similarity level (along with the average similarity level) is very important in discriminating the high and low residual chemicals. The absolute difference between MaxPos and MaxNeg (especially for Data set 3) is also found somewhat important.

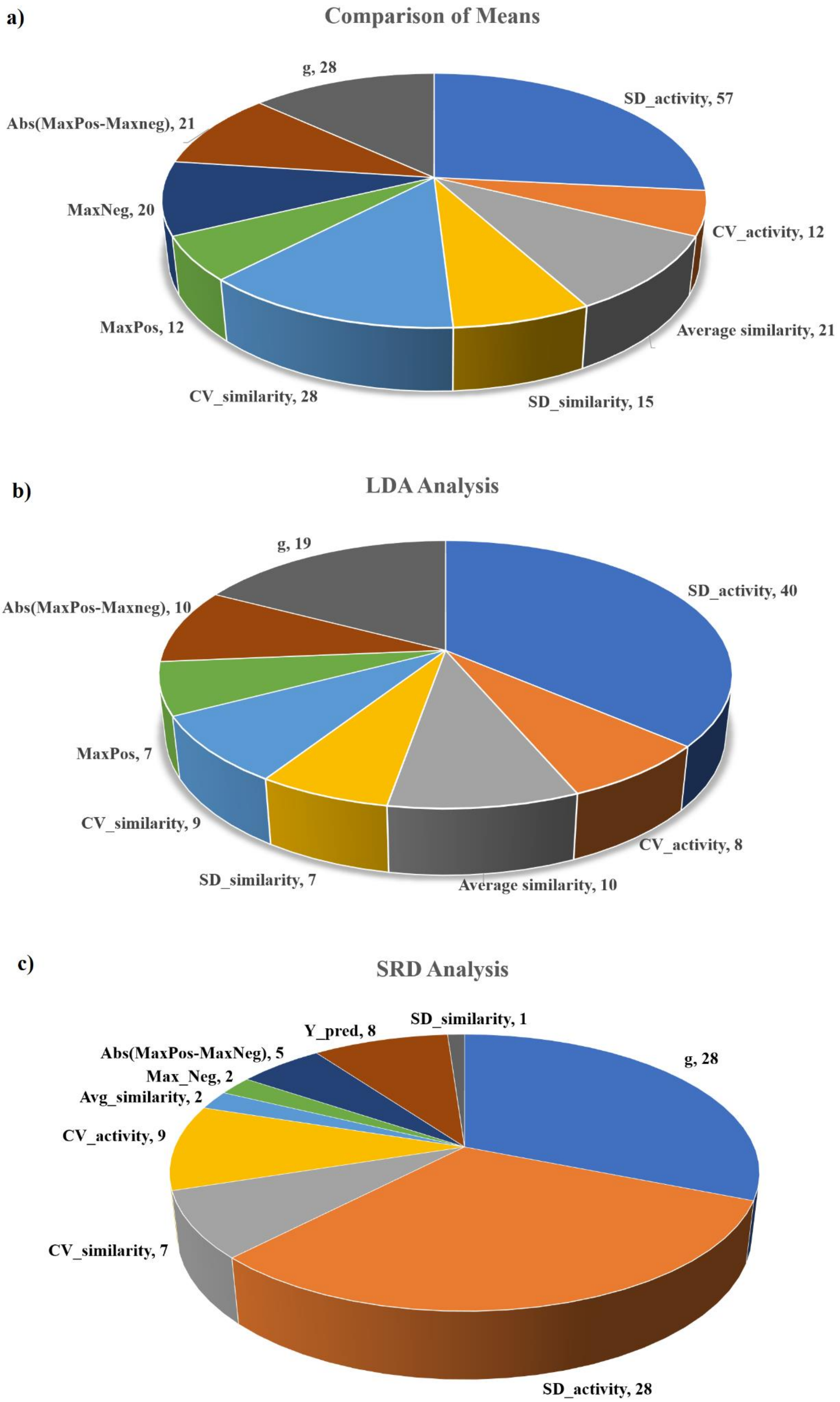

a)
Comparison of Means
g, 28
SD_activity, 57
Abs(MaxPos-Maxneg), 21
MaxNeg, 20
CV_activity, 12
Average similarity, 21
MaxPos, 12
CV_similarity, 28
SD_similarity, 15
b)
LDA Analysis
g, 19
SD_activity, 40
Abs(MaxPos-Maxneg), 10
MaxPos, 7
CV_similarity, 9
CV_activity, 8
SD_similarity, 7
Average similarity, 10
c)
SRD Analysis
SD_similarity, 1
Y_pred, 8
Abs(MaxPos-MaxNeg), 5
Max_Neg, 2
Avg_similarity, 2
g, 28
CV_activity, 9
CV_similarity, 7
SD_activity, 28

Figure 6. Frequency of occurrences of different important error and similarity measures found from (a) comparison of means between high and low residual compounds; (b) LDA models; (c) SRD analysis

Study 2.

The trends obtained from the LDA analysis cannot be expected to be identical with the previous analysis (Study 1) which was based on the quantitative residual values of two sets of samples (high and low residual compounds) drawn from the individual query set while Study 2 was performed for the graded residuals of the whole query set. In spite of this, one clear trend we find from the frequency of occurrences of different measures (**Figure 6b**) that SD_activity and *g* occur most in the obtained LDA equations followed by average similarity, CV_activity, CV_similarity and Abs(MaxPos-MaxNeg).

Study 3.

Study 3 applies the sum of ranking differences approach in identifying the most suitable measures that rank the query compounds most similar to the ranking based on the absolute residuals (reference).

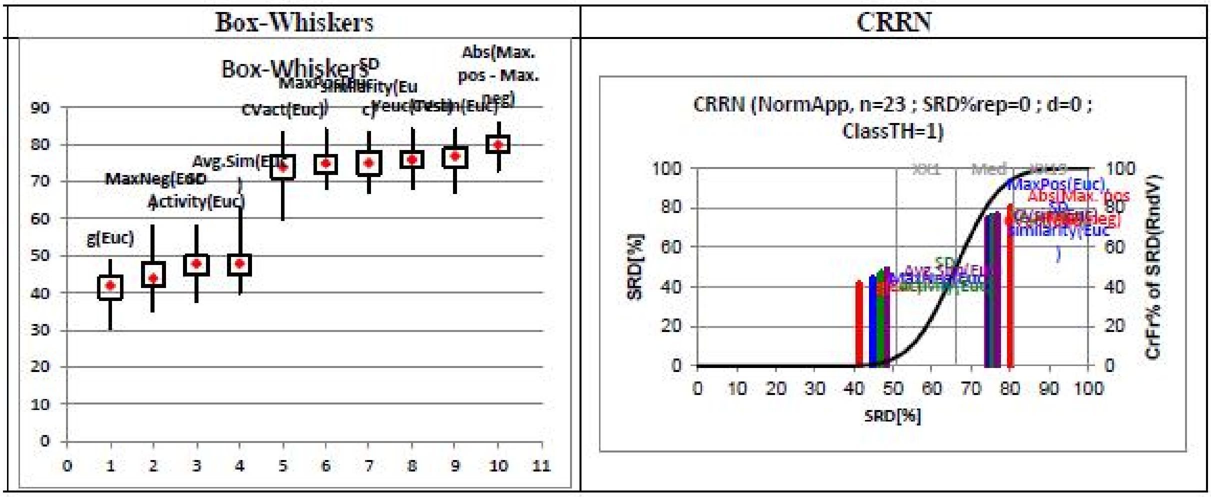


Figure 7. Sevenfold cross-validated SRD results (Box-Whiskers plot) and normalized SRD values (between 0 and 100) compared to random ranking (CRRN or comparison of ranks with ranking numbers) for division 1 of Data set 1.

Figure 7 shows that ranking of the cases as per *g*(ED) is the closest to that of the reference ranking using absolute residuals for Division 1 of Data set 1. The remaining images of other divisions and other data sets are available in **Supplementary Materials SI-1**. Figure 5 shows the frequency of occurring different measures as determinant of rank order similar to the reference ranking. In line with the observations from Studies 1 and 2, SD_activity and g occur most frequently as the most influential indicator of the quality of predictions followed by CV_activity, Y_{Pred} and CV_similarity (**Figure 6c**).

Considering the results from all three studies as discussed above **(Figure 6)**, for read-across predictions of new query compounds, the quality of predictions is thought to be dependent on the following factors:

1. First line diagnostic measures: Dispersion of activity of close source compounds (SD_activity and CV_activity)
2. Second line diagnostic measures: Similarity measures (*g*, CV_similarity and average similarity and Abs(MaxPos-MaxNeg)
3. Level of the predicted activity

Other measures as discussed above would be related to one or more of the above-mentioned measures.

Dispersion of response values of close source compounds: If the dispersion (in the form of standard deviation and coefficient of variation, but mainly standard deviation) of the response values of the source compounds is high, the precision level of the prediction for a query compound will be low. For example, in Data set 1, Division 3 (Laplacian kernel method), the

10 compounds showing highest predicted residuals have mean SD_activity value of 0.847 vs. the 10 compounds having lowest predicted residuals showing SD_activity of 0.600. This difference is significant at p <0.05. This is an inversely proportional relationship suggesting the requirement of selection of optimum number of close source compounds to avoid a high dispersion or error value.

Similarity measures: "*g*" is a concordance measure to indicate whether the close source compounds selected belong to the same class or the other (based on the mean response of the original list of source compounds as the threshold) [23]. If all of them belong to either positive or negative class, the concordance is higher, and the residual is expected to be low. For example, Division 1 (Euclidean distance method) of Dataset 2 shows that the 10 compounds having high residual values have the mean *g* value of 0.44 compared to 0.18 for the 10 compounds having low residual values. This difference is significant at p < 0.05. When all (or most of the) close source compounds belong to the same class (either positive or negative), it is expected that the predicted value will be precise and it will not at least misclassify the prediction for the query chemical.

The average similarity level to the close source compounds is also an important indicator of reliability. From Division 9 of Data set 2, it is seen that the group of 10 query compounds having high residuals have the average similarity level (Euclidean) of 0.882 compared to 0.946 of the group of 10 query compounds having low residual values. If the similarity level increases, reliability of predictions also increases. This is to be noted here that the similarity level of the close source compounds for any query compound is usually higher in case of the Euclidean distance-based approach than the Gaussian kernel based similarity followed by Laplacian kernel based similarity. For this reason, the difference in similarity between the two classes is significant at lower confidence level for the latter two cases. In addition, the interpretation of average similarity is also dependent on data structure (distribution of positive and negative compounds in the data set while the classification is based on the training set response mean as the threshold).

The coefficient of variation of the similarity also plays an important role. For example, in Division 2 (Euclidean distance method) of Data set 3, the average CV_similarity level of the set of 10 compounds having high residual values is 0.066 compared to 0.041 for the set of compounds with low residuals. As the CV of similarity values increases, the reliability of predictions decreases. Similar results are also seen for SD_similarity, but its significance is observed in lower number of cases.

We may note that the trend mentioned here with regard to the average similarity with respect to high and low residual compounds may also be opposite if the level of dispersion of similarity of close source compounds is high. This happens when the number of close source compounds belonging to either positive or negative class is similar (i.e., *PosFrac* is close to 0.5). For example, in case of the Euclidean distance-based similarity of Division 7 of Data set 1, the dispersion of the similarity values of close similar source compounds is relatively higher and the SD_similarity value of lower residual compounds is actually higher than the high residual compounds, while the average similarity value of lower residual compounds is thus lower than the high residual compounds. This depends on the data structure showing the relative number of close similar source compounds belonging to either positive or negative class and in such cases, SD_activity is the main determining factor for the quality of predictions.

The absolute difference between maximum similarity to positive compounds and maximum similarity to negative compounds is also found important in several cases. This difference may be thought to be a perplexity measure. It may be expected that for low residual compounds, this difference may be higher for a more deterministic prediction as observed in case of Laplacian kernel-based similarity of Division 4 of Data set 1. Here, the absolute difference value for the low residual compounds is 0.138 compared to 0.021 in case of high residual compounds. However, the opposite trend is found in Data set 3 where in case of Division 2

(Euclidean distance-based similarity), the absolute difference value for low residual compounds is 0.055 compared to 0.123 in case of high residual compounds. In this case, the SD_similarity value is lower for low residual compounds, while in case of Data set 1, Division 4 (Laplacian similarity), the SD_similarity value is higher for low residual compounds. This explains the observed difference in the impact of absolute difference value which is in turn dependent on the data structure.

Other similarity-based measures like maximum similarity to positive compounds and maximum similarity to negative compounds are also found important in some cases. But their significance depends on the data structure and they are related to other similarity measures already discussed.

Level of predicted values: In some cases, especially in case of Data set 3, either or both of observed or/and predicted response values show statistically significant differences between high and low residual compounds. For example, in case of Division 1 (Euclidean distance method) of Data set 3, the average predicted value of the high residual compounds is -1.266 while that for the low residual compounds is -2.174. This difference is significant at p <0.05. This indicates that a compound predicted to be lower active has more confidence of predictions than a compound predicted as higher active. The uncertainty level of higher level of quantitative predictions is also higher.

Based on the results obtained from the three data sets with their 10 division pattern, we propose here at a preliminary level a set of diagnostic thresholds of different similarity measures (based on Euclidean based similarity) to identify the quality of quantitative predictions (**Table 2**). The first and some of the rest criteria as mentioned in **Table 2** are expected to be met for reliable predictions. Apart from the above, a compound predicted to be more active will have in general less confidence level. However, the indicated thresholds may be more refined in the future with the availability with additional results with other data sets.

Table 2. Desired level of different dispersion/similarity measures for good reliability of quantitative read-across predictions (based on Euclidean distance-based similarity)

Sl.	Dispersion/Similarity measure	Desired range	Reliability

1.	SD_activity (Euclidean)	<=0.75	Very good (All criteria met);
			Good (Criterion 1 and at least one of the rest but not all);
			Moderate (Any one met);
			Bad (None of the criteria met)

2.	*g* (Euclidean)	<=0.4*	
3(a)	Average similarity (Euclidean)	>=0.85	
3(b)	CV_similarity (Euclidean)	<=0.05	

*Corresponds to *PosFrac* >= 0.8 or *PosFrac* <= 0.2

Overview and Conclusion

In absence of experimental data for toxicity or property of any query chemical, a chemical similarity-based approach is an ideal alternative to bridge the data gaps. Chemical read-across has emerged as a proven method for efficient prediction in this regard which is also recognized and accepted by different regulatory bodies like OECD, US EPA, etc. and regulations like REACH [29]. Although chemical read-across may quickly predict the target property or toxicity of the query chemicals, in absence of the experimental values, it may be challenging to attach a level of uncertainty to the compound-specific predictions. We have discussed this aspect in the context of the Read-Across-v4.0 tool developed by us, but in general the principles should be applicable to other chemical read-across predictions also. From the present analysis, dispersion of the response values of selected close source compounds (specifically standard deviation) emerges to be the most deterministic feature for the reliability of predictions. In the discussed tool, read-across predictions are made using a weighted average approach. Naturally, weighted standard deviation and weighted standard error values are also reported. Based on this, a confidence interval of each predicted value may be presented as below:

$$95\%\ confidence\ interval\ of\ read - across\ predictions\ =\ weighted\ average + t_{95\%} \times \frac{s_{weighted}}{\sqrt{n}} \qquad (2)$$

Apart from the dispersion measures, chemical similarity metrics like concordance measure g, which indicates whether the close source compounds belong to either a definite class (positive or negative, leading to more reliability) or a mixed class (less reliability), average similarity level (higher reliability for higher similarity level) and coefficient of variation of similarity (a greater value leads to lower reliability) have been found to important contributing factors. The difference between the maximum similarity levels of query compounds to positive and negative source compounds is also found important in some cases depending on the data structure. The interpretation of the similarity-based measures depends on the data structures. In case of a high dispersion of similarity of close source compounds to a query compound and/or equal proportion of close positive and negative source compounds for a query compound, the dispersion of observed responses is the main deterministic measure for the reliability of predictions. We have also made a preliminary recommendation about the desired values of different dispersion/similarity measures for good reliability of read-across predictions; however, this may be refined further with the availability of additional results.

Finally, a higher range of predicted response values has been found to be associated with higher uncertainty of predictions in some cases. It appears that a compound is predicted to be less active with more certainty than a compound predicted to be higher active.

The dispersion and similarity features as listed above may be considered to ascertain the level of confidence during quantitative read-across predictions of query compounds without having experimental response values. These measures will definitely enhance usability of chemical read-across quantitative predictions in absence of observed data. The similarity and error-based measures discussed here are also suitable for a novel kind of modeling (quantitative read-across structure-activity relationship or q-RASAR) which is discussed elsewhere [30].

Conflict of interest

Declared none.

Acknowledgements

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. MC and PD thank AICTE, New Delhi and ICMR, New Delhi, respectively, for Senior Research Fellowships.

References

[1] S. Kar, K. Roy, Predictive toxicology using QSAR: A perspective, J. Indian Chem. Soc. 87 (2010) 1455–1515.

[2] S. Kar, K. Roy, Risk assessment for ecotoxicity of pharmaceuticals – an emerging issue, Expert Opin. Drug Saf. 11 (2012) 235–274. doi:10.1517/14740338.2012.644272.

[3] A.B. Raies, V.B. Bajic, In silico toxicology: computational methods for the prediction of chemical toxicity, WIREs Comput. Mol. Sci. 6 (2016) 147–172. doi:10.1002/wcms.1240.

[4] S. Kar, H. Sanderson, K. Roy, E. Benfenati, J. Leszczynski, Ecotoxicological assessment of pharmaceuticals and personal care products using predictive toxicology approaches, Green Chem. 22 (2020) 1458–1516. doi:10.1039/C9GC03265G.

[5] S. Klatte, H.C. Schaefer, M. Hempel, Pharmaceuticals in the environment – A short review on options to minimize the exposure of humans, animals and ecosystems, Sustain. Chem. Pharm. 5 (2017) 61–66. doi:10.1016/j.scp.2016.07.001.

[6] F. Mansour, M. Al-Hindi, W. Saad, D. Salam, Environmental risk analysis and prioritization of pharmaceuticals in a developing world context, Sci. Total Environ. 557–558 (2016) 31–43. doi:10.1016/j.scitotenv.2016.03.023.

[7] A. Cherkasov, E.N. Muratov, D. Fourches, A. Varnek, I.I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y.C. Martin, R. Todeschini, V. Consonni, V.E. Kuz'Min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, A. Tropsha, QSAR modeling: Where have you been? Where are you going to?, J. Med. Chem. 57 (2014) 4977–5010. doi:10.1021/JM4004285

[8] E. Berggren, P. Amcoff, R. Benigni, K. Blackburn, E. Carney, M. Cronin, H. Deluyker, F. Gautier, R.S. Judson, G.E.N. Kass, D. Keller, D. Knight, W. Lilienblum, C. Mahony, I. Rusyn, T. Schultz, M. Schwarz, G. Schüürmann, A. White, J. Burton, A.M. Lostia, S. Munn, A. Worth, Chemical safety assessment using read-across: Assessing the use of novel testing methods to strengthen the evidence base for decision making, Environ. Health Perspect. 123 (2015) 1232–1240. doi:10.1289/ehp.1409342.

[9] S. Kovarich, L. Ceriani, M. Fuart Gatnik, A. Bassan, M. Pavan, Filling data gaps by read-across: A mini review on its application, developments and challenges, Mol. Inform. 38 (2019) 1800121. doi:10.1002/minf.201800121.

[10] A. Gajewicz, K. Jagiello, M.T.D. Cronin, J. Leszczynski, T. Puzyn, Addressing a bottle neck for regulation of nanomaterials: Quantitative read-across (Nano-QRA) algorithm for cases when only limited data is available, Environ. Sci. Nano. 4 (2017) 346–358. doi:10.1039/c6en00399k.

[11] A. Gajewicz, Development of valuable predictive read-across models based on "real-life" (sparse) nanotoxicity data, Environ. Sci. Nano. 4 (2017) 1389–1403. doi:10.1039/c7en00102a.

[12] T.W. Schultz, P. Amcoff, E. Berggren, F. Gautier, M. Klaric, D.J. Knight, C. Mahony, M. Schwarz, A. White, M.T.D. Cronin, A strategy for structuring and reporting a read-across prediction of toxicity, Regul. Toxicol. Pharmacol. 72 (2015) 586–601. doi:10.1016/j.yrtph.2015.05.016.

[13] G. Schüürmann, R.U. Ebert, R. Kühne, Quantitative read-across for predicting the acute

fish toxicity of organic compounds, Environ. Sci. Technol. 45 (2011) 4616–4622. doi:10.1021/ES200361R.

[14] B. van Ravenzwaay, S. Sperber, O. Lemke, E. Fabian, F. Faulhammer, H. Kamp, W. Mellert, V. Strauss, A. Strigun, E. Peter, M. Spitzer, T. Walk, Metabolomics as read-across tool: A case study with phenoxy herbicides, Regul. Toxicol. Pharmacol. 81 (2016) 288–304. doi:10.1016/J.YRTPH.2016.09.013.

[15] R. Kühne, R.U. Ebert, P.C. Vonderohe, N. Ulrich, W. Brack, G. Schüürmann, Read-across prediction of the acute toxicity of organic compounds toward the water flea Daphnia magna, Mol. Inform. 32 (2013) 108–120. doi:10.1002/MINF.201200085.

[16] S.J. Enoch, M.T.D. Cronin, T.W. Schultz, J.C. Madden, Quantitative and mechanistic read across for predicting the skin sensitization potential of alkenes acting via Michael addition, Chem. Res. Toxicol. 21 (2008) 513–520. doi:10.1021/TX700322G.

[17] M. Chatterjee, A. Banerjee, P. De, A. Gajewicz-Skretna, K. Roy, A novel quantitative read-across tool designed purposefully to fill the existing gaps in nanosafety data, Environ. Sci. Nano 9 (2022) 189–203. doi:10.1039/d1en00725d.

[18] P.R. Bevington, D.K. Robinson, Data reduction and error analysis, McGraw-Hill, New York, 1969.

[19] R.K. Mukherjee, V. Kumar, K. Roy, Chemometric modeling of plant protection products (PPPs) for the prediction of acute contact toxicity against honey bees (A. mellifera): A 2D-QSAR approach, J. Hazard. Mater. 423 (2022) 127230. doi:10.1016/j.jhazmat.2021.127230.

[20] R.K. Mukherjee, V. Kumar, K. Roy, Ecotoxicological QSTR and QSTTR modeling for the pediction of acute oral toxicity of pesticides against multiple avian species, Environ. Sci. Technol. 56 (2022) 335–348. doi:10.1021/acs.est.1c05732.

[21] A. Banerjee, P. De, V. Kumar, S. Kar, K. Roy, Quick and efficient quantitative predictions of androgen receptor binding affinity for screening endocrine disruptor chemicals using 2D-QSAR and chemical read-across. ChemRxiv (2022). https://doi.org/10.26434/chemrxiv-2022-gcrjg

[22] K. Roy, S. Kar, R. Das, Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment, Academic Press, New York, 2015.

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

[24] G. Snedecor, W. Cochran, Statistical Methods, 8th ed., Iowa State University Press, Ames, IA, 1989.

[25] SPSS Statistics - India, IBM (2022). https://www.ibm.com/in-en/products/spss-statistics (accessed April 6, 2022).

[26] H. van de Waterbeemd, Discriminant analysis for activity prediction, in: H. van de Waterbeemd (Ed.), Chemom. Methods Mol. Des., VCH, Weinheim, Germany, 1995: pp. 283–293.

[27] A. Rácz, A. Gere, D. Bajusz, K. Héberger, Is soft independent modeling of class analogies a reasonable choice for supervised pattern recognition?, RSC Adv. 8 (2017) 10–21. doi:10.1039/C7RA08901E.

[28] K. Héberger, Sum of ranking differences compares methods or models fairly, Trends Anal. Chem. 29 (2010) 101–109. doi:10.1016/J.TRAC.2009.09.009.

[29] H Foth, A.W. Hayes, Background of REACH in EU regulations on evaluation of chemicals, Hum. Exp. Toxicol. 27 (2008) 443-461. doi:10.1177%2F0960327108092296

[30] A. Banerjee, K. Roy, First report of q-RASAR modeling towards an approach of easy interpretability and efficient transferability, ChemRxiv Cambridge Open Engag. (2022). doi:10.26434/chemrxiv-2022-0qclt.