q-RASAR vs. QSAR: EFFICIENT PREDICTIONS OF ACTIVITY/PROPERTY/TOXICITY ENDPOINTS

Kunal Roy
Drug Theoretics and Cheminformatics Lab
Division of Medicinal and Pharmaceutical Chemistry
Department of Pharmaceutical Technology
Jadavpur University, Kolkata 700 032 (India)
Email: kunalroy_in@yahoo.com
URL: http://sites.google.com/site/kunalroyindia/
Input Data
(Activity/Property/Toxicity)

Descriptor calculation
- PaDEL
- Dragon
- AlvaDesc
- Ochem
- Rdkit
- One hot coding
- CBVM

Dataset Division
- Random Division
- Kennard-Stone based
- Activity based
- Euclidean distance based
- K-medoids

Feature Selection
- Genetic algorithm
- Lasso feature selection
- Stepwise selection

Learning Algorithms
- Supervised
- Unsupervised
- Regression
- Classification
- Clustering
- Dimensionality reduction

Model analysis
- Predictive variance
- Applicability domain
- Mechanistic interpretation
- Model quality

Online databases
Data retrieval from the literature
Handbooks
Scientific articles

Biological and Chemical curation

9/20/2023
Applications of QSAR

- Biomedical science
- Pharmaceutical science
- Materials Science
- Food Sciences
- Agricultural Sciences
- Nano Sciences
- Predictive toxicology
- Regulatory aspects
• Read across (RA) is a **prediction method** of unknown chemicals from the chemical analogues with known toxicity from the **same chemical category**.

• It is accepted by **REACH** and **US EPA**.

• Used for **data gap filling**.

• Defined chemical category is necessary.

  • Strategies: One → One; One → Many
  Many → One; Many → Many

• **Analog** approach

• **Category** approach

---

Chatterjee et al., Environ Sci Nano, 2022, 9, 189-203
Workflow for read-across predictions

*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
Unsupervised learning (Similarity -based)

Supervised learning (Response - dependent)

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

<table>
<thead>
<tr>
<th>Measure</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dispersion measures</strong></td>
<td></td>
</tr>
<tr>
<td>SD_activity</td>
<td>Standard deviation of the (observed) activity values of the selected close source compounds for each query compound</td>
</tr>
<tr>
<td>CV_activity</td>
<td>Coefficient of variation of the response</td>
</tr>
<tr>
<td><strong>Similarity measures</strong></td>
<td></td>
</tr>
<tr>
<td>Average similarity</td>
<td>Mean similarity to the close source compounds for each query compound</td>
</tr>
<tr>
<td>SD_similarity</td>
<td>Standard deviation of the similarity values of the selected close source compounds for each query compound</td>
</tr>
<tr>
<td>MaxPos</td>
<td>Maximum Similarity level to the Positive close source compounds (based on source set observed mean)</td>
</tr>
<tr>
<td>MaxNeg</td>
<td>Maximum Similarity level to the Negative close source set compounds (based on source set observed mean)</td>
</tr>
<tr>
<td>AbsDiff</td>
<td>Absolute difference between MaxPos and MaxNeg</td>
</tr>
<tr>
<td><strong>Concordance measure</strong></td>
<td></td>
</tr>
<tr>
<td>$g$</td>
<td>$g = 1 - 2 \times</td>
</tr>
<tr>
<td>Measure</td>
<td>Expression</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Weighted average activity</td>
<td>$\overline{x_{wtd}} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$</td>
</tr>
<tr>
<td>SD_activity</td>
<td>$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}$</td>
</tr>
<tr>
<td>CV_activity</td>
<td>$CV_{activity} = \frac{s_{weighted}}{\overline{x_{wtd}}}$</td>
</tr>
<tr>
<td>ED-based similarity function</td>
<td>$f(ED) = 1 - d(X,Y)_{scaled}$</td>
</tr>
<tr>
<td></td>
<td>$d(X,Y) = \sqrt{\sum_{i=0}^{n} (X_i - Y_i)^2}$</td>
</tr>
<tr>
<td>Measure</td>
<td>Expression</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-------------------------------------------------</td>
</tr>
<tr>
<td>GK-based similarity function</td>
<td>[ f(GK) = e^{-\frac{</td>
</tr>
<tr>
<td>LK-based similarity function</td>
<td>[ f(LK) = e^{-\gamma |x-y|_1} ]</td>
</tr>
<tr>
<td>Average similarity</td>
<td>[ \text{Similarity}<em>{average} = \frac{\sum</em>{i=1}^{n} f_i}{n} ]</td>
</tr>
<tr>
<td>SD_similarity</td>
<td>[ s_{\text{similarity}} = \sqrt{\frac{\sum_{i=1}^{n} (f_i - \overline{f})^2}{n-1}} ]</td>
</tr>
<tr>
<td>Measure</td>
<td>Expression</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------------------------------</td>
</tr>
<tr>
<td>MaxPos</td>
<td></td>
</tr>
<tr>
<td>MaxNeg</td>
<td></td>
</tr>
<tr>
<td>AbsDiff</td>
<td>$AbsDiff =</td>
</tr>
<tr>
<td>Concordance measure</td>
<td>$g = 1 - 2 \times</td>
</tr>
</tbody>
</table>
\[ g = 1 - 2 \times |\text{PosFrac} - 1/2| \]

\[ g_m = (-1)^n \times 2|\text{PosFrac} - 0.5| \]
RASAR Algorithm linked with AOP

Chemical Datasets: Chemical Identifiers, Endpoint values

Known inhibitors

Target 1 MIE1.1 Target 2 MIE1.2 Target 3 MIE1.3

Unsupervised

Similarity matrix generation

Feature vector (Similarity descriptors based on Fingerprint/property, Toxicity values against multiple endpoints)

Explore maximum similarity

Supervised learning like Random Forest

Query compounds

AOP1

AOP2

AOP3

Endpoint 1

Endpoint 2

Endpoint 3

Model predictions for multiple endpoints
RASAR algorithm combining QSAR and Read-across

- Chemical Dataset with quantitative response values
- Computations 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)
RASAR Descriptor Calculation

Example:
Training Set = Train.xlsx
Test/Query Set = Test.xlsx

Feature Selection from the
Training set
(Structural + Physicochemical)

Training Set
Optimized hyperparameters from the Validation set(s)

Test Set
Training set = Train
Query set = Test

RASAR descriptor calculation (for the Training set)

RASAR-Desc-Calc-v2.0

Clubbing of the RASAR descriptors and the previously selected features

Prediction of test set compounds

Feature Selection (Training set)
(Structural + Physicochemical + RASAR descriptors)

RASAR Model Development
Modeling androgen receptor binding affinity

\[
\text{LogRBA} = -3.23 + 0.49 \times S_{\text{SSS}}\text{CH} - 0.41 \times M_{\text{Max}}\text{CH} + 0.23 \times n\text{C}_{\text{C}}\text{conj} \\
+ 0.35 \times \log P_{99} - 0.17 \times F_{10}[C - O] + 0.06 \times \text{minsOH} + 0.06 \times N\% + 0.67 \times F_{08}[O - F] 
\]

\[R^2_{\text{Train}} = 0.74, R^2_{\text{LOO}} = 0.68, R^2_{\text{FI}} = 0.58, R^2_{\text{F2}} = 0.58 \]

Scaled average \( r^2_{\text{N}} \text{(Train)} = 0.57 \), Scaled average \( r^2_{\text{N}} \text{(Test)} = 0.50 \)

Scaled delta \( r^2_{\text{N}} \text{(Train)} = 0.18 \), Scaled delta \( r^2_{\text{N}} \text{(Test)} = 0.07 \)

\[\text{MAE}_{\text{Train}} = 0.46, \text{MAE}_{\text{Test}} = 0.54, n_{\text{Train}} = 103, n_{\text{Test}} = 44\]
Modeling androgen receptor binding affinity

We have used a data set androgen receptor binding affinity (RBA) originally collected from the Endocrine Disruptor Knowledge Base (EDKB) database (https://www.fda.gov/science-research/bioinformatics-tools/endocrinendisruptor-knowledge-base), and chemical curation of the compounds was performed by the application of a KNIME workflow (https://sites.google.com/site/dtclabdrc/) taking the single.sdf file as input.

\[ \text{LogRBA} = -3.23 + 0.49 \times \text{SsstrCH} - 0.41 \times \text{MaxcCH} + 0.23 \times nCconj + 0.35 \times \text{LogP99} - 0.17 \times \text{F10}[C - \text{O}] + 0.06 \times \text{minsOH} + 0.06 \times \text{N}\% + 0.67 \times \text{F08}[O - F] \]

\[ R^2_{(Train)} = 0.74, \quad Q^2_{(LOO)} = 0.68, \quad Q^2_{(T1)} = 0.58, \quad Q^2_{(T2)} = 0.58 \]

Scaled average \( r^2_{(Train)} = 0.57 \), \( r^2_{(Test)} = 0.50 \)

Scaled delta \( r^2_{(Train)} = 0.18 \), \( r^2_{(Test)} = 0.07 \)

\( MAE_{(Train)} = 0.46, \quad MAE_{(Test)} = 0.54, \quad n_{(Train)} = 103, \quad n_{(Test)} = 44 \)

\( n_{\text{Training}} = 102, \quad n_{\text{Test}} = 44 \)

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.
### Table 2  List of physicochemical features selected from the previously reported QSAR model [12]

<table>
<thead>
<tr>
<th>Measure</th>
<th>Description</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>SsssCH</td>
<td>Sum of E-state value of tertiary carbon atoms of type &gt;CH–</td>
<td>E-state index</td>
</tr>
<tr>
<td>MaxaaCH</td>
<td>Maximum E-state value of the carbon atom of type aaCH</td>
<td>E-state index</td>
</tr>
<tr>
<td>nCconj</td>
<td>Number of non-aromatic conjugated carbons (sp²)</td>
<td>Constitutional descriptor</td>
</tr>
<tr>
<td>LOGP99</td>
<td>Wildmann-Crippen octanol–water partition coefficient</td>
<td>Hydrophobicity measure</td>
</tr>
<tr>
<td>F10[C–O]</td>
<td>Frequency of C and O at the topological distance 10</td>
<td>Atom pair index</td>
</tr>
<tr>
<td>minsOH</td>
<td>Minimum Estate of the –OH hydroxyl group</td>
<td>E-state index</td>
</tr>
<tr>
<td>N%</td>
<td>The percentage of nitrogen present in the molecular structure</td>
<td>Constitutional descriptor</td>
</tr>
<tr>
<td>F08[O–F]</td>
<td>The frequency of O and F atoms at the topological distance of 8</td>
<td>Atom pair index</td>
</tr>
</tbody>
</table>
RASAR: Modeling androgen receptor binding affinity

Table 3  List of q-RASAR models

<table>
<thead>
<tr>
<th>Model no.</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Individual q-RASAR models</strong></td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td>logRBA = −1.33 + 2.27MaxPos(GK) − 3.57Avg.Sim(GK) − 1.02g(GK) + 0.04minsOH − 0.14N% − 0.06F10[C − O]</td>
</tr>
<tr>
<td>M2</td>
<td>logRBA = −2.38 − 1.66MaxNeg(GK) + 0.78MaxPos(GK) + 4.32SDSimilarity(GK) + 0.06minsOH − 0.09N% − 0.05F10[C − O]</td>
</tr>
<tr>
<td>M3</td>
<td>logRBA = −1.97 + 0.35SssSCH + 1.55MaxPos(GK) − 0.34MaxaaCH − 1.31Avg.Sim(GK) + 0.01minsOH − 0.04F10[C − O]</td>
</tr>
<tr>
<td>M4</td>
<td>logRBA = −2.93 − 1.25MaxNeg(GK) + 1.22MaxPos(GK) + 0.73SDActivity(GK) + 0.05nCconj + 2.47SDSimilarity(GK) + 0.03minsOH</td>
</tr>
<tr>
<td><strong>Pooled descriptor q-RASAR models</strong></td>
<td></td>
</tr>
<tr>
<td>P1 (M1 + M2)</td>
<td>logRBA = −1.71 − 1.47MaxNeg(GK) + 1.06MaxPos(GK) + 2.88SDSimilarity(GK) − 0.86Avg.Sim(GK) + 0.05minsOH − 0.41g(GK) − 0.10N% − 0.05F10[C − O]</td>
</tr>
<tr>
<td>P2 (M1+M2+M3)</td>
<td>logRBA = −1.76 − 1.00MaxNeg(GK) + 0.29sSssSCH + 0.91MaxPos(GK) − 0.24MaxaaCH − 0.40Avg.Sim + 1.32SDSimilarity(GK) + 0.03minsOH − 0.04F10[C − O] − 0.05% + 0.17g(GK)</td>
</tr>
<tr>
<td>P3 (M1+M2+M4)</td>
<td>logRBA = −2.55 − 1.13MaxNeg(GK) + 1.10MaxPos(GK) + 0.72SDActivity(GK) + 0.08nCconj − 0.48Avg.Sim(GK) + 1.81SDSimilarity(GK) + 0.03minsOH − 0.05F10[C − O] − 0.06% + 0.13g(GK)</td>
</tr>
</tbody>
</table>
RASAR: Modeling androgen receptor binding affinity

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

Arkaprava Banerjee - Kumal Roy

Received: 31 April 2022 / Accepted: 3 June 2022
© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
RASAR: Modeling androgen receptor binding affinity
RASAR: Modeling androgen receptor binding affinity
RASAR: Modeling androgen receptor binding affinity

\[
\log \text{RBA} = -1.21 - 1.31 \text{MaxNeg(GK)} + 0.58 g_m(GK) + 0.21 \text{MaxPos(GK)} + 2.23 \text{SD Similarity (GK)} - 0.67 \text{Avg.Sim(GK)} + 0.06 \text{min sOH} - 0.10 \text{N\%} - 0.13 F10[C - O]
\]  

\( n_{\text{Training}} = 102 \quad n_{\text{Test}} = 44 \quad \text{LV} = 4 \)

\[
R^2 = 0.753 \quad Q^2_{(\text{LOO})} = 0.698 \quad Q^2_{F1} = 0.674 \quad Q^2_{F2} = 0.674 \quad \text{MAE}_{\text{TEST}} = 0.461
\]

\[
g_m = (-1)^n \times 2 \left| \text{PosFrac} - 0.5 \right|
\]

- \( n = 1 \) if \( \text{MaxPos} < \text{MaxNeg} \),
- \( n = 2 \) if \( \text{MaxPos} > \text{MaxNeg} \)

![VIP Plot Model F1](image1)

![VIP Plot Model F1m](image2)
RASAR: Modeling five toxicity endpoints

On Some Novel Similarity-Based Functions Used in the ML-Based q-RASAR Approach for Efficient Quantitative Predictions of Selected Toxicity End Points

Arkaprava Banerjee and Kunal Roy*

Cite This: https://doi.org/10.1021/acs.chemrestox.2c00374

Read Online
RASAR: Modeling five toxicity endpoints

- Rat androgen receptor binding affinity (147 compounds)
- Acute oral toxicity data (bobwhite quail) (128 pesticides)
- Acute contact toxicity (honey bee) (113 plant protection products)
- Acute oral toxicity data (mallard duck) (62 pesticides)
- Inhalation toxicity data (126 VOCs)
RASAR: Modeling five toxicity endpoints
RASAR: Modeling five toxicity endpoints
### RASAR: Modeling five toxicity endpoints

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Model</th>
<th>$R^2$</th>
<th>$Q^2_{(LOO)}$</th>
<th>$Q^2_{P1}$</th>
<th>$Q^2_{P2}$</th>
<th>$\text{MAE}_{(TEST)}$ *</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dataset 1</strong></td>
<td>q-RASAR (GK)</td>
<td>0.71</td>
<td>0.63</td>
<td>0.70</td>
<td>0.70</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>QSAR$^{13}$</td>
<td>0.74</td>
<td>0.68</td>
<td>0.58</td>
<td>0.58</td>
<td>0.54</td>
</tr>
<tr>
<td><strong>Dataset 2</strong></td>
<td>q-RASAR (GK)</td>
<td>0.68</td>
<td>0.54</td>
<td>0.77</td>
<td>0.77</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>QSAR$^{18}$</td>
<td>0.66</td>
<td>0.58</td>
<td>0.65</td>
<td>0.65</td>
<td>0.46</td>
</tr>
<tr>
<td><strong>Dataset 3</strong></td>
<td>q-RASAR (LK)</td>
<td>0.62</td>
<td>0.53</td>
<td>0.83</td>
<td>0.83</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>QSAR$^{19}$</td>
<td>0.67</td>
<td>0.59</td>
<td>0.65</td>
<td>0.65</td>
<td>0.58</td>
</tr>
<tr>
<td><strong>Dataset 4</strong></td>
<td>q-RASAR (GK)</td>
<td>0.68</td>
<td>0.53</td>
<td>0.68</td>
<td>0.60</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>QSAR$^{18}$</td>
<td>0.66</td>
<td>0.57</td>
<td>0.66</td>
<td>0.58</td>
<td>0.58</td>
</tr>
<tr>
<td><strong>Dataset 5</strong></td>
<td>q-RASAR (GK)</td>
<td>0.73</td>
<td>0.64</td>
<td>0.74</td>
<td>0.74</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>QSAR$^{20}$</td>
<td>0.74</td>
<td>0.66</td>
<td>0.68</td>
<td>0.68</td>
<td>0.49</td>
</tr>
</tbody>
</table>
RASAR: Modeling five toxicity endpoints
RASAR: Modeling five toxicity endpoints
RASAR: Modeling five toxicity endpoints
RASAR: Modeling five toxicity endpoints

Data set 4
Train

Test

Data set 5
Train

Test
RASAR: Modeling five toxicity endpoints
RASAR: Modeling Cardiotoxicity data

Collection of the cardiotoxicity data
Data curation, descriptor calculation and data pre-treatment
Dataset division

Training  Test

Feature selection  Default settings of the hyperparameters
Calculation of the RASAR descriptors  Data fusion  Feature selection & model development

261 Compounds

Predictions for the test set

Compounds from the DrugBank DB

Predictions

Machine Learning algorithms

Partial Least Squares model
Detection of less confident predictions using the DTC Plot

hERG_Toxicity_Calculator_v2.0

RR  LSVR  SVR  RF
GB  AdaBoost  MLP  k-NN

DTC LAB

9/20/2023
Banerjee and Roy, Unpublished, 2023
**RASAR: Modeling Cardiotoxicity data**

**DTC Plot for the RASAR model Applicability Domain**

- A high difference in the similarity level naturally leads to a high difference in response values.
- A considerable extent of similarity along with a large difference in response values makes the impact of RASAR descriptors unreliable for these compounds.
- A low level of maximum similarity leads to a lower impact of RASAR descriptors.
RASAR: Modeling Cardiotoxicity data

Radar Plot of the std. regression coefficients of the PLS q-RASAR model

\[ n_{Train} = 196 \quad n_{Test} = 63 \quad R^2_{Train} = 0.608 \quad Q^2_{(LOO)} = 0.546 \]
\[ Q^2_{F1} = 0.660 \quad Q^2_{F2} = 0.660 \quad MAE_{Train} = 0.581 \quad MAE_{Test} = 0.548 \]
RASAR: Modeling Cardiotoxicity data

Variable Importance Plot
VIP[Comp. 4]

Var ID (Primary)
RASAR: Modeling Cardiotoxicity data

Scatter Plot of the PLS q-RASAR model
# RASAR: Modeling Cardiotoxicity data

<table>
<thead>
<tr>
<th>Model</th>
<th>Training set statistics</th>
<th>Optimum hyperparameters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2_{\text{Train}}$, MAE, MAE (CV)</td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>$\text{MAE}<em>{\text{Train}}$, $\text{MAE}</em>{(C V)}$</td>
<td></td>
</tr>
<tr>
<td>PLS</td>
<td>0.608, 0.452, 0.502 = 0.006</td>
<td>$n_\text{components} = 4$</td>
</tr>
<tr>
<td>RR</td>
<td>0.608, 0.454, 0.504 = 0.006</td>
<td>$\alpha = 0.5$</td>
</tr>
<tr>
<td>LSVR</td>
<td>0.593, 0.433, 0.504 = 0.006</td>
<td>$C=15.0$, $\text{max}_\text{iter} = 1000000$</td>
</tr>
<tr>
<td>SVR</td>
<td>0.676, 0.479, 0.513 = 0.007</td>
<td>$\text{degree}=2$, $\text{gamma}='\text{auto}'$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test set statistics</th>
<th>$R^2_{\text{Test}}$, MAE, Q_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLS</td>
<td>0.508 = 0.012, 0.518 = 0.003, 0.66</td>
</tr>
<tr>
<td>RR</td>
<td>0.509 = 0.002, 0.515 = 0.003, 0.66</td>
</tr>
<tr>
<td>LSVR</td>
<td>0.515 = 0.002, 0.502 = 0.004, 0.64</td>
</tr>
<tr>
<td>SVR</td>
<td>0.524 = 0.002, 0.486 = 0.003, 0.63</td>
</tr>
</tbody>
</table>

---

9/20/2023

Banerjee and Roy, Unpublished, 2023
### RASAR: Modeling Cardiotoxicity data

<table>
<thead>
<tr>
<th>Method</th>
<th>m &amp; (std)</th>
<th>m &amp; (std)</th>
<th>m &amp; (std)</th>
<th>m &amp; (std)</th>
<th>m &amp; (std)</th>
<th>m &amp; (std)</th>
<th>m &amp; (std)</th>
<th>m &amp; (std)</th>
<th>m &amp; (std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>0.733</td>
<td>0.398</td>
<td>0.548</td>
<td>0.551</td>
<td>0.550 ± 0.007</td>
<td>0.427 ± 0.013</td>
<td>0.554 ± 0.002</td>
<td>0.433 ± 0.003</td>
<td>0.58</td>
</tr>
<tr>
<td>Gradboost</td>
<td>0.803</td>
<td>0.337</td>
<td>0.536</td>
<td>0.521</td>
<td>0.551 ± 0.007/0.014</td>
<td>0.422 ± 0.018</td>
<td>0.559 ± 0.018</td>
<td>0.418 ± 0.009</td>
<td>0.65</td>
</tr>
<tr>
<td>Adabooost</td>
<td>0.685</td>
<td>0.45</td>
<td>0.568</td>
<td>0.558</td>
<td>0.557 ± 0.007</td>
<td>0.425 ± 0.013</td>
<td>0.565 ± 0.002</td>
<td>0.424 ± 0.003</td>
<td>0.58</td>
</tr>
<tr>
<td>MLP regression</td>
<td>0.608</td>
<td>0.465</td>
<td>0.499</td>
<td>0.45</td>
<td>0.501 ± 0.006</td>
<td>0.524 ± 0.012</td>
<td>0.505 ± 0.002</td>
<td>0.524 ± 0.003</td>
<td>0.65</td>
</tr>
<tr>
<td>KNN regression</td>
<td>0.572</td>
<td>0.489</td>
<td>0.562</td>
<td>0.565</td>
<td>0.573 ± 0.007</td>
<td>0.396 ± 0.016</td>
<td>0.583 ± 0.002</td>
<td>0.398 ± 0.004</td>
<td>0.52</td>
</tr>
</tbody>
</table>

---

**Note:** The table above shows the performance metrics of different machine learning models used for modeling cardiotoxicity data. Each model's performance is evaluated based on accuracy metrics such as mean (m) and standard deviation (std). The models include Random Forest (RF), Gradient Boosting (Gradboost), Adaboost, Multi-Layer Perceptron (MLP) regression, and K-Nearest Neighbors (KNN) regression. The parameters used for each model are also provided, including max_depth, n_estimators, and random_state for RF and Gradboost, learning_rate, loss, and n_estimators for Adaboost, activation, alpha, hidden_layer_sizes, learning_rate_init, max_iter, solver, and random_state for MLP regression, and leaf_size and n_neighbors for KNN regression. The models are trained and evaluated on a dataset with specific features and labels relevant to cardiotoxicity assessment.
RASAR: Modeling Cardiotoxicity data

hERG_Toxicity Calculator v 2.0

This tool quickly provides the quantitative prediction (along with AD) of the potential cardiotoxicity induced by a compound by interacting with the hERG K⁺ channel using a q-RASAR model developed by the DTC Laboratory (Banerjee and Roy, 2023, Unpublished).

Predictions to be used for research purposes only.
DTC Lab Tools

Quantitative Read-Across
V4.1

Chatterjee M., Banerjee A., De P., Gajewicz A., Roy K.
Banerjee A., Roy K., Mol Divers, 2022, DOI: 10.1007/s11030-022-10478
Software developed by Aditya Saha
Email: aditya18narat@gmail.com

RASAR
Descriptor Calculator
v2.0
Banerjee A., Roy K., Mol Divers, 2022, DOI: 10.1007/s11030-022-10478-6
Banerjee A., Chatterjee M., De P., Roy K., Chemom Intell Lab Sys, 227, 2022,
DOI: 10.1016/j.chemolab.2022.104613
Software developed by Aditya Saha
Email: aditya18narat@gmail.com

Auto RA Optimizer

Auto RA Optimizer
v1.0

Machine Learning Regression
Beta version

Chatterjee M., Banerjee A., De P., Gajewicz A., Roy K.
Software developed by Aditya Saha
Email: aditya18narat@gmail.com

These GUIs use scikit-learn libraries to optimize hyperparameters and develop machine learning regression models
Software developed by Souvik Purkait
Email: souvikpurkait22@gmail.com

https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home
RASAR: Modeling Property data

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

Arkaprava Banerjee¹ | Agnieszka Gajewicz-Skretna² | Kunal Roy³

¹Drug Theorics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, India
²Laboratory of Environmental Cheminformatics, Faculty of Chemistry, University of Gdansk, Gdansk, Poland

Correspondence
Kunal Roy, Drug Theorics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India.
Email: kunal.roy@jadavpuruniversity.in

Funding information
Science and Engineering Research Board (SERB), Grant/Award Number: MTH/2019/000000/ Life Science Research Board, DBDIO, India, Grant/Award Number: LSRB/01/15001/M/LSRB-394/ SH&D5/2022

Abstract
In this study, the specific surface area of various perovskites was modeled using a novel quantitative read-across structure-property relationship (q-RASPR) approach, which clubs both Read-Across (RA) and quantitative structure-property relationship (QSPR) together. After optimization of the hyper-parameters, certain similarity-based error measures for each query compound were obtained. Clubbing some of these error-based measures with the previously selected features along with the Read-Across prediction function, a number of machine learning models were developed using Partial Least Squares (PLS), Ridge Regression (RR), Linear Support Vector Regression (LSVR), Random Forest (RF) regression, Gradient Boost (GBoost), Adaptive Boosting (AdaBoost), Multiple Layer Perceptron (MLP) regression and k-Nearest Neighbor (kNN) regression. Based on the repeated cross-validation as well as external prediction quality and interpretability, the PLS model ($R^2_{tr} = 38$, $R^2_{cv} = 12$, $R^2_{pred} = 0.737$, $Q^2_{LOO} = 0.637$, $R^2_{tr} = 0.879$, $Q^2_{(kitored)} = 0.901$) was selected as the best predictor which underscored the previously reported results. The finally selected model should efficiently predict specific surface areas of other perovskites for their use in photocatalysis. The new q-RASPR method also appears promising for the prediction of several other property endpoints of interest in materials science.

Keywords
machine learning, perovskites, photocatalysis, q-RASPR, specific surface area
Machine learning-based q-RASPR modeling of power conversion efficiency of organic dyes in dye-sensitized solar cells†

Souvik Pore, Arkaprava Banerjee and Kunal Roy

Different computational tools are now popularly used as an alternative to experiments for predicting several property endpoints of industrial importance. Recently, read-across and quantitative structure-property relationship (QSPR) have been merged to develop a new modeling technique read-across structure–property relationship (RASPR) which appears to have much potential in predictive modeling. This approach is also promising for modeling relatively smaller data sets as the similarity-based RASPR descriptors are computed from multiple structural and physicochemical features. To understand the potential of RASPR in data gap filling, we have undertaken a case study of modeling Power Conversion Efficiency (PCE) of different classes of organic dyes used in Dye-Sensitized Solar Cells (DSSCs) for renewable energy generation. We have used a large dataset of 429 compounds covering 4 classes of organic dyes. We initially performed read-across analysis using different similarity measures with structural analogues for query compounds and calculated the weighted average predictions. Based on the read-across optimized settings, RASPR descriptors were calculated, and these were then merged with the chemical descriptors, and finally, a single partial least squares (PLS) model was developed for each of the dye classes after feature selection, followed by additional Machine Learning (ML) models. The external prediction quality of the final RASPR models superseded those of the previously developed QSPR models using the same level of chemical information. The important structural features and similarity measures contributing to the PCE have been extracted using the RASPR method which can be used to enhance the PCE values in the newly designed dyes. The RASPR method may also be efficiently applied in modeling other properties of interest in a similar manner.
Prediction-Inspired Intelligent Training for the Development of Classification Read-across Structure–Activity Relationship (c-RASAR) Models for Organic Skin Sensitizers: Assessment of Classification Error Rate from Novel Similarity Coefficients

Arkaprava Banerjee and Kunal Roy*

Abstract: The advancements in the field of chemoinformatics have led to a reduction in animal testing to estimate the activity, property, and toxicity of query chemicals. Read-across structure–activity relationship (RASAR) is an emerging concept that utilizes various similarity functions derived from chemical information to develop highly predictive models. Unlike quantitative structure–activity relationship (QSAR) models, RASAR descriptors of a query compound are computed from its close congeners instead of the compound itself, thus targeting predictions in the model training phase. The objective of the present study is not to propose new QSAR models for skin sensitization but to demonstrate the enhancement in the quality of predictions of the skin-sensitizing potential of organic compounds by developing classification-based RASAR (c-RASAR) models. A diverse, previously curated data set was collected from the literature for which 2D descriptors were computed. The extracted essential features were then used to develop a classification-based linear discriminant analysis (LDA) QSAR model. Furthermore, from the read-across-based predictions, RASAR descriptors were calculated using the basic settings of the hyperparameters for the LightGBM-based optimum similarity measure. After feature selection, an LDA c-RASAR model was developed, which superseded the prediction quality of the LDA–QSAR model. Various other combinations of RASAR descriptors were also taken to develop additional c-RASAR models, all showing better prediction quality than the LDA QSAR model. Various machine learning c-RASAR models were also developed for comparison purposes. In this work, we have proposed and analyzed three new similarity metrics: $d_{SA}$, $d_{LO}$, and $d_{AN}$. The first one is an indicator variable used to generate a simple univariate c-RASAR model with good prediction ability, while the remaining two are similarity indices used to analyze plausible activity cliffs in the training and test sets and are believed to play an important role in the modelability analysis of data sets.
Read-across-based intelligent learning: development of a global q-RASAR model for the efficient quantitative predictions of skin sensitization potential of diverse organic chemicals†

Arkaprava Banerjee and Kunal Roy

Environmental chemicals and contaminants cause a wide array of harmful implications to terrestrial and aquatic life which ranges from skin sensitization to acute oral toxicity. The current study aims to assess the quantitative skin sensitization potential of a large set of industrial and environmental chemicals acting through different mechanisms using the novel quantitative Read-Across Structure–Activity Relationship (q-RASAR) approach. Based on the identified important set of structural and physicochemical features, Read-Across-based hyperparameters were optimized using the training set compounds followed by the calculation of similarity and error-based RASAR descriptors. Data fusion, further feature selection, and removal of prediction confidence outliers were performed to generate a partial least squares (PLS) q-RASAR model, followed by the application of various Machine Learning (ML) tools to check the quality of predictions. The PLS model was found to be the best among different models. A simple user-friendly Java-based software tool was developed based on the PLS model, which efficiently predicts the toxicity value(s) of query compound(s) along with their status of Applicability Domain (AD) in terms of leverage values. This model has been developed using structurally diverse compounds and is expected to predict efficiently and quantitatively the skin sensitization potential of environmental chemicals to estimate their
Research Paper

Machine learning - based q-RASAR modeling to predict acute contact toxicity of binary organic pesticide mixtures in honey bees

Mainak Chatterjee\textsuperscript{a}, Arkaprava Banerjee\textsuperscript{a}, Simone Tosi\textsuperscript{b}, Edoardo Carnesecchi\textsuperscript{c}, Emilio Benfenati\textsuperscript{d}, Kunal Roy\textsuperscript{a,*}

\textsuperscript{a} Drug Theoretical and Chemoinformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India
\textsuperscript{b} Department of Agriscience, Forestry and Food Sciences, University of Turin, Turin, Italy
\textsuperscript{c} Institute for Risk Assessment Sciences, Utrecht, the Netherlands
\textsuperscript{d} Department of Environmental Health Sciences, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, via Mario Negri 2, 20157 Milano, Italy

\textbf{HIGHLIGHTS}

- A novel q-RASAR model has been developed for the prediction of toxicity of organic mixtures in honey bees.
- Three different mixing rules have been used to calculate the mixture descriptors.
- The developed model has been validated following the recent OECD guidelines.
- The use of machine learning-based algorithms further enhanced the predictability of the q-RASAR model.
- The toxicity of environmentally relevant untreated organic mixtures has been predicted by this new model.

\textbf{GRAPHICAL ABSTRACT}
**RASAR: Modeling Nanotoxicity data**

**Efficient predictions of cytotoxicity of TiO₂-based multi-component nanoparticles using a machine learning-based q-RASAR approach**

Arkaprava Banerjee, Supratik Kar, Souvik Pore, and Kunal Roy

Drug Theorics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, India; Chemometrics & Molecular Modeling Laboratory, Department of Chemistry, Kean University, Union, NJ, USA

**ABSTRACT**

The availability of experimental nanotoxicity data is in general limited which warrants both the use of *in silico* methods for data gap filling and exploring novel methods for effective modeling. Read-Across Structure-Activity Relationship (RASAR) is an emerging cheminformatic approach that combines the usefulness of a QSAR model and similarity-based Read-Across predictions. In this work, we have generated simple, interpretable, and transferable quantitative-RASAR (q-RASAR) models which can efficiently predict the cytotoxicity of TiO₂-based multi-component nanoparticles. A data set of 29 TiO₂-based nanoparticles with specific amounts of noble metal precursors was rationally divided into training and test sets, and the Read-Across-based predictions for the test set were generated. The optimized hyperparameters and the similarity approach, which yield the best predictions, were used to calculate the similarity and error-based RASAR descriptors. A data fusion of the RASAR descriptors with the chemical descriptors was done followed by the best subset feature selection. The final set of selected descriptors was used to develop the q-RASAR models, which were validated using the stringent OECD criteria. Finally, a random forest model was also developed with the selected descriptors, which could efficiently predict the cytotoxicity of TiO₂-based multi-component nanoparticles superseding previously reported models in the prediction quality thus showing the merits of the q-RASAR approach. To further evaluate the usefulness of the approach, we have applied the q-RASAR approach also to a second cytotoxicity data set of 34 heterogeneous TiO₂-based nanoparticles which further confirmed the enhancement of external prediction quality of QSAR models after incorporation of RASAR descriptors.
Conclusion

• These studies report the development of simple, interpretable, and reproducible q-RASAR models for various toxicity (activity/property) endpoints.
• The q-RASAR models reported here thus deliver lower prediction errors for the query sets than corresponding QSAR models, suggesting that they are the potential models of choice for efficient predictions using a given level of chemical information.
• Based on the variable importance analysis, the RASAR descriptors “RA score”, “gm” and “average similarity” appear efficient similarity-based determinants for the prediction of toxicity which warrants further extensive studies on these functions.
More about q-RASAR

https://sites.google.com/site/kunalroyindia/home/راسار

https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home#h.i79rttmog6nl

Tools developed by Arkaprava Banerjee

Publications with q-RASAR modeling from other laboratories

- Ecotoxicological QSAR study of fused/non-fused polycyclic aromatic hydrocarbons (FNFPAs): Assessment and priority ranking of the acute toxicity to Pimephales promelas by QSAR and consensus modeling methods. Science of The Total Environment, 876, 162736 (2023)
- Data driven toxicity assessment of organic chemicals against Gammarus species using QSAR approach. Chemosphere 328, 138433 (2023)
- QSAR and Chemical Read-Across Analysis of 370 Potential MGMT Inactivators to Identify the Structural Features Influencing Inactivation Potency. Pharmaceutics 15, 2170 (2023)
Thank you