**Comparative Machine Learning and Deep Learning Frameworks for Robust Carcinogenicity Prediction and Activity Cliffs Analysis**

**Arkaprava Banerjee, Vinay Kumar, Kunal Roy\***

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

**Supplementary Information SI-2**

***\**Correspondence to: Kunal Roy (**[kunal.roy@jadavpuruniversity.in](mailto:kunal.roy@jadavpuruniversity.in); [kunalroy\_in@yahoo.com](mailto:kunalroy_in@yahoo.com)**)**

**Table S1.** List of descriptors used to develop the QSAR and c-RASAR models.

|  |  |  |
| --- | --- | --- |
| **Descriptors** | **Class** | **Description** |
| DBI | Topological indices | Dragon branching index |
| S3K | Topological indices | 3-path Kier alpha-modified shape index |
| MAXDP | Topological indices | Maximal electrotopological positive variation |
| piPC08 | Walk and path counts | Molecular multiple path count of order 8 |
| Yindex | Information indices | Balaban Y index |
| IC3 | Information indices | Information Content index (neighborhood symmetry of 3-order) |
| J\_D | 2D matrix-based descriptors | Balaban-like index from topological distance matrix (Balaban distance connectivity index) |
| SpMax\_L | 2D matrix-based descriptors | Leading eigen value from Laplace matrix |
| J\_D/Dt | 2D matrix-based descriptors | Balaban-like index from distance/detour matrix |
| Chi\_Dz(Z) | 2D matrix-based descriptors | Randic-like index from Barysz matrix weighted by atomic number |
| EE\_B(m) | 2D matrix-based descriptors | Estrada-like index (log function) from Burden matrix weighted by mass |
| SpMax\_B(v) | 2D matrix-based descriptors | Leading eigen value from Burden matrix weighted by van der Waals volume |
| SpMax\_B(s) | 2D matrix-based descriptors | Leading eigen value from Burden matrix weighted by I-State |
| SpMAD\_B(s) | 2D matrix-based descriptors | Spectral mean absolute deviation from Burden matrix weighted by I-State |
| ATSC0m | 2D autocorrelations | Centred Broto-Moreau autocorrelation of lag 0 weighted by mass |
| GATS6e | 2D autocorrelations | Geary autocorrelation of lag 6 weighted by Sanderson electronegativity |
| P\_VSA\_LogP\_4 | P\_VSA-like descriptors | P\_VSA-like on LogP, bin 4 |
| P\_VSA\_LogP\_6 | P\_VSA-like descriptors | P\_VSA-like on LogP, bin 6 |
| P\_VSA\_MR\_5 | P\_VSA-like descriptors | P\_VSA-like on Molar Refractivity, bin 5 |
| P\_VSA\_v\_3 | P\_VSA-like descriptors | P\_VSA-like on van der Waals volume, bin 3 |
| P\_VSA\_e\_2 | P\_VSA-like descriptors | P\_VSA-like on Sanderson electronegativity, bin 2 |
| P\_VSA\_s\_6 | P\_VSA-like descriptors | P\_VSA-like on I-state, bin 6 |
| SpMax\_AEA(dm) | Edge adjacency indices | Leading eigen value from augmented edge adjacency mat. weighted by dipole moment |
| SdssC | Atom-type E-state indices | Sum of dssC E-states |
| CATS2D\_01\_AA | Pharmacophore descriptors | CATS2D Acceptor-Acceptor at lag 01 |
| B01[C-N] | 2D Atom Pairs | Presence/absence of C - N at topological distance 1 |
| B01[N-N] | 2D Atom Pairs | Presence/absence of N - N at topological distance 1 |
| B03[C-Cl] | 2D Atom Pairs | Presence/absence of C - Cl at topological distance 3 |
| B04[C-N] | 2D Atom Pairs | Presence/absence of C - N at topological distance 4 |
| MLOGP | Molecular properties | Moriguchi octanol-water partition coeff. (logP) |
| MDEN-23 | MDE descriptors | Molecular distance edge between all secondary and tertiary nitrogens |
| MACCSFP102 | MACCS fingerprint | QO (Q - heteroatom, O - oxygen) |
| MACCSFP163 | MACCS fingerprint | Six-membered ring |
| RA function(GK) | RASAR Descriptors | A composite function derived from Read-Across that efficiently encapsulates the information of all the selected structural and physicochemical features. |
| CVsim(GK) | RASAR Descriptors | It represents the coefficient of variation of the similarity values for compounds constituting the close source neighbors. |
| MaxPos(GK) | RASAR Descriptors | It represents the maximum similarity value to a positive close source neighbor (a “positive” indicates a data point having an experimental response value greater than the training set mean response). |
| MaxNeg(GK) | RASAR Descriptors | It represents the maximum similarity value to a negative close source neighbor (a “negative” indicates a data point having an experimental response value lower than the training set mean response). |
| Abs MaxPos-MaxNeg | RASAR Descriptors | It represents the absolute difference in the MaxPos and MaxNeg values. |
| Avg.Sim(GK) | RASAR Descriptors | It represents the average similarity values of the close congeners for a particular query compound. |
| SD similarity(GK) | RASAR Descriptors | It represents the standard deviation in the similarity values among the close congeners for a particular query compound. |
| gm(GK)[Banerjee-Roy Coefficient] | RASAR Descriptors | A novel concordance measure representing the propensity of a query compound to be positive or negative. |
| gm\*Avg.Sim | RASAR Descriptors | It is a derived RASAR descriptor, which is a product of gm and Avg. Sim. |
| gm\*SD Similarity | RASAR Descriptors | It is a derived RASAR descriptor, which is a product of gm and SD\_Similarity. |
| Pos.Avg.Sim | RASAR Descriptors | It represents the average similarity value of the positive compounds present in the list of close source neighbors. |
| Neg.Avg.Sim | RASAR Descriptors | It represents the average similarity value of the negative compounds present in the list of close source neighbors. |
| sm1(GK)[Banerjee-Roy similarity coefficient 1] | RASAR Descriptors | A similarity coefficient that can detect activity cliffs present in the dataset. |
| sm2(GK)[Banerjee-Roy similarity coefficient 2] | RASAR Descriptors | A similarity coefficient that can detect activity cliffs present in the dataset. |
| gm\_class(GK) | RASAR Descriptors | A binary RASAR descriptor derived from gm. |