ARKA: A framework of dimensionality reduction for machine-learning classification modeling, risk assessment, and data gap-filling of sparse environmental toxicity data

Arkaprava Banerjee, Kunal Roy*

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

Supplementary Materials SI-3

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

Table S1. List of QSAR descriptors involved in the computation of the ARKA descriptors

Datasets	*ARKA_1*	*ARKA_2*
Dataset 1	gmin, minsssCH	B06[C-N], B04[N-O], B10[C-O], B09[C-C], B05[N-O], MaxaaN, Eta_B, nR05, gmax, C-040, SAscore, B03[O-O]
Dataset 2	B03[O-O], ETA_Psi_1, B02[O-S], X3A	B09[C-C], MLOGP2, ETA_Beta, MLOGP
Dataset 3	ETA_EtaP_B_RC, nCrs, S_tsc	nAB, nRCONHR, Jurs-DPSA-1
Dataset 4	MW, SaaaC, 2χv, Atype_C_24	
Dataset 5	MAXDP	WAP, nRNNOx, Cl-086

Table S2. LDA Coefficient values of QSAR and ARKA models (Datasets 1-5)

Dataset	*Descriptor types*	*Models*
1	**QSAR**	$df =- 0.214 \times nR05 - 0.047 \times$ … $- 0.271 \times MaxaaN + 0.246 \times$ … $[N - O] + 0.307 \times B05[N - O$ … $- 0.134 \times B10[C - O] - 0.12$ …
	ARKA	$df = 0.322 \times ARKA_1 - 0.530 \times ARKA_2$
2	**QSAR**	$df = 1.941 \times X3A + 1.391 \times ET$ … $[O - O] - 1.853 \times B09[C - C]$ …
	ARKA	$df = 1.724 \times ARKA_1 - 1.097 \times ARKA_2$
3	**QSAR**	$df =- 2.380 \times ETA_{EtaP_{B_{RC}}} - 2.0$ … $S_{tsC} - 3.456 \times Jurs - DPSA -$ …
	ARKA	$df = 1.101 \times ARKA_1 - 2.480 \times ARKA_2$

4	**QSAR**	$df = 5.339 \times MW - 2.539 \times SaaaC - 2.704 \times 2_{\chi V} + 2.989 \times Aty$
	ARKA	$df = 1.844 \times ARKA_1$
5	**QSAR**	$df =- 0.664 \times Wap + 0.895 \times MAXDP - 1.472 \times nRNNO_X - 0.7$
	ARKA	$df = 0.895 \times ARKA_1 - 1.553 \times ARKA_2$

Table S3. Optimized hyperparameter* settings for the different ML models developed

Models	***Dataset 1***	***Dataset 2***	***Dataset 3***	***Dataset 4***	***Dataset 5***
SVM_QSAR	C: 10, gamma: scale, kernel: poly	C: 1, gamma: scale, kernel: linear	C: 0.1, gamma: scale, kernel: linear	C: 1, gamma: scale, kernel: rbf	C: 10, gamma: scale, kernel: poly
SVM_ARKA	C: 0.1, gamma: scale, kernel: rbf	C: 1, gamma: scale, kernel: rbf	C: 0.1, gamma: scale, kernel:	C: 1, gamma: scale, kernel:	C: 10, gamma: scale, kernel:

			linear	linear	linear
RF_QSAR	'criterion': 'entropy', 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 500	'criterion': 'gini', 'max_depth': 3, 'min_samples_leaf': 5, 'min_samples_split': 2, 'n_estimators': 500	'criterion': 'entropy', 'max_depth': 3, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 500	'criterion': 'gini', 'max_depth': None, 'min_samples_leaf': 5, 'min_samples_split': 2, 'n_estimators': 50	'criterion': 'gini', 'max_depth': 2, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 500
RF_ARKA	'criterion': 'entropy', 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 3, 'n_estimators': 100	'criterion': 'gini', 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 3, 'n_estimators': 100	'criterion': 'entropy', 'max_depth': 2, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50	'criterion': 'gini', 'max_depth': 1, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 500	'criterion': 'entropy', 'max_depth': 2, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100
LR_QSAR	'C': 0.1, 'penalty': 'l1', 'solver': 'liblinear'	'C': 0.1, 'penalty': None, 'solver': 'lbfgs'	'C': 1.0, 'penalty': 'l2', 'solver': 'lbfgs'	'C': 0.1, 'penalty': None, 'solver': 'saga'	'C': 10.0, 'penalty': 'l1', 'solver': 'liblinear'
LR_ARKA	'C': 0.1, 'penalty': 'l2', 'solver': 'lbfgs'	'C': 0.1, 'penalty': None, 'solver': 'lbfgs'	'C': 0.1, 'penalty': 'l2', 'solver': 'liblinear'	'C': 0.1, 'penalty': 'l2', 'solver': 'liblinear'	'C': 0.1, 'penalty': 'l2', 'solver': 'lbfgs'

*As per Scikit learn (J. Mach. Learn. Res. 12, 2825-2830).

Table S4. Results of 20-times 5-fold cross-validation of all the developed ML models.

Desc	***Models***	***Dataset1***		***Dataset2***		***Dataset3***		***Dataset4***		***Dataset5***	
		Acc.	***Acc.CV***	***Acc.***	***Acc.CV***	***Acc.***	***Acc.CV***	***Acc.***	***Acc.CV***	***Acc.***	***Acc.CV***
QSAR	***LDA***	0.662	0.621	0.863	0.821	0.887	0.876	0.905	0.896	0.741	0.702
	SVM	0.874	0.755	0.863	0.833	0.887	0.868	0.878	0.857	0.759	0.713
	RF	1.000	0.849	0.880	0.787	0.897	0.811	0.905	0.835	0.833	0.672
	LR	0.684	0.635	0.872	0.821	0.897	0.869	0.892	0.877	0.778	0.688
ARKA	***LDA***	0.651	0.645	0.812	0.798	0.825	0.809	0.851	0.838	0.722	0.702
	SVM	0.665	0.662	0.821	0.785	0.825	0.822	0.851	0.851	0.741	0.724
	RF	0.996	0.822	1.000	0.811	0.835	0.755	0.892	0.855	0.833	0.724
	LR	0.658	0.651	0.829	0.802	0.825	0.812	0.851	0.837	0.741	0.665

Table S5. Accuracy values (training and test sets) of ARKA models for the five data sets

Modeling algorithm	Data set 1		Data set 2		Data set 3		Data set 4		Data set 5	
	Trainin g	Test	Training	Test	Training	Test	Trainin g	Test	Training	Test
LDA	0.651	0.678	0.812	0.761	0.825	0.489	0.851	0.806	0.722	0.688
SVM	0.665	0.678	0.821	0.783	0.825	0.489	0.851	0.839	0.741	0.875
RF	0.996	0.624	1	0.674	0.835	0.466	0.892	0.903	0.833	0.813
LR	0.658	0.683	0.829	0.761	0.825	0.477	0.851	0.807	0.741	0.813

Table S6. Analysis of the number of outliers using the leverage approach

Datasets	***QSAR outliers***	$n_{Descriptors}$ ***(QSAR)***	***ARKA outliers***	$n_{Descriptors}$ ***(ARKA)***	$\boldsymbol{n_{TRAIN}}$	$\boldsymbol{n_{TEST}}$

	TRAIN	*TEST*		*TRAIN*	*TEST*			
1	2	4	14	11	3	2	269	202
2	6	6	8	0	3	2	117	46
3	3	12	6	3	11	2	97	88
4	4	0	4	3	0	1	74	31
5	3	1	4	0	0	2	54	16

$n_{Descriptors}$ = The number of descriptors

n_{TRAIN} = The number of data points in the training set

n_{TEST} = The number of data points in the test set

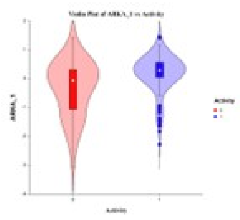
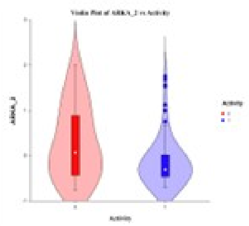
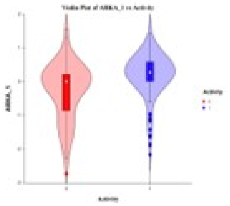
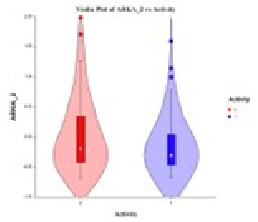
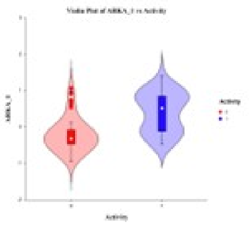
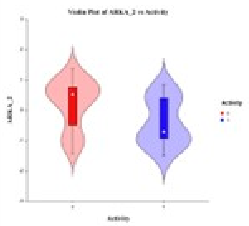
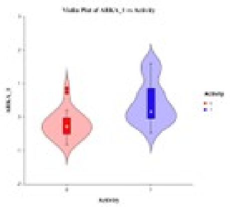
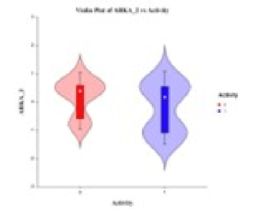
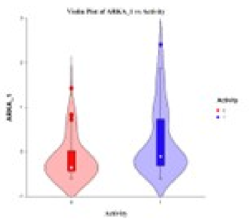
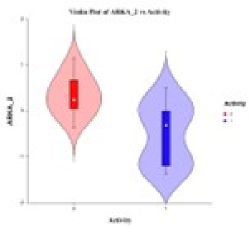
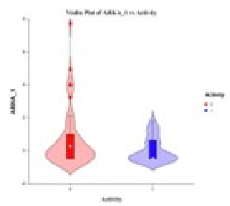
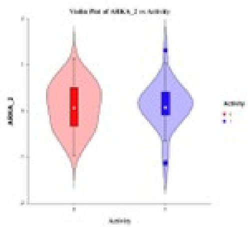
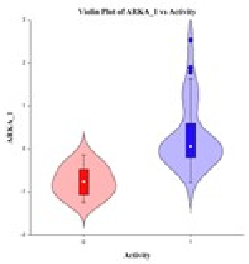
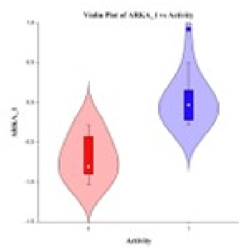
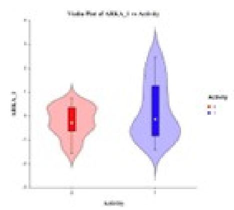
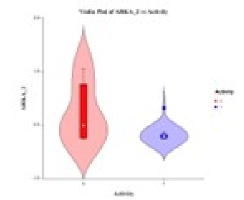
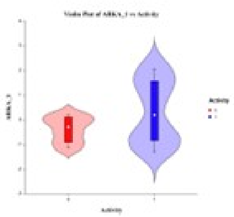
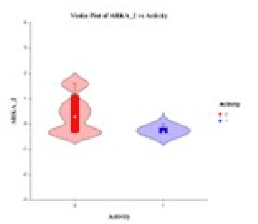
DS	*TRAINING*		*TEST*	
	ARKA_1 vs Activity	*ARKA_2 vs Activity*	*ARKA_1 vs Activity*	*ARKA_2 vs Activity*
1				
2				
3				
4				
5				

Figure S1. Violin Plots of ARKA_1 vs Activity and ARKA_2 vs Activity for all the five datasets (DS = Data set).

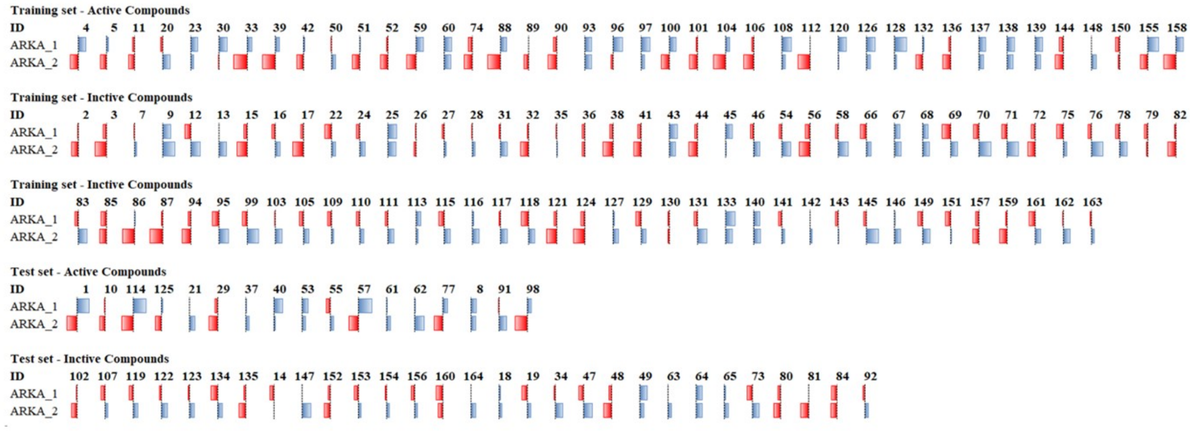


Figure S2. Analysis of the values of ARKA_1 and ARKA_2 for the active and inactive compounds in the training and test sets of a representative example (Dataset 2; blue indicates a positive value while red indicates a negative value).

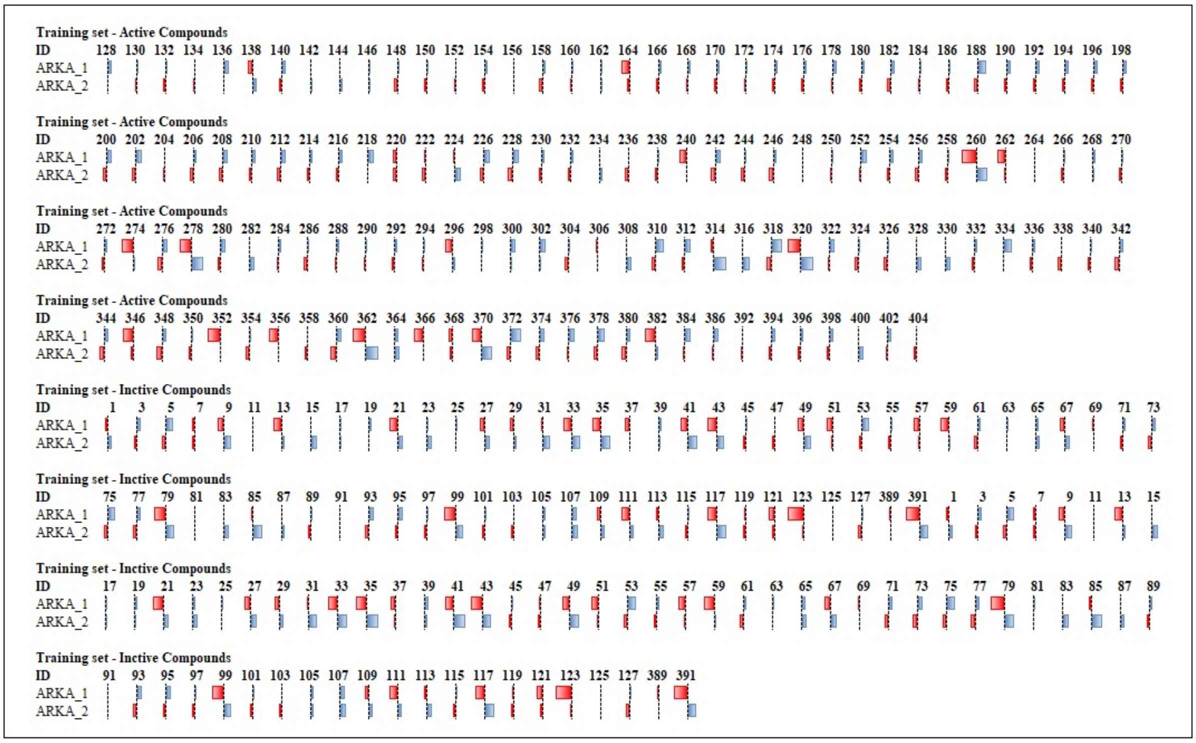


Figure S3. Data Bar Plots for the training set compounds demonstrating the variation of the ARKA_1 and ARKA_2 values (Dataset 1).

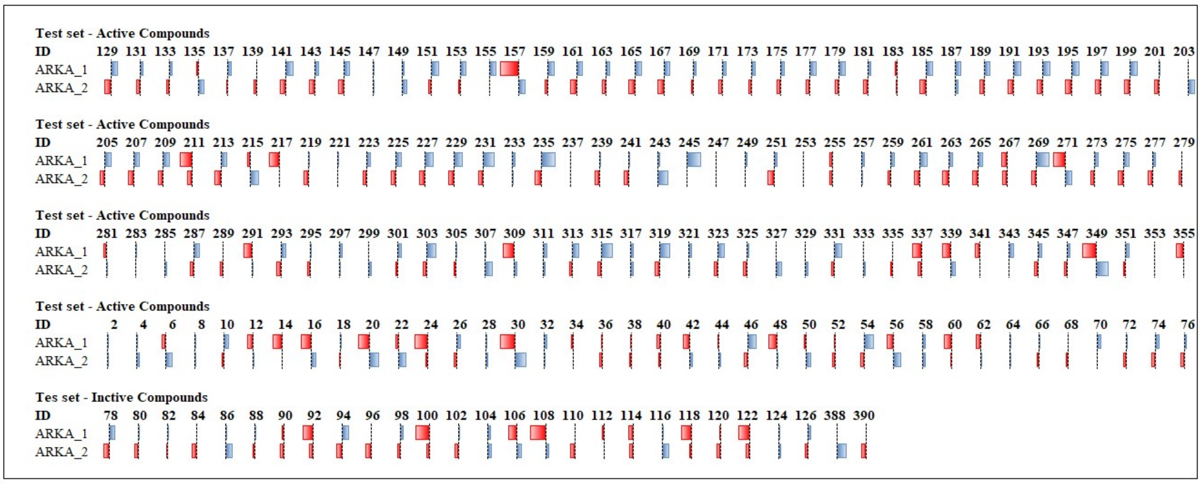


Figure S4. Data Bar Plots for the test set compounds demonstrating the variation of the ARKA_1 and ARKA_2 values (Dataset 1).

Data set modelability

The data set modelability was studied using ARKA descriptors for Dataset 2 (**Figure S5**), showing similar results to Dataset 1 (**Figure 8** of the main manuscript). Thus, the ARKA descriptors may be used in detecting serious activity/prediction cliffs and understanding the modelability of a data set. This may be exercised even in case of regression modeling problems considering the whole data set response mean as the threshold and then classifying the response values as positives or negatives. For initial modelability analysis (when the contributing features are not known), one may proceed with computing the ARKA descriptors starting with the (whole) pretreated pool of descriptors while for the final model development, one should use only the selected features.

In the case of Dataset 3, the scatter plot of ARKA_2 vs. ARKA_1 for the training set (**Figure S6**) shows poor modelability of the data set (the negative data points in the fourth quadrant are near the Y-axis) while the plot for the test set (**Figure S6**) shows poor quality of projections in the fourth quadrant suggesting poor information content of the QSAR descriptors for predictions of the endpoint (consistent with the poor model performance as per **Figure 4**). This is evident just from the ARKA descriptor scatter plot even without performing any modeling.

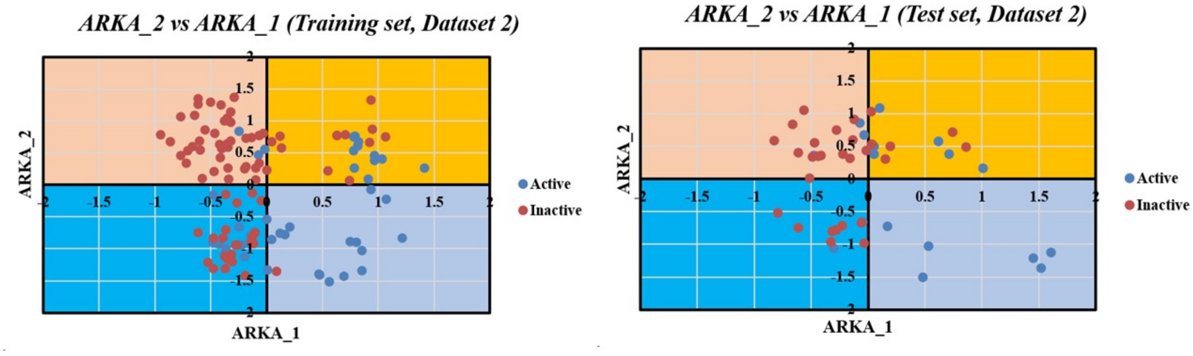


Figure S5. Scatter Plots of ARKA_2 vs ARKA_1 of Dataset 2.

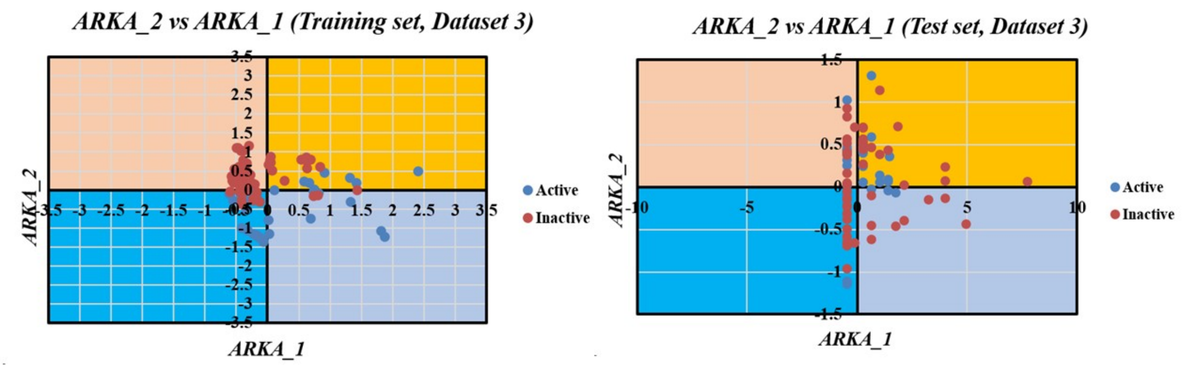


Figure S6. Scatter Plots of ARKA_2 vs ARKA_1 of Dataset 3.