Supporting Information

Improving Molecular Machine Learning Through Adaptive Subsampling with Active Learning

Yujing Wen[a], Zhixiong Li[a], Yan Xiang[a], Daniel Reker*[a]

[a] *Department of Biomedical Engineering, Duke University, Durham, North Carolina 27705, United States*

AUTHOR INFORMATION

* Corresponding Author

E-mail address: daniel.reker@duke.edu

Table S1. Percentage increase of different performance metrics when using active learning-based subsampling relative to training models on the full dataset.

Dataset	Percentage increase in MCC	Percentage increase in F1	Percentage increase in ACC	Percentage increase in BAS
BBBP	4.25%	0.40%	0.90%	3.40%
BACE	5.07%	3.71%	1.29%	2.19%
Clintox	129%	139%	2.72%	3.01%
HIV	5.52%	6.01%	0.55%	0.93%

Table S2. Class balance in the active learning-based dataset compared to the original class imbalance.

Dataset	Imbalance (% positive data)	
	Full dataset	maxIter set $T_{n_{max}}$
BBBP	77.92%	58.38%
BACE	53.04%	51.57%
ClinTox	7.03%	29.68%
HIV	3.81%	21.22%

Table S3: Number of unique molecules selected by the 20 active learning runs.

Dataset	Number of unique molecules selected		Total number of training data molecules
	in every of the 20 active learning runs	at least once during 20 active learning runs	
BBBP	70	700	1019
BACE	224	702	756
Clintox	14	342	739
HIV	647	6784	20563

Table S4: Performance of sampling methods on datasets without error introduction. Performance is shown as percentage of maximum performance achieved per dataset. The last column shows the median performance of a method across all datasets. The best performing method per column (i.e., per dataset or across all datasets as measured by median) is highlighted in bold.

	BBBP	BACE	Clintox	HIV	B. Cancer	Median
Full model	0.96	0.95	0.29	0.85	0.97	0.95
AllKNN	0.98	0.90	0.81	0.94	0.96	0.95
ClusterCentroids	0.78	0.98	0.35	-0.02	0.98	0.94
CondensedNearestNeighbour	0.58	0.71	0.89	0.98	0.93	0.78
EditedNearestNeighbours	0.97	0.78	0.72	0.96	0.96	0.89
InstanceHardnessThreshold	0.78	0.90	0.66	**1.00**	0.87	0.96
NearMiss	0.30	0.78	0.10	0.10	0.97	0.87
NeighbourhoodCleaningRule	0.99	0.88	0.72	0.96	0.96	0.30
OneSidedSelection	0.96	0.95	0.34	0.91	0.95	0.96
RandomUnderSampler	0.92	0.95	0.70	0.52	**1.00**	0.95
RepeatedEditedNearestNeighbours	0.96	0.78	0.74	0.94	0.92	0.92
RandomOverSampler	0.97	0.95	**1.00**	0.96	0.94	0.95
SMOTEN	0.93	0.97	0.30	0.74	0.95	0.95
Balanced	0.87	0.89	0.61	0.61	0.92	0.96
Diverse	0.90	1.00	0.20	0.14	NA	0.87
Balanced-Diverse	0.82	0.93	0.45	0.44	NA	0.55
Diverse-Balanced	0.85	0.99	0.47	0.44	NA	0.64
Active Learning	**1.00**	**1.00**	0.67	0.90	0.98	**0.98**

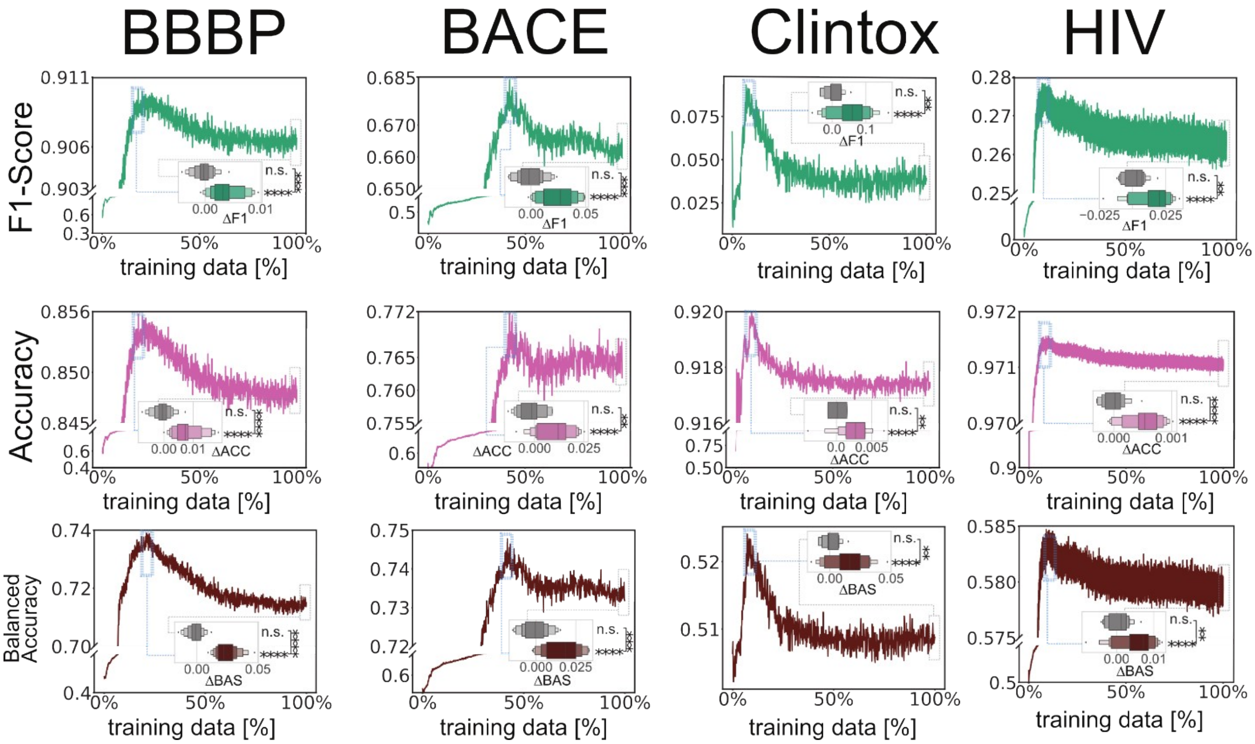

Figure S1. Evaluation of active learning subsampling using other evaluation metrics: F1 score, accuracy, and balanced accuracy.

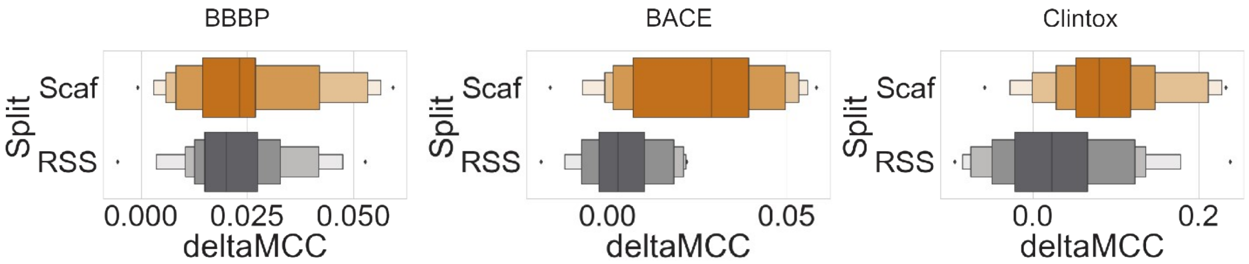

Figure S2. Improvements in performance for active learning-based subsampling when using scaffold-based train-test splits ("scaf") compared to random stratified splits ("RSS").