Electronic Supplementary Material (ESI) for Molecular Systems Design & Engineering.

Applying local interpretable model-agnostic explanations to identify substructures that are responsible for mutagenicity of chemical compounds

Lucca Caiaffa Santos Rosa and Andre Silva Pimentel*

Departamento de Química, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, RJ 22453-900, Brazil.

*Corresponding author: a_pimentel@puc-rio.br

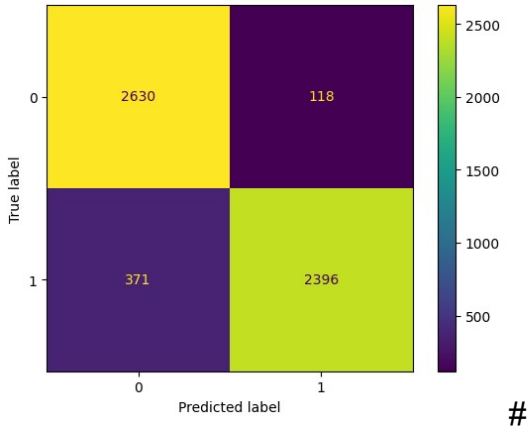
Supplementary Material

Table S1. Metrics (mean ROC-AUC) of the training and validation datasets for the Rain Forest (RF), and Extra Trees (ET) classifier models for three different runs (#1, #2, and #3) with 10x5-nested cross-validation after resampling for the mutagenicity of compounds using the Hansen and Bursi Ames mutagenicity datasets.

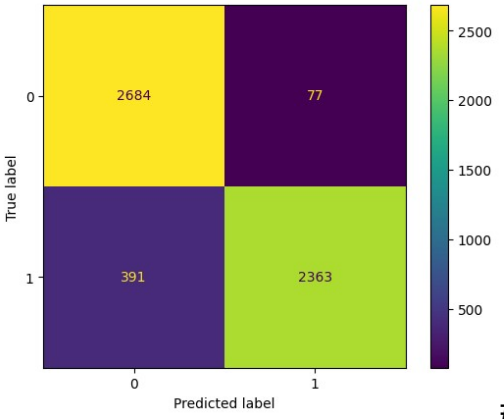
		ROC-AUC score	
Dataset	Model	Precision of Training	Precision of validation
Hansen	RF	0.90	0.83
	ET	0.92	0.84
Bursi	RF	0.92	0.90
	ET	0.94	0.91

Table S2. The precision, recall, F1, and accuracy, Matthew correlation coefficient (MCC), and Cohen´s kappa (κ) scores for the Random Forest (RF) and Extra Trees (ET) classifier models in three different runs (#1, #2, and #3) for the training dataset of mutagenicity of compounds using resampling method with 5-fold cross-validation.

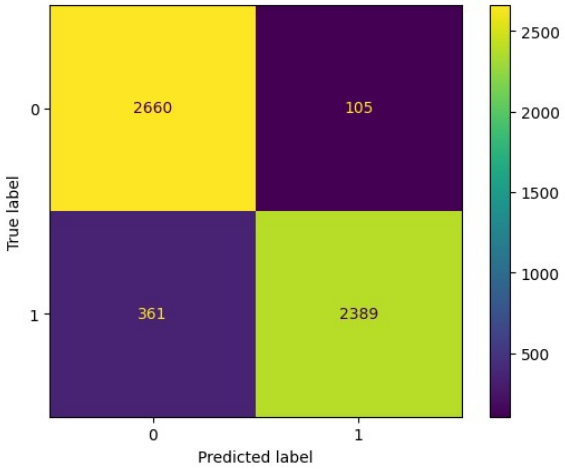
Dataset	Model	Scores					
		Precision	Recall	F1	Accuracy	MCC	κ
Bursi	RF #1	0.978	0.878	0.929	0.929	0.863	0.859
	RF #2	0.977	0.891	0.935	0.935	0.874	0.870
	RF #3	0.974	0.893	0.933	0.933	0.870	0.867
	ET #1	0.988	0.906	0.948	0.948	0.898	0.895
	ET #2	0.991	0.919	0.955	0.955	0.913	0.910
	ET #3	0.990	0.916	0.953	0.953	0.909	0.907
Hansen	RF #1	0.953	0.866	0.911	0.911	0.826	0.823
	RF #2	0.968	0.858	0.915	0.915	0.835	0.830
	RF #3	0.958	0.868	0.915	0.915	0.834	0.830
	ET #1	0.970	0.877	0.925	0.925	0.854	0.851
	ET #2	0.973	0.895	0.935	0.935	0.874	0.871
	ET #3	0.976	0.887	0.933	0.933	0.870	0.867



#1

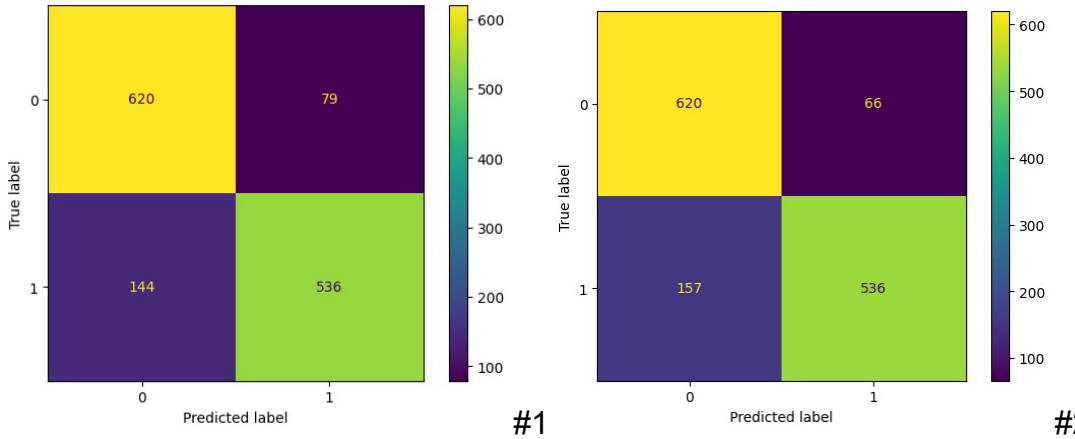


#2

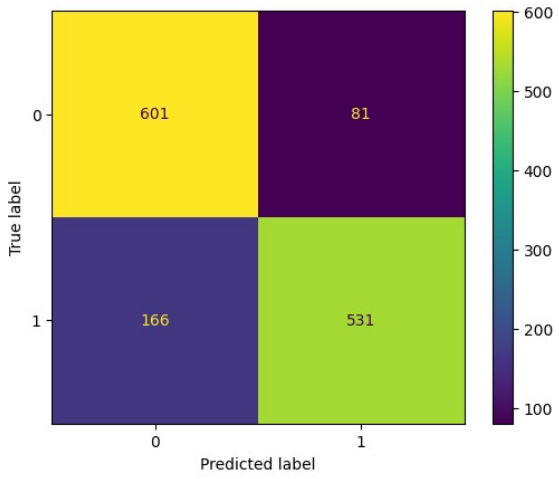


#3

Figure S1. The confusion matrix for the RF classification models using resampling method and 5-fold cross-validation of the Hansen mutagenicity training dataset for the runs #1, #2, and #3.

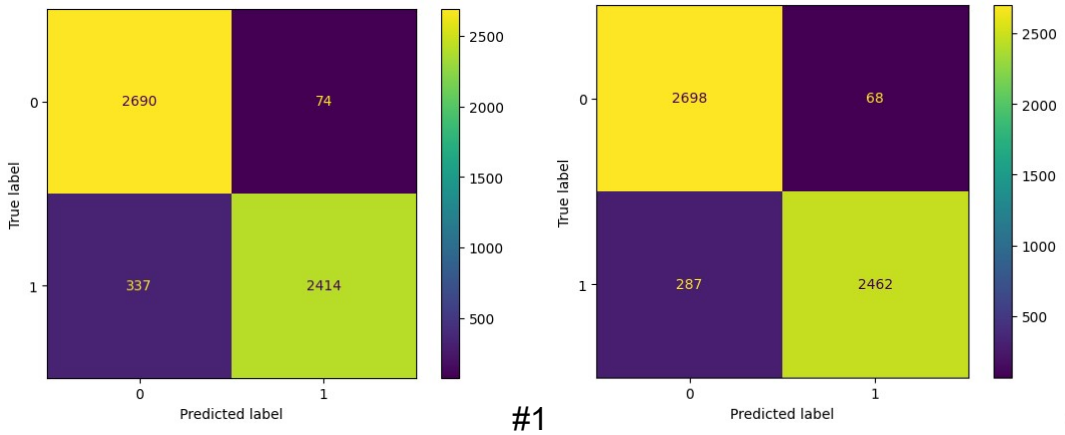


#1 #2



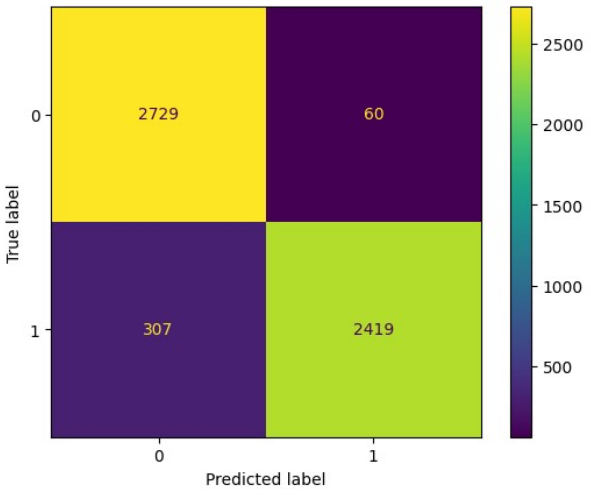
#3

Figure S2. The confusion matrix for the RF classification models using resampling method and 5-fold cross-validation of the Hansen mutagenicity validation dataset for the runs #1, #2, and #3.



#1

#2



#3

Figure S3. The confusion matrix for the ET classification models using resampling method and 5-fold cross-validation of the Hansen mutagenicity training dataset for the runs #1, #2, and #3.

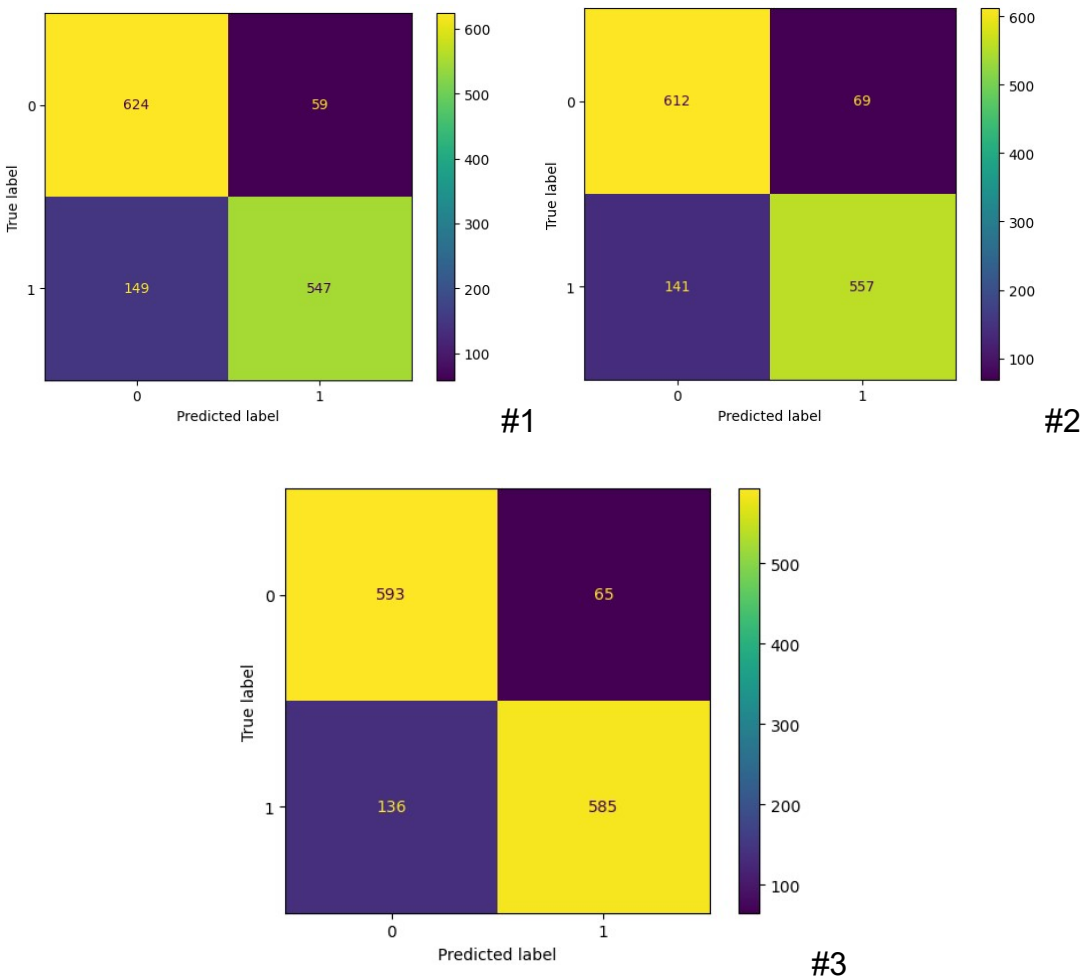
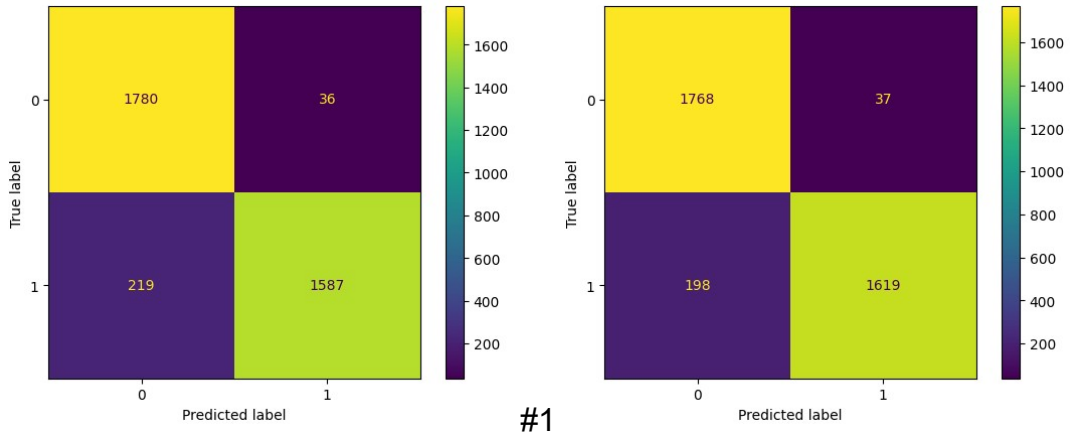


Figure S4. The confusion matrix for the ET classification models using resampling method and 5-fold cross-validation of the Hansen mutagenicity validation dataset for the runs #1, #2, and #3.



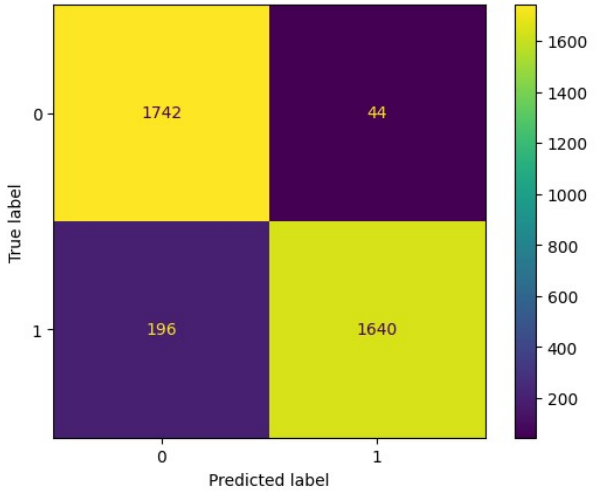
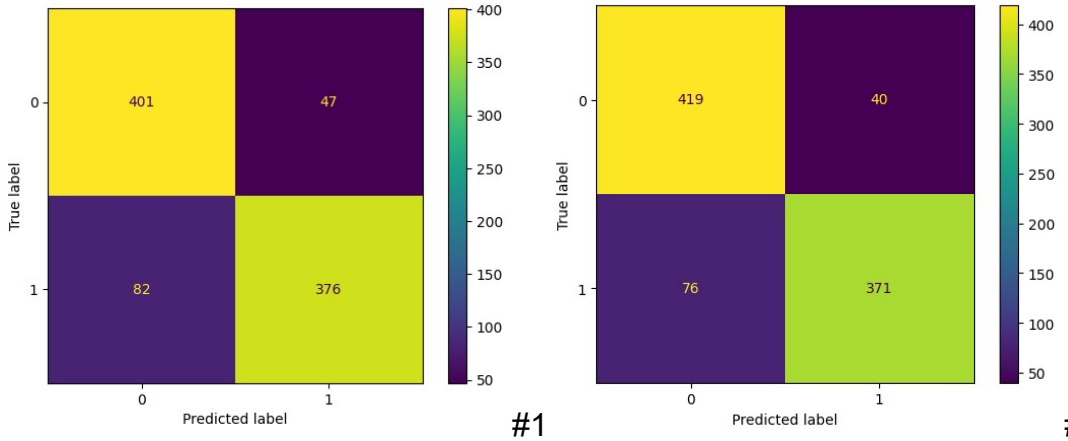


Figure S5. The confusion matrix for the RF classification models using resampling method and 5-fold cross-validation of the Bursi mutagenicity training dataset for the runs #1, #2, and #3.



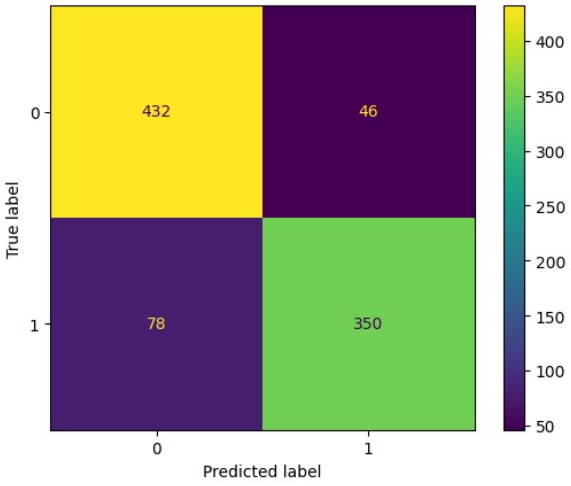
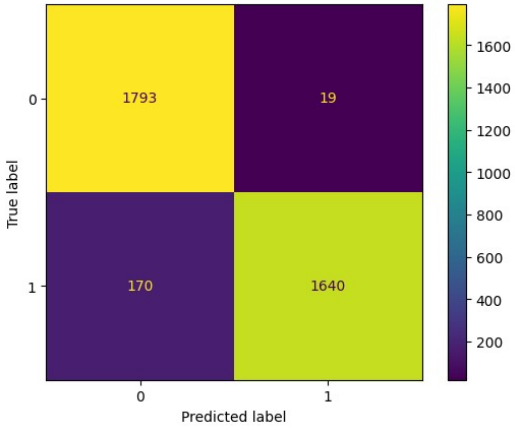
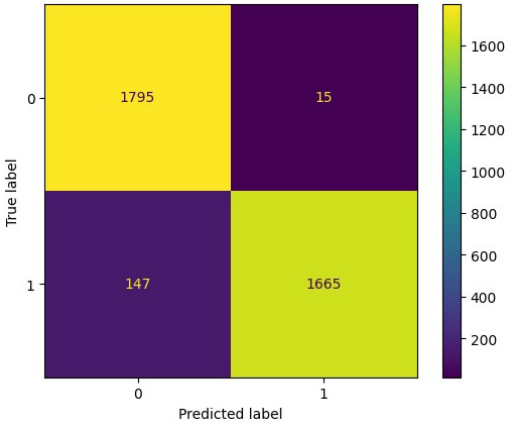


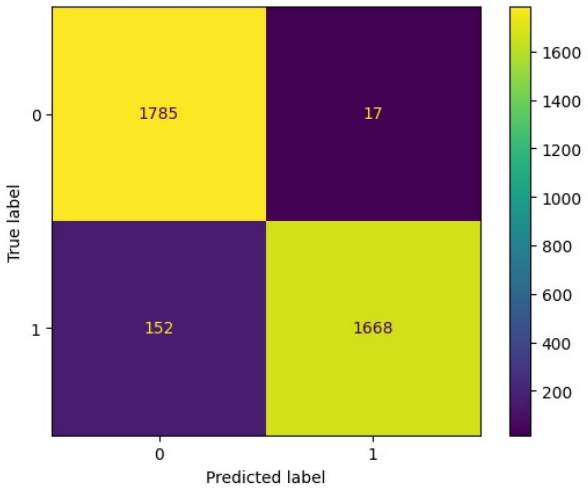
Figure S6. The confusion matrix for the RF classification models using resampling method and 5-fold cross-validation of the Bursi mutagenicity validation dataset for the runs #1, #2, and #3.



#1

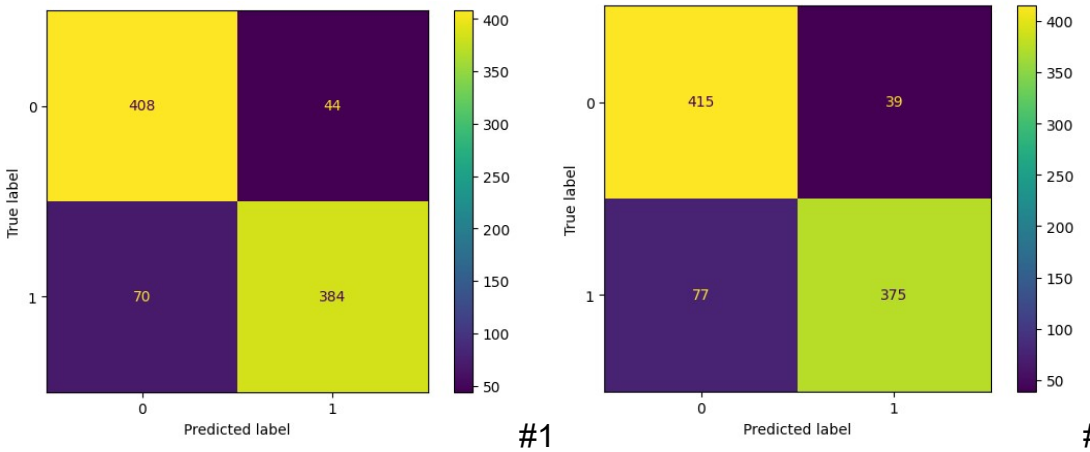


#2

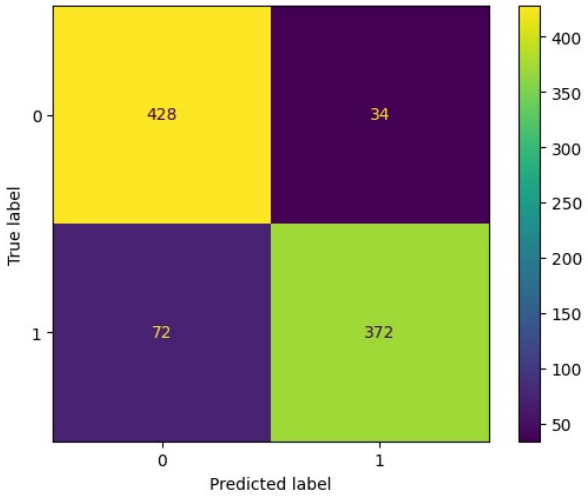


#3

Figure S7. The confusion matrix for the ET classification models using resampling method and 5-fold cross-validation of the Bursi mutagenicity training dataset for the runs #1, #2, and #3.

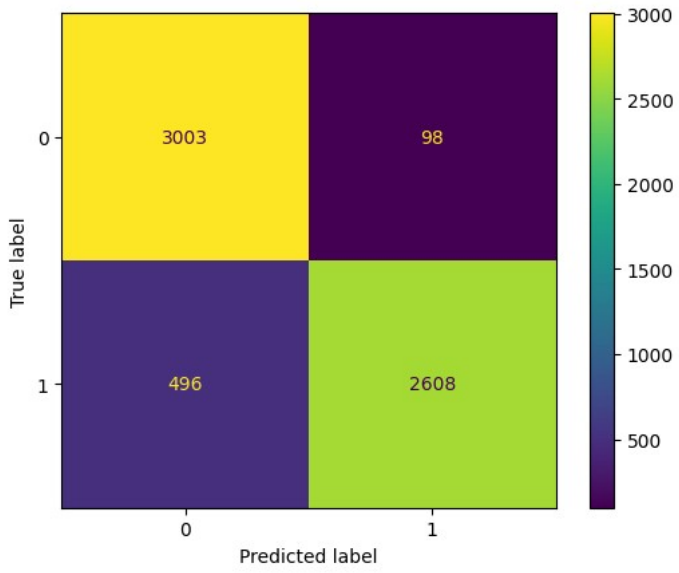


#1 #2

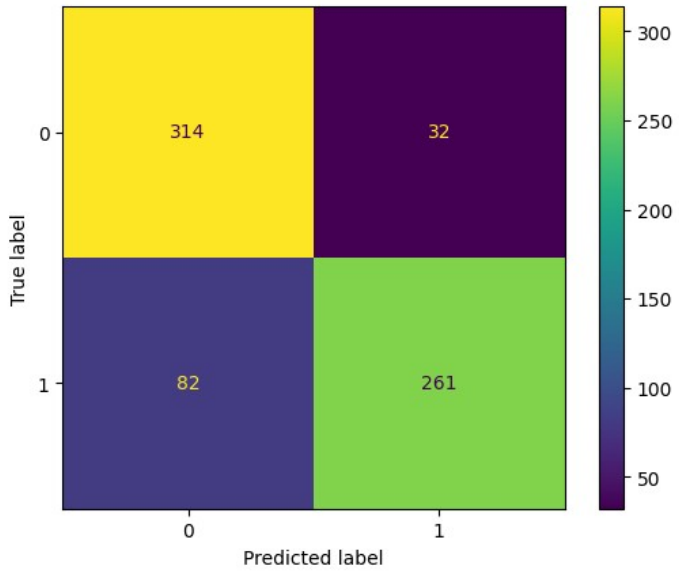


#3

Figure S8. The confusion matrix for the ET classification models using resampling method and 5-fold cross-validation of the Bursi mutagenicity validation dataset for the runs #1, #2, and #3.

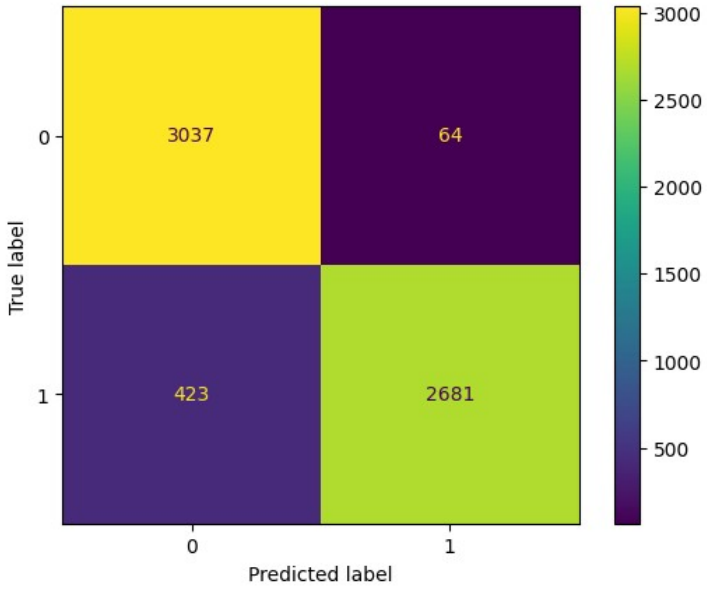


(A)

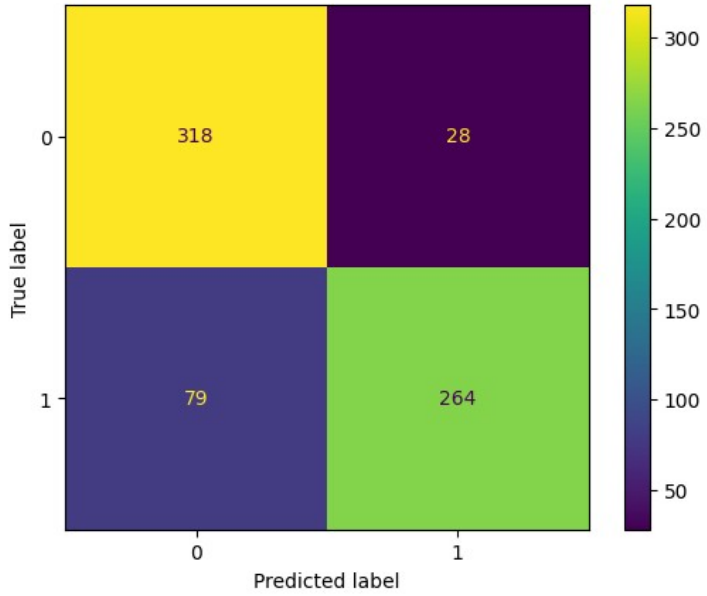


(B)

Figure S9. The confusion matrix for the RF classification models using resampling method and 10x5-nested cross-validation of the Hansen mutagenicity training (A) and validation (B) datasets.

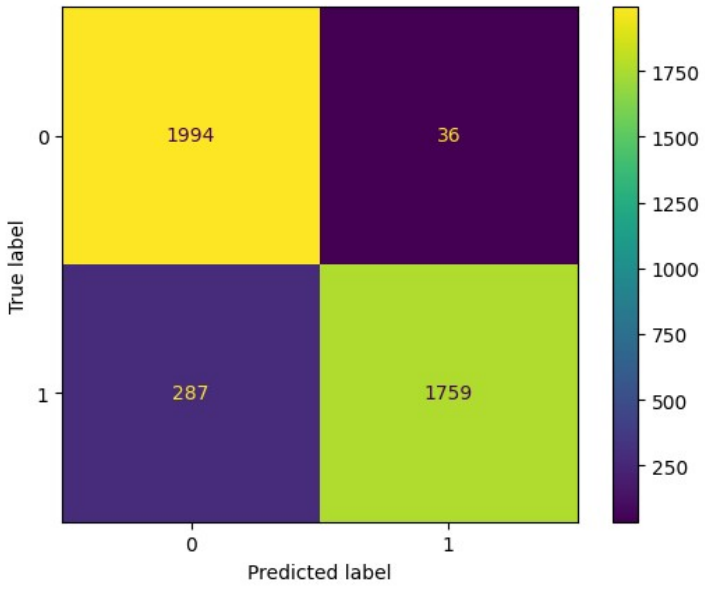


(A)

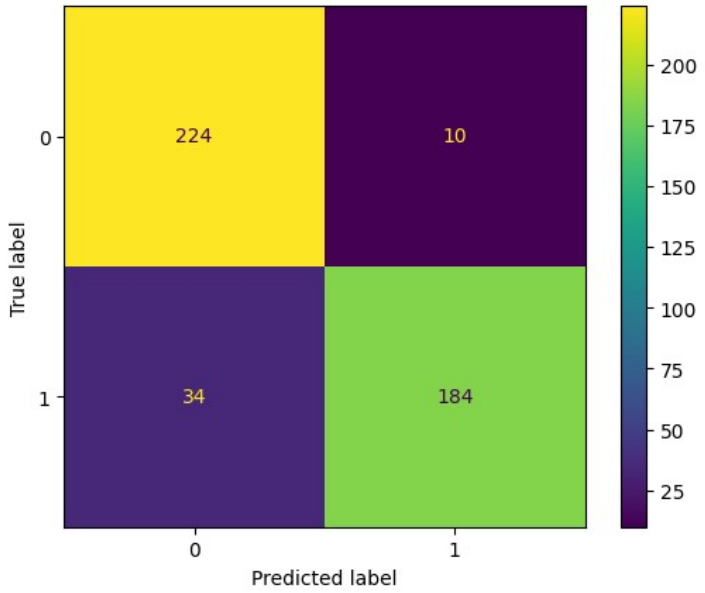


(B)

Figure S10. The confusion matrix for the ET classification models using resampling method and 10x5-nested cross-validation of the Hansen mutagenicity training (A) and validation (B) datasets.

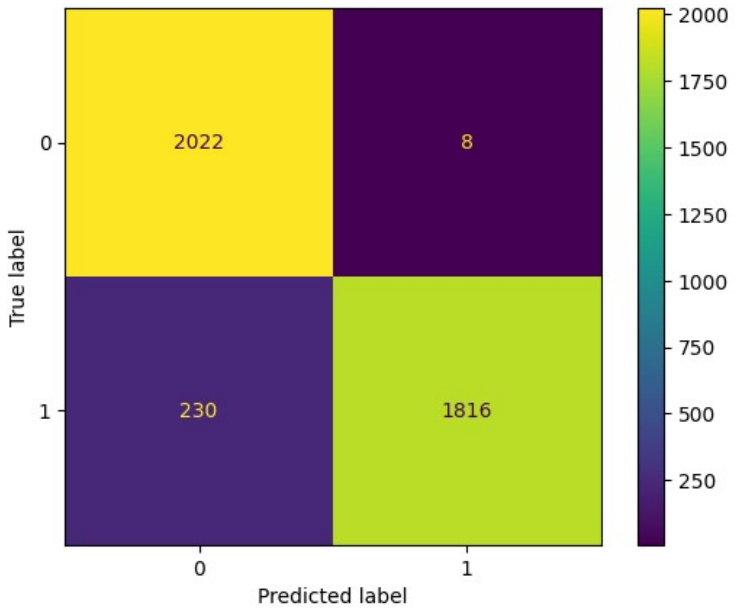


(A)

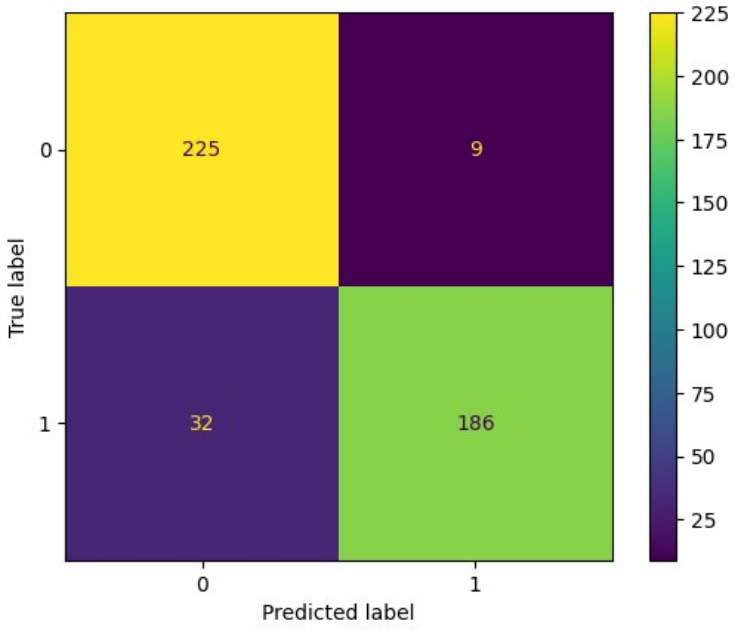


(B)

Figure S11. The confusion matrix for the RF classification models using resampling method and 10x5-nested cross-validation of the Bursi mutagenicity training (A) and validation (B) datasets.



(A)



(B)

Figure S12. The confusion matrix for the ET classification models using resampling method and 10x5-nested cross-validation of the Bursi mutagenicity training (A) and validation (B) datasets.

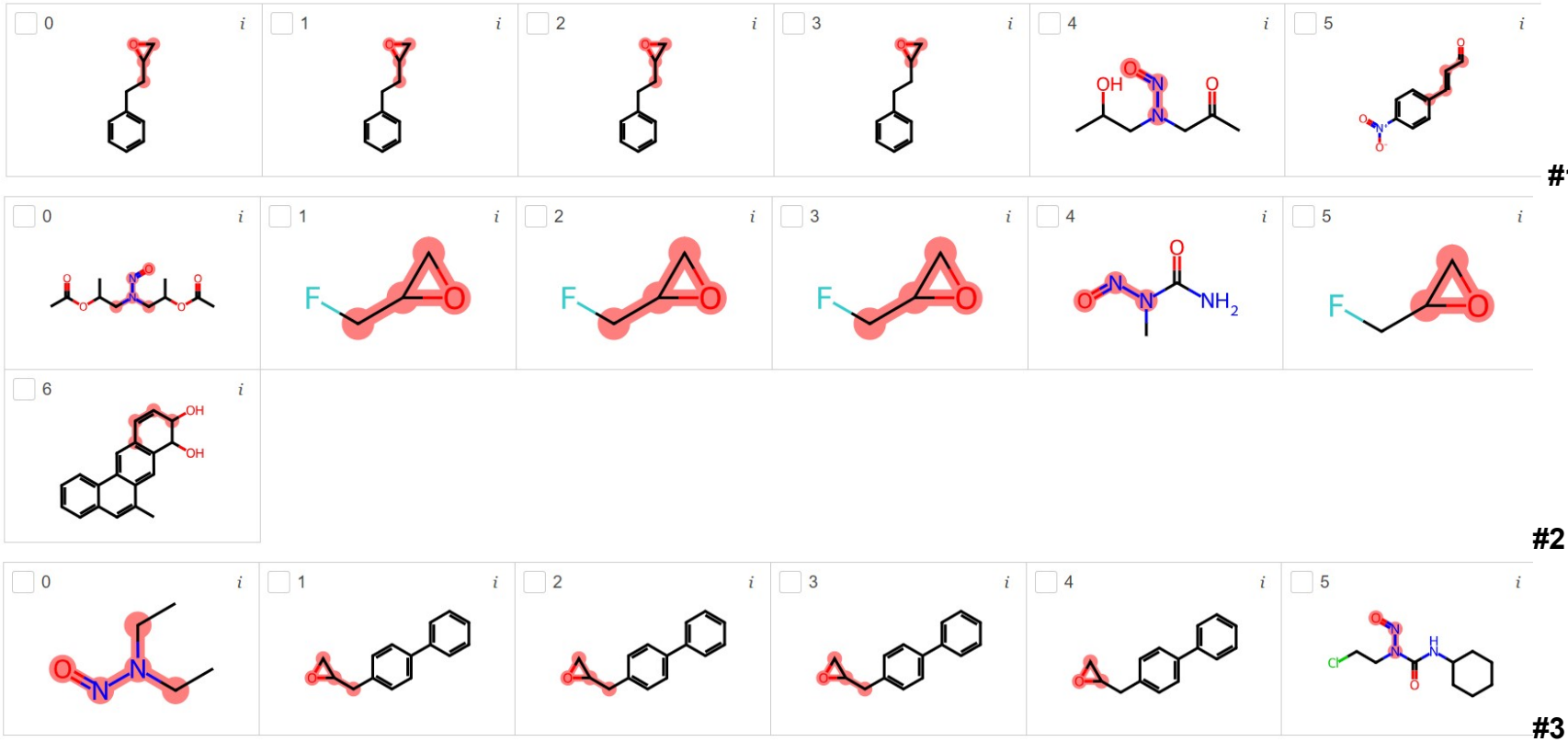


Figure S13. The most important substructures found by LIME for Bursi mutagenicity dataset using resampling method, 5-fold cross-validation, and the RF classification model for runs #1, #2, and #3.

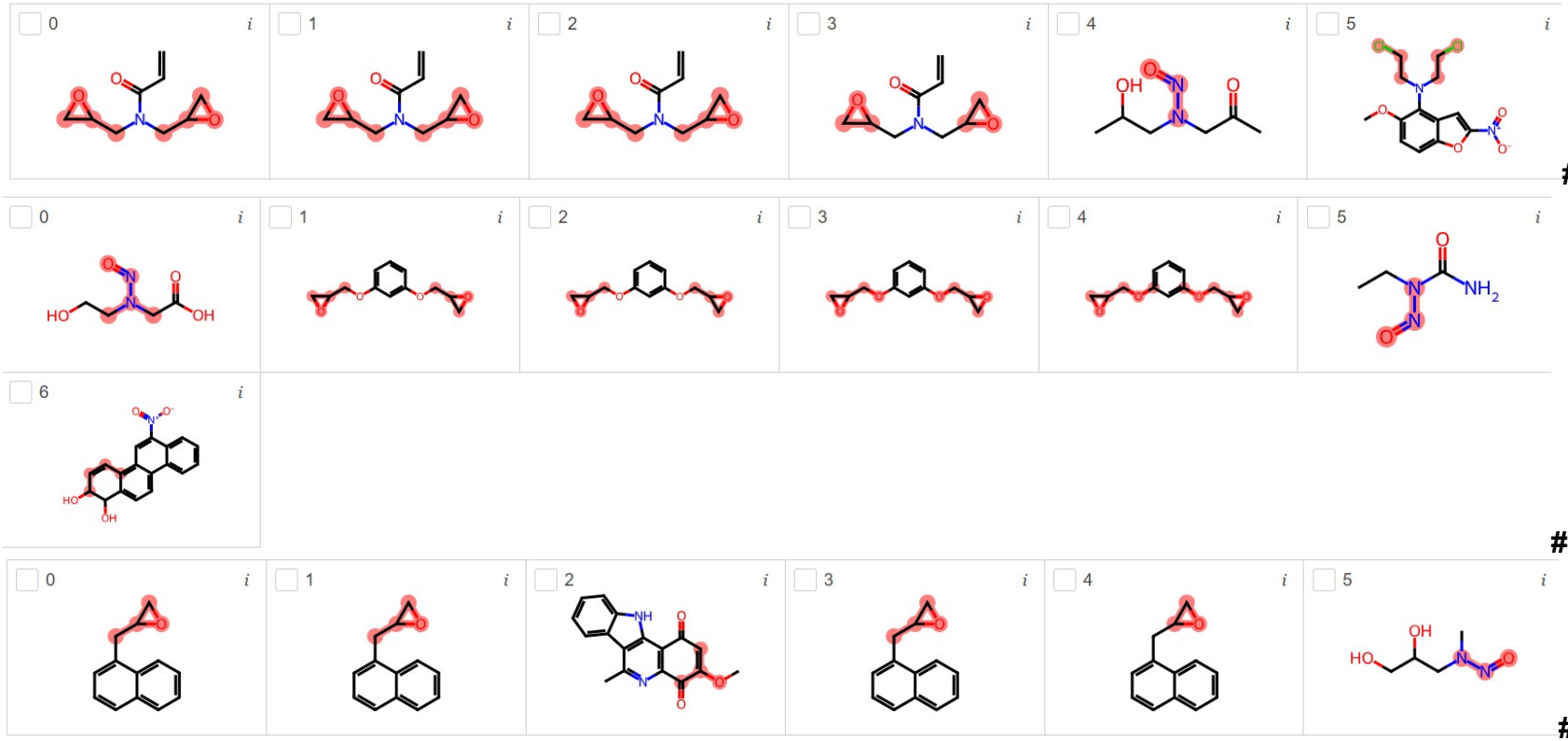


Figure S14. The most important substructures found by LIME for Bursi mutagenicity dataset using resampling method, 5-fold cross-validation, and the ET classification model for runs #1, #2, and #3.

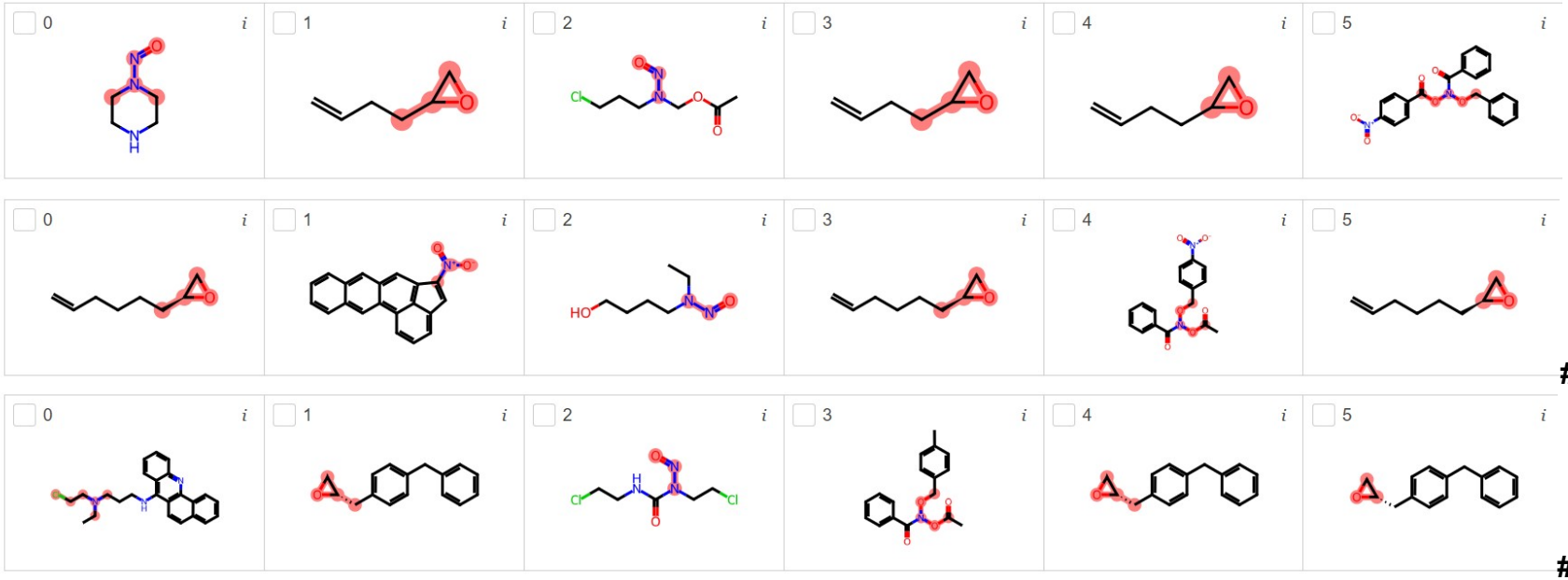


Figure S15. The most important substructures found by LIME for Hansen mutagenicity dataset using resampling method, 5-fold cross-validation, and the RF classification model for runs #1, #2, and #3.

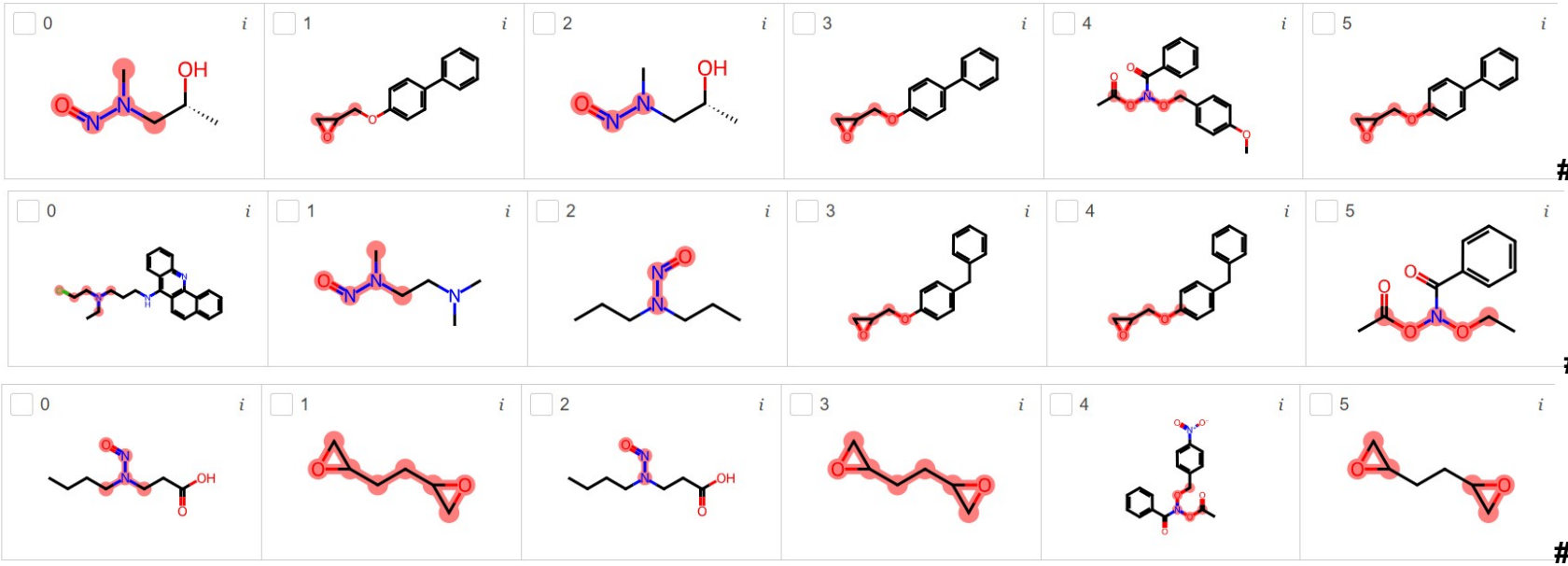
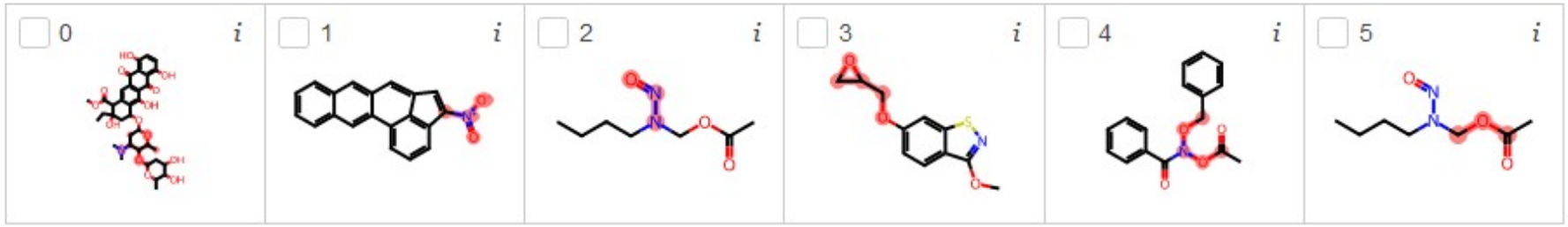
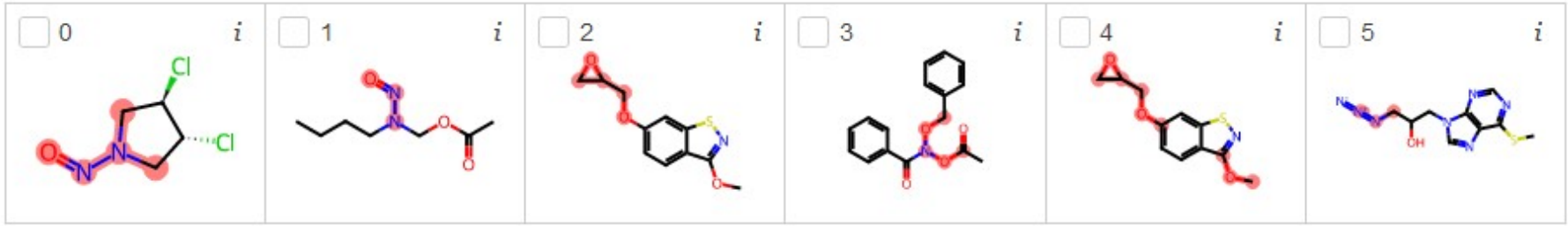


Figure S16. The most important substructures found by LIME for Hansen mutagenicity dataset using resampling method, 5-fold cross-validation, and the ET classification model for runs #1, #2, and #3.

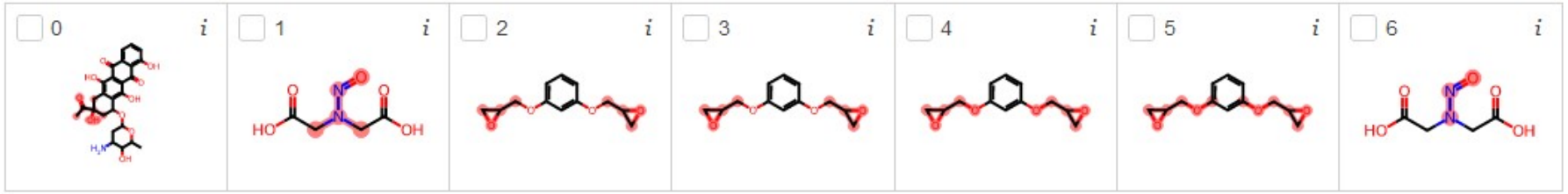


(A)

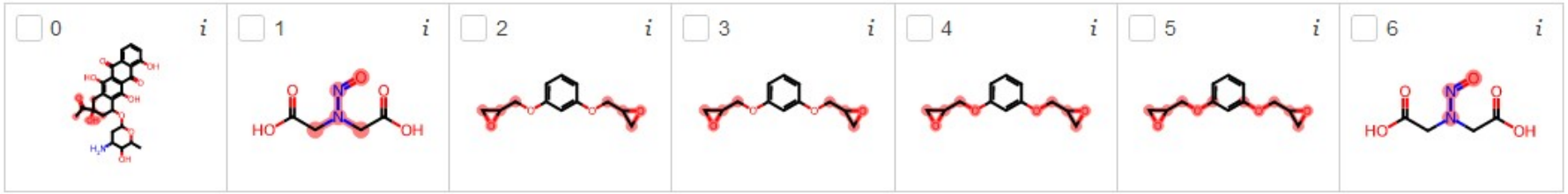


(B)

Figure S17. The most important substructures found by LIME for Hansen mutagenicity dataset using resampling method, 10x5-nested cross-validation, and the RF (A) and ET (B) classification models.



(A)



(B)

Figure S18. The most important substructures found by LIME for Bursi mutagenicity dataset using resampling method, 10x5-nested cross-validation, and the RF (A) and ET (B) classification models.