

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

4

5 Lucca Caiaffa Santos Rosa and Andre Silva Pimentel*

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

8

9 *Corresponding author: a_pimentel@puc-rio.br

10

11

Supplementary Material

12

13

14 **Table S1.** Metrics (mean ROC-AUC) of the training and validation datasets for the
15 Rain Forest (RF), and Extra Trees (ET) classifier models for three different runs (#1,
16 #2, and #3) with 10x5-nested cross-validation after resampling for the mutagenicity
17 of compounds using the Hansen and Bursi Ames mutagenicity datasets.

18

19

20

21

22

23

24

25

26

27

28

29

30

31

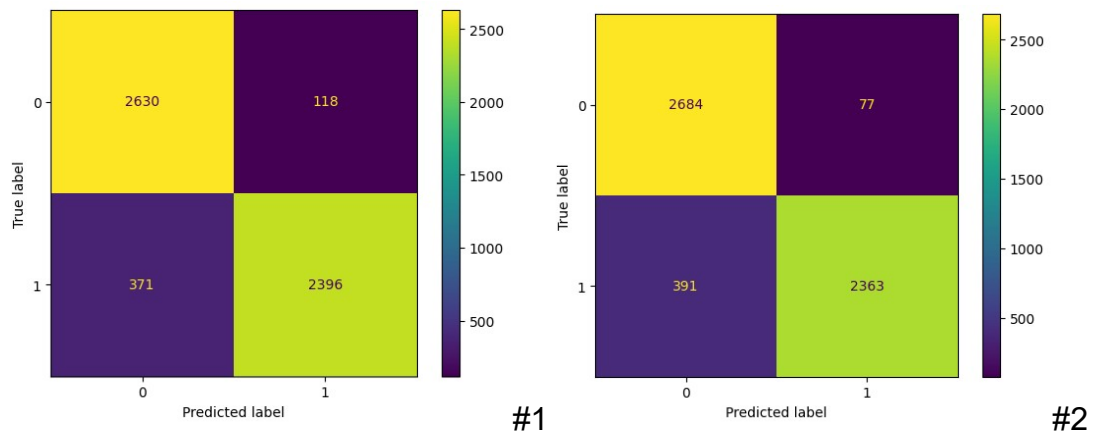
		ROC-AUC score	
Dataset	Model	Precision of Training	Precision of validation
Hansen	RF	0.90	0.83
	ET	0.92	0.84
Bursi	RF	0.92	0.90
	ET	0.94	0.91

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

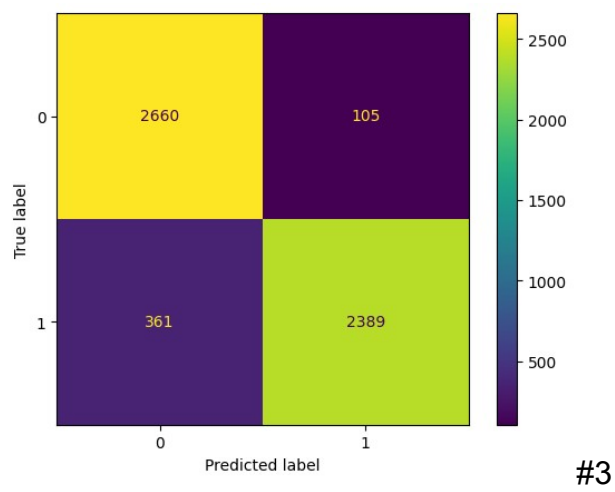
Dataset	Model	Scores					
		Precision	Recall	F1	Accuracy	MCC	κ
Bursi	RF #1	0.978	0.878	0.929	0.929	0.863	0.859
	RF #2	0.977	0.891	0.935	0.935	0.874	0.870
	RF #3	0.974	0.893	0.933	0.933	0.870	0.867
	ET #1	0.988	0.906	0.948	0.948	0.898	0.895
	ET #2	0.991	0.919	0.955	0.955	0.913	0.910
	ET #3	0.990	0.916	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

37
 38
 39

40



41

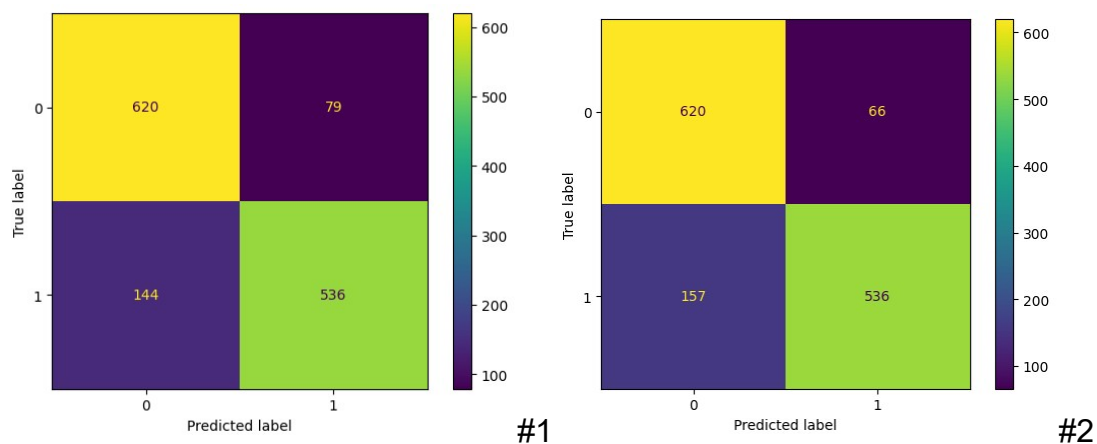


42

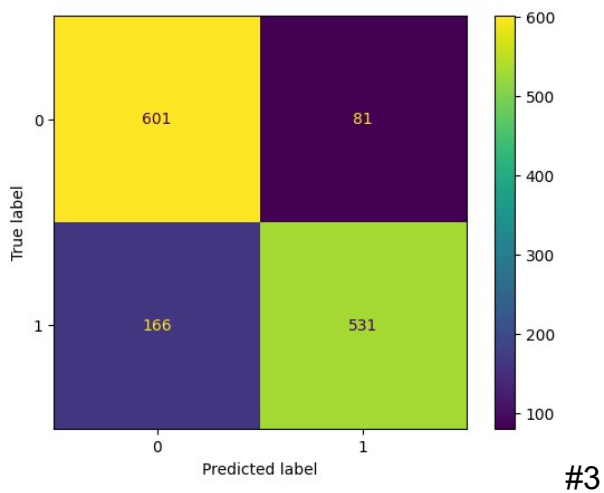
43 **Figure S1.** The confusion matrix for the RF classification models using resampling
44 method and 5-fold cross-validation of the Hansen mutagenicity training dataset for
45 the runs #1, #2, and #3.

46

47



48

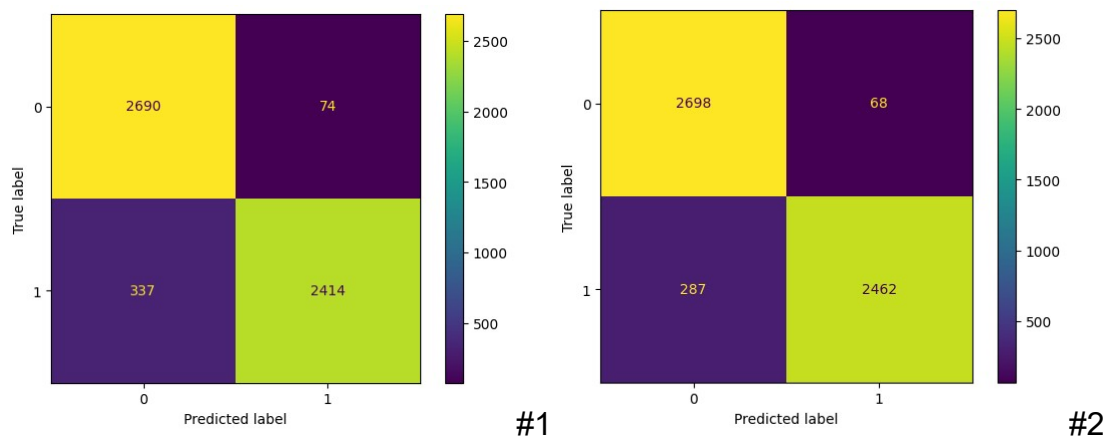


49

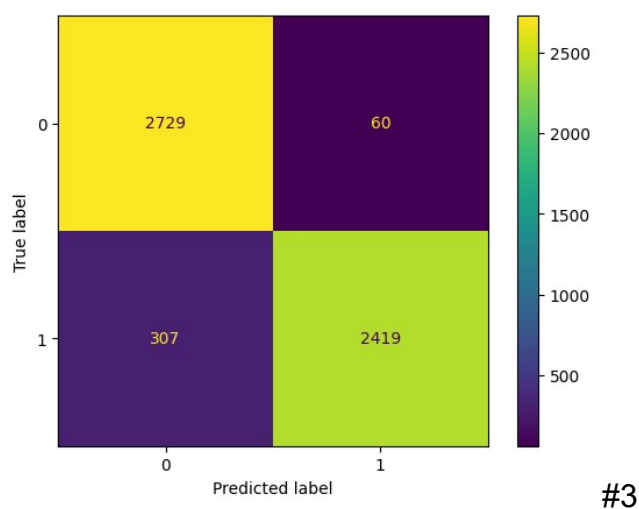
50 **Figure S2.** The confusion matrix for the RF classification models using resampling
51 method and 5-fold cross-validation of the Hansen mutagenicity validation dataset
52 for the runs #1, #2, and #3.

53

54



55

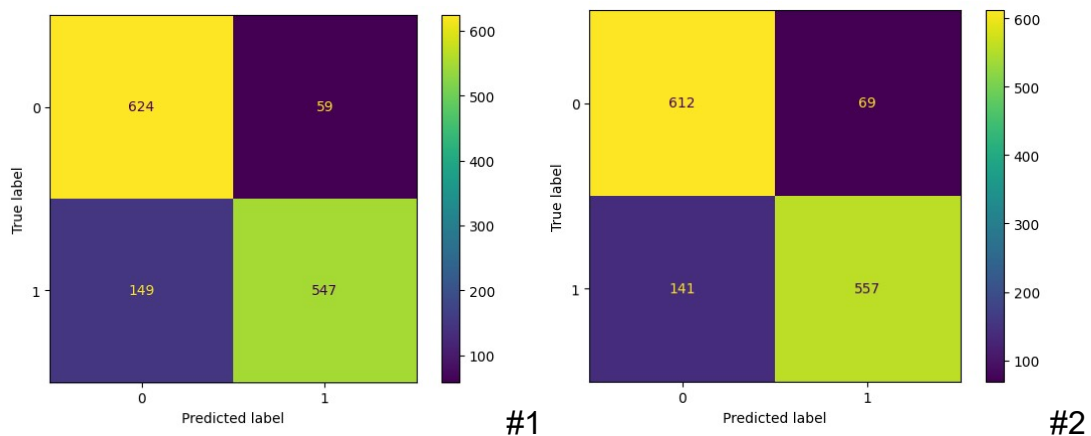


56

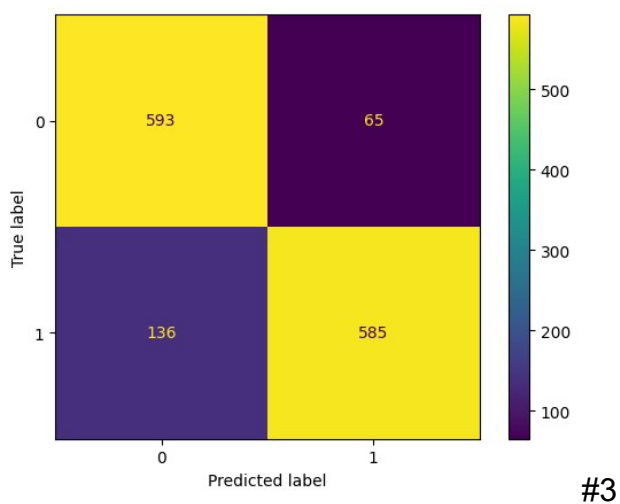
57 **Figure S3.** The confusion matrix for the ET classification models using resampling
58 method and 5-fold cross-validation of the Hansen mutagenicity training dataset for
59 the runs #1, #2, and #3.

60

61



62



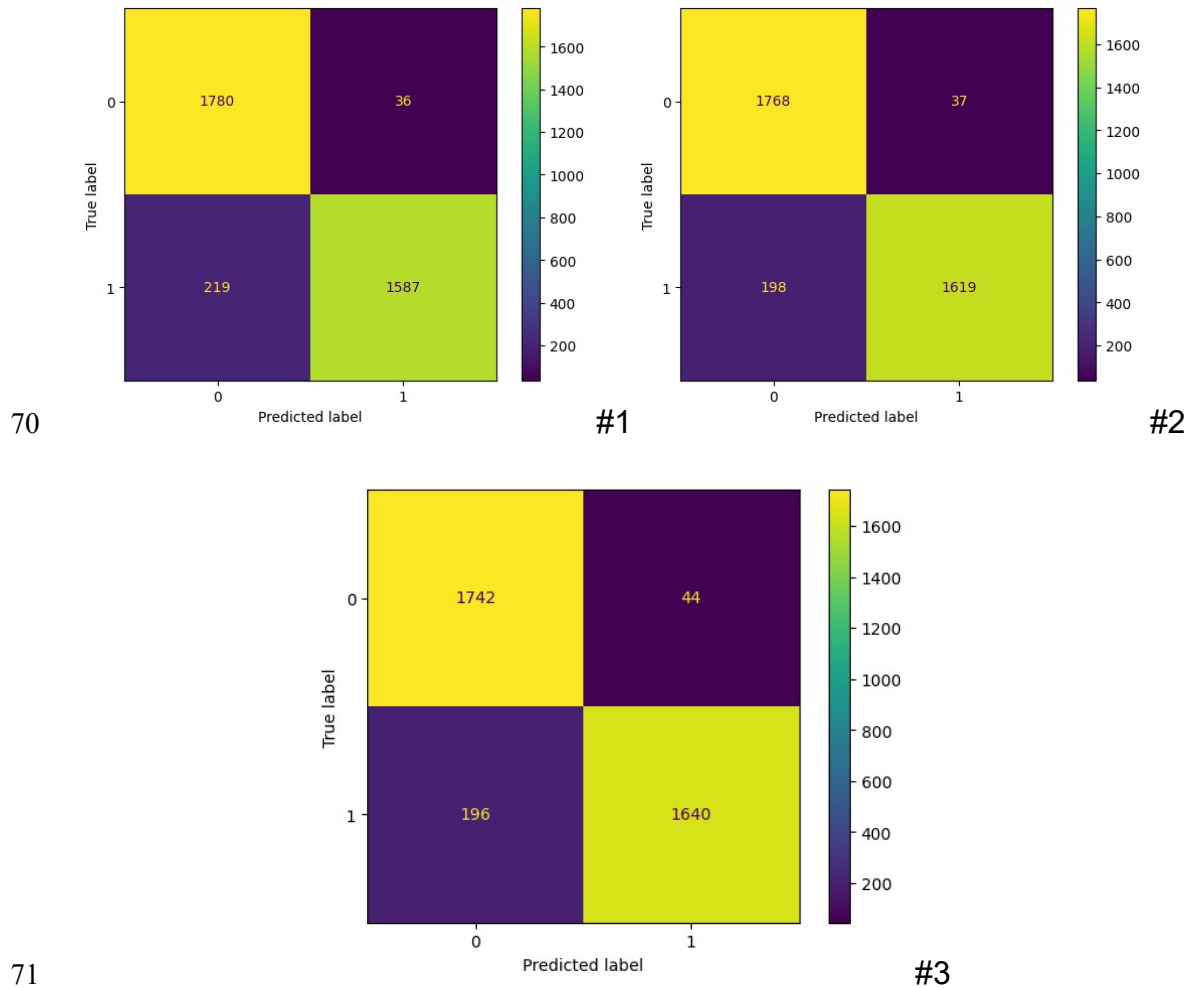
63

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

67

68

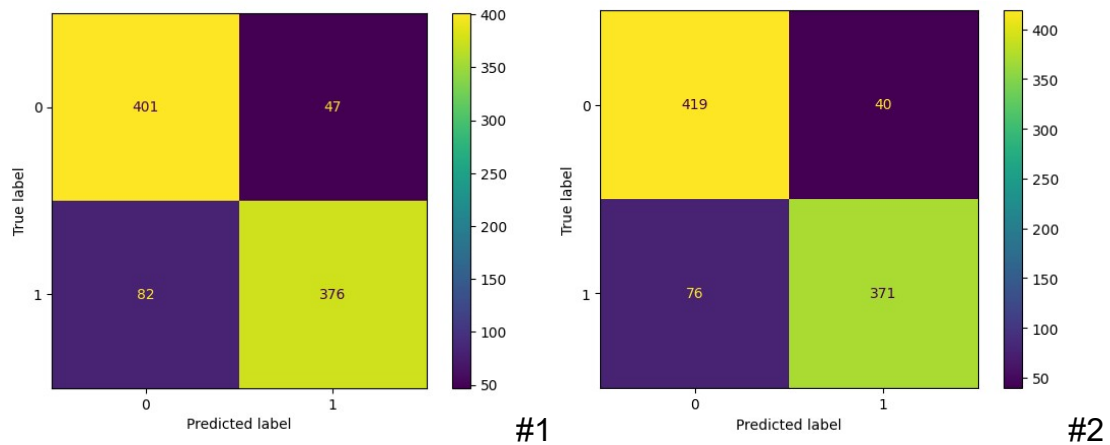
69



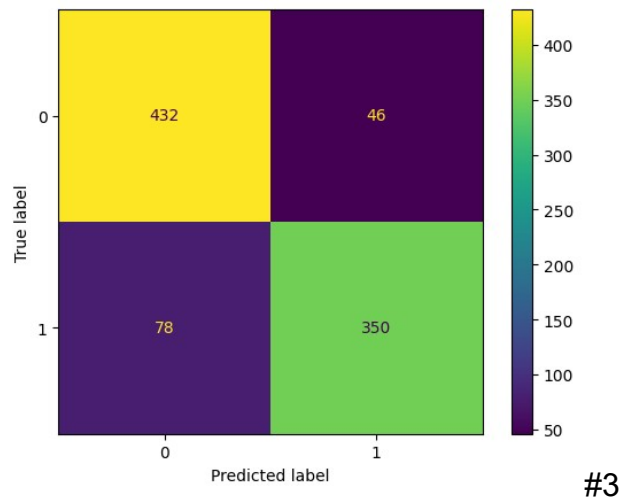
72 **Figure S5.** The confusion matrix for the RF classification models using resampling
73 method and 5-fold cross-validation of the Bursi mutagenicity training dataset for
74 the runs #1, #2, and #3.

75

76



77



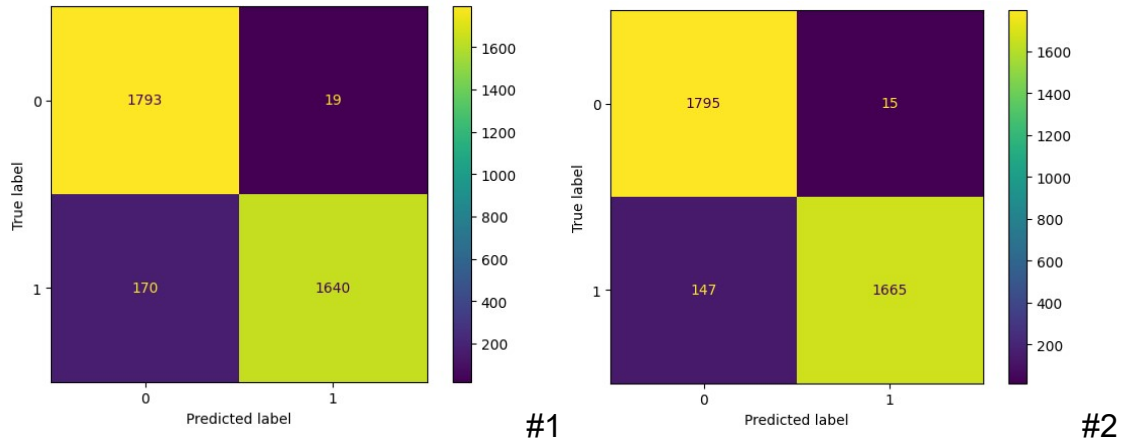
78

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

82

83

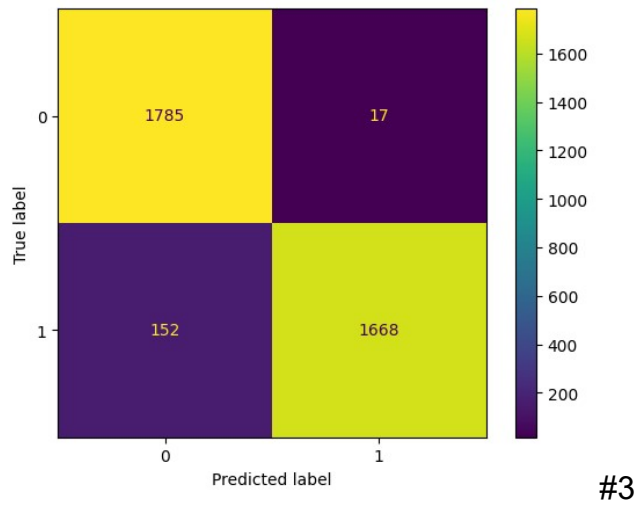
84



85

#1

#2



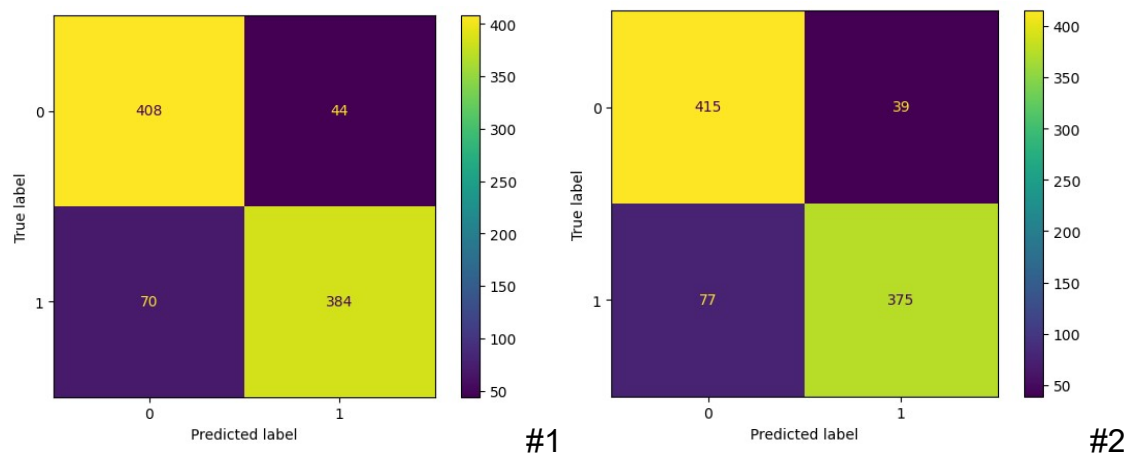
86

#3

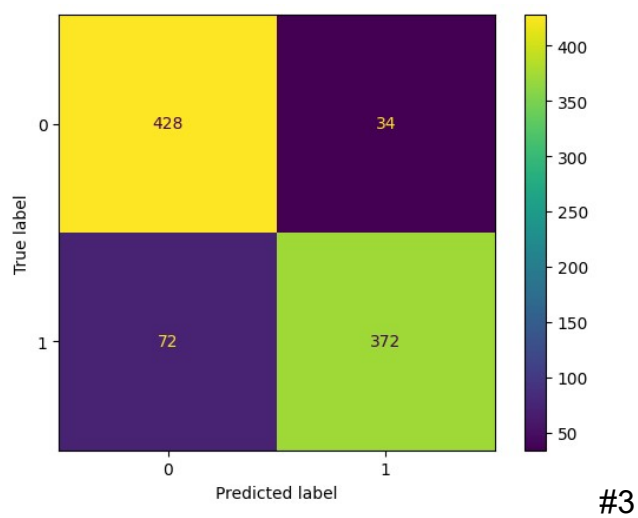
87 **Figure S7.** The confusion matrix for the ET classification models using resampling
88 method and 5-fold cross-validation of the Bursi mutagenicity training dataset for
89 the runs #1, #2, and #3.

90

91



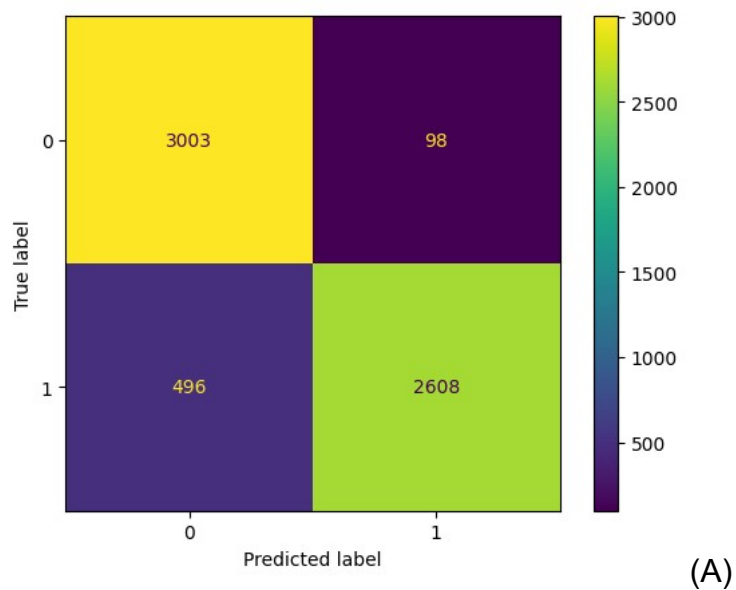
93



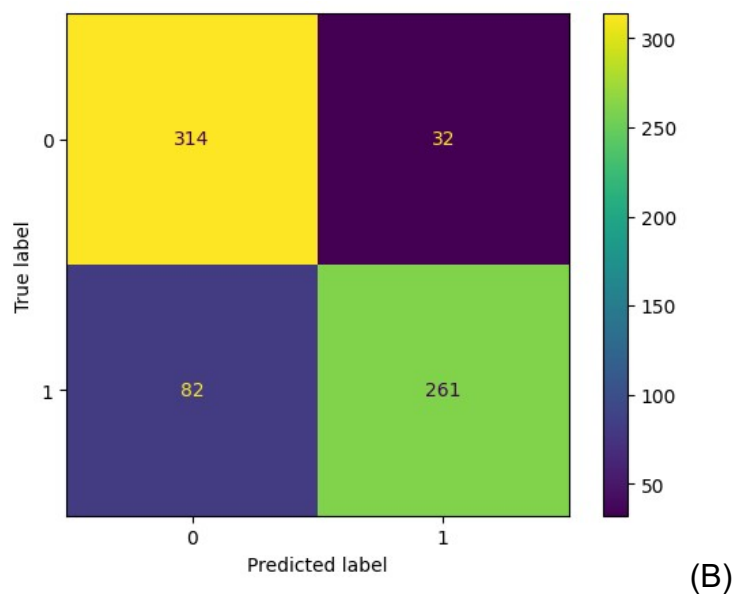
94 **Figure S8.** The confusion matrix for the ET classification models using resampling
95 method and 5-fold cross-validation of the Bursi mutagenicity validation dataset for
96 the runs #1, #2, and #3.

97

98



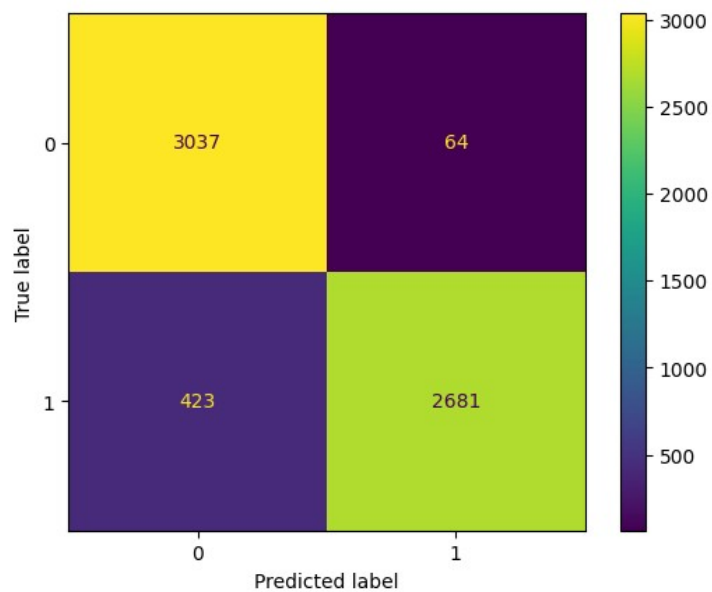
99



100

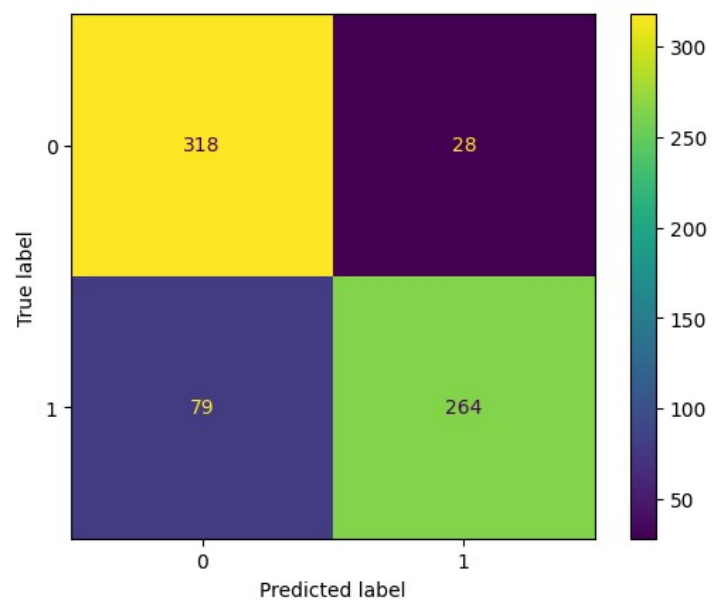
101 **Figure S9.** The confusion matrix for the RF classification models using resampling
102 method and 10x5-nested cross-validation of the Hansen mutagenicity training (A)
103 and validation (B) datasets.

104



105

(A)



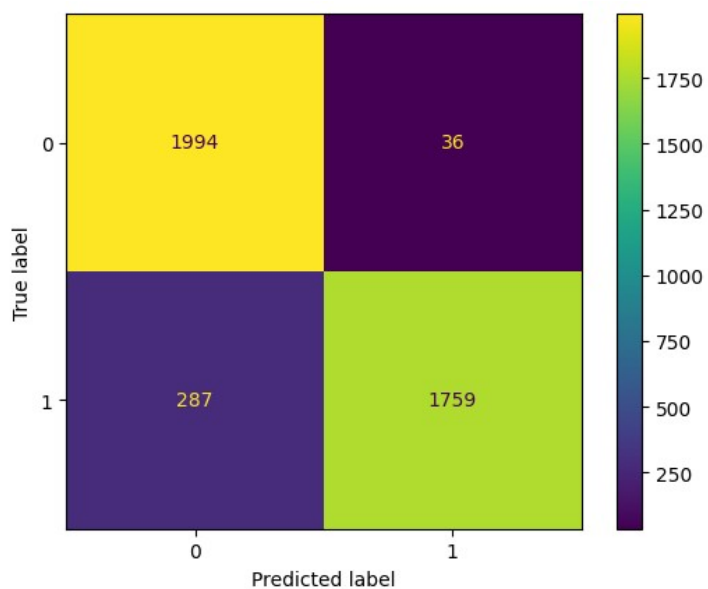
106

(B)

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

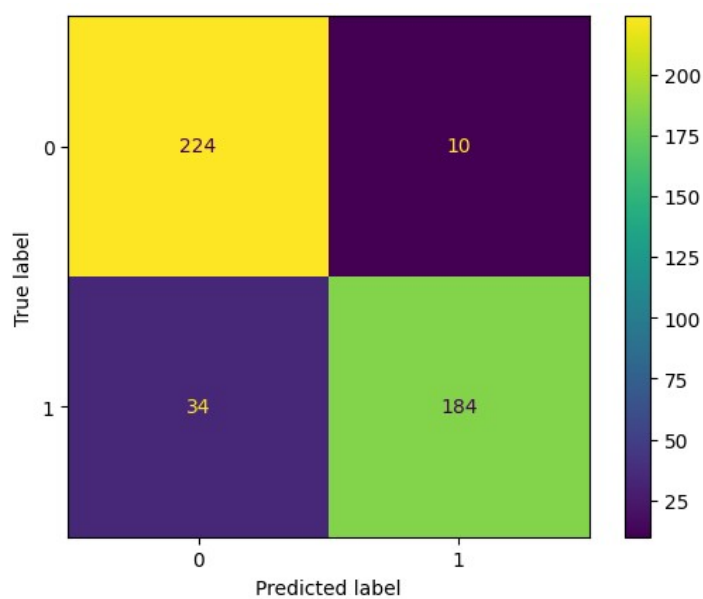
110

111



112

(A)



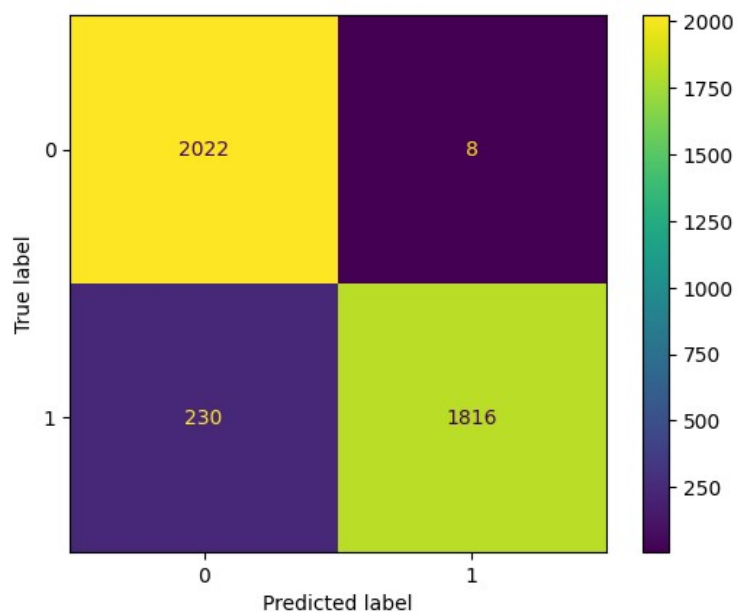
113

(B)

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

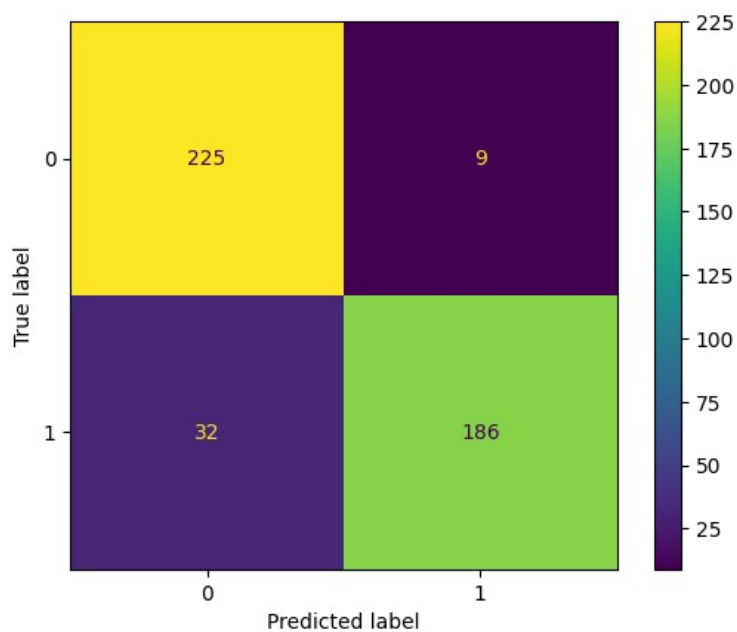
117

118



119

(A)

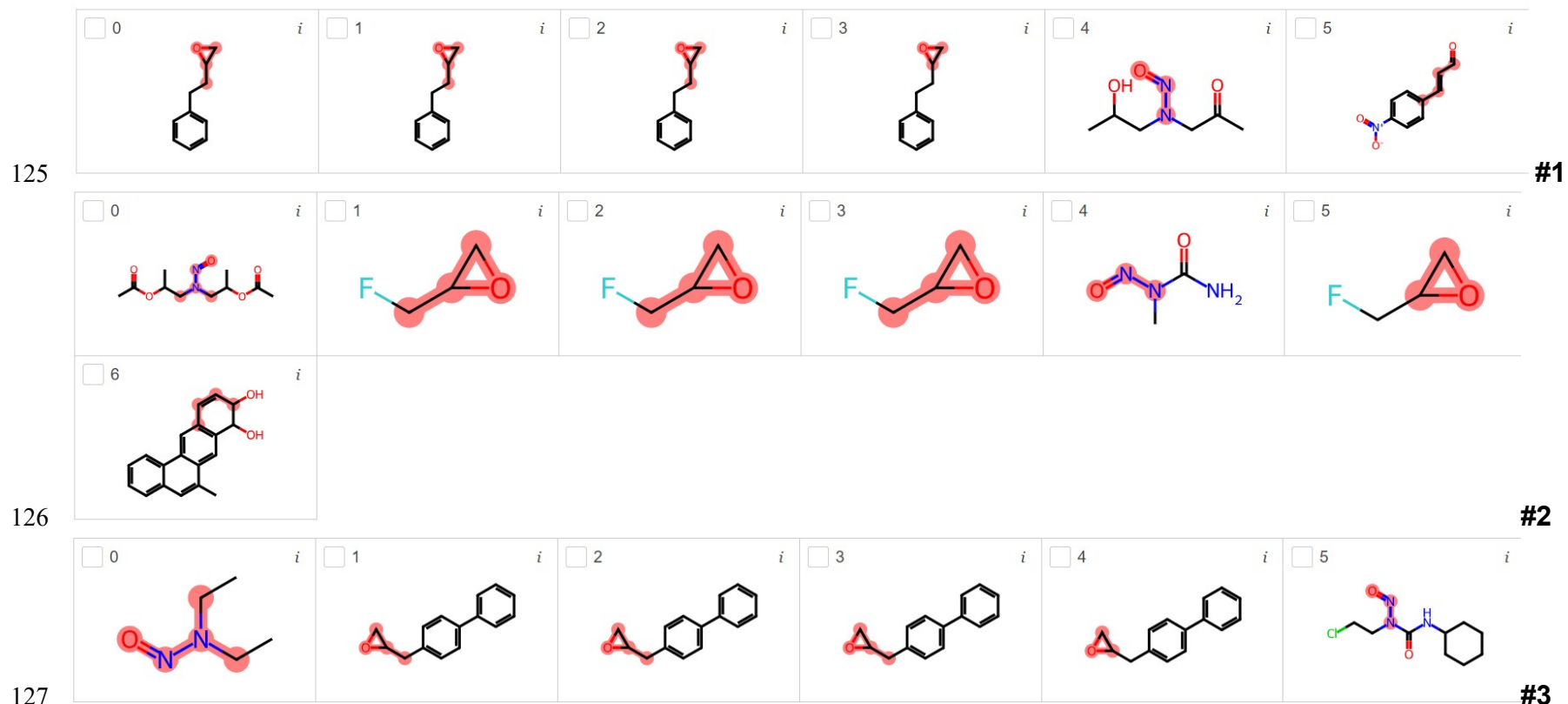


120

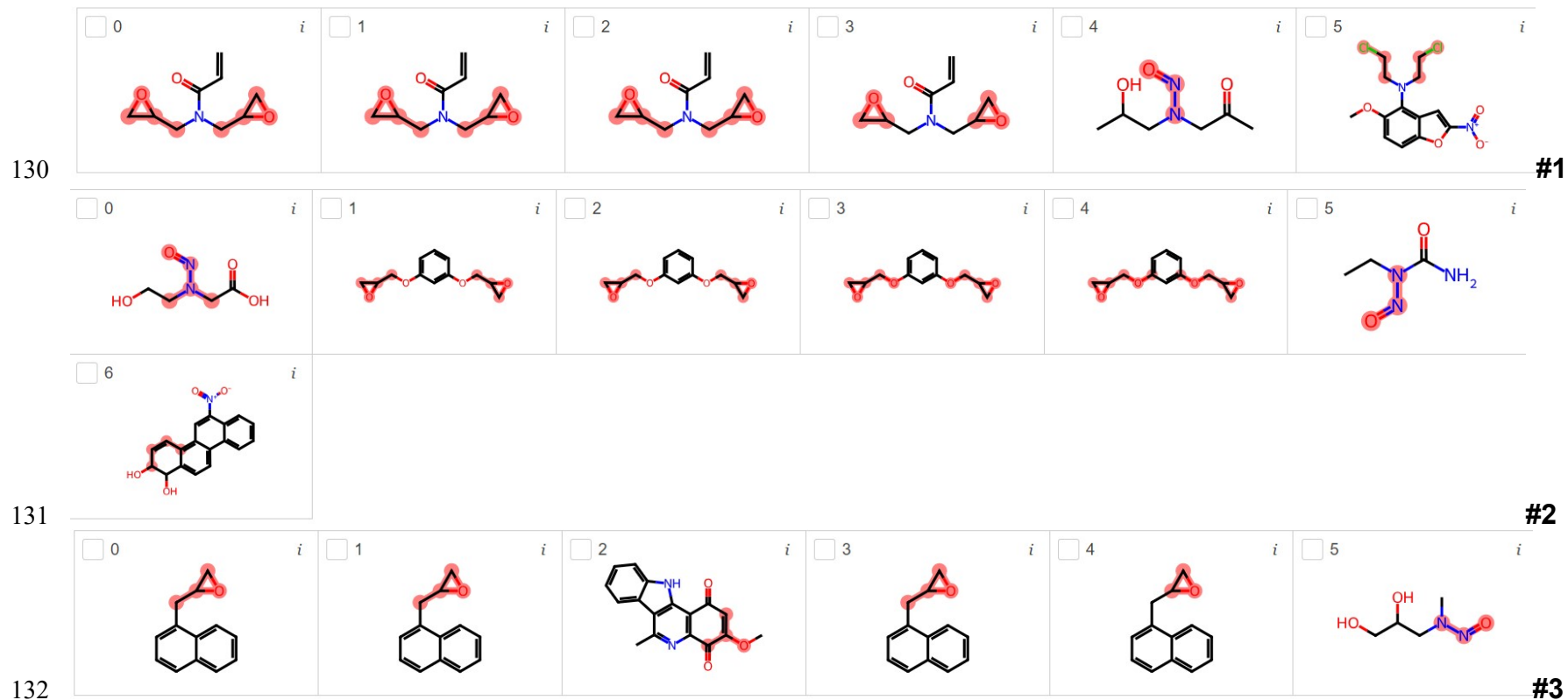
(B)

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

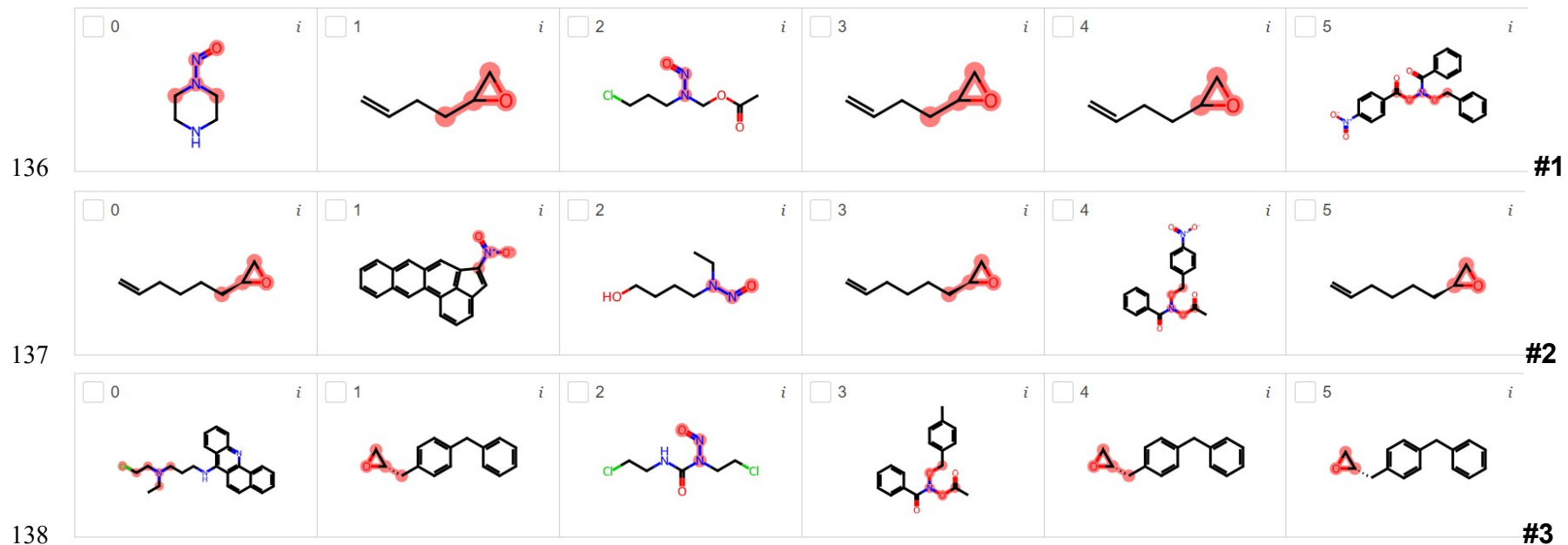
124



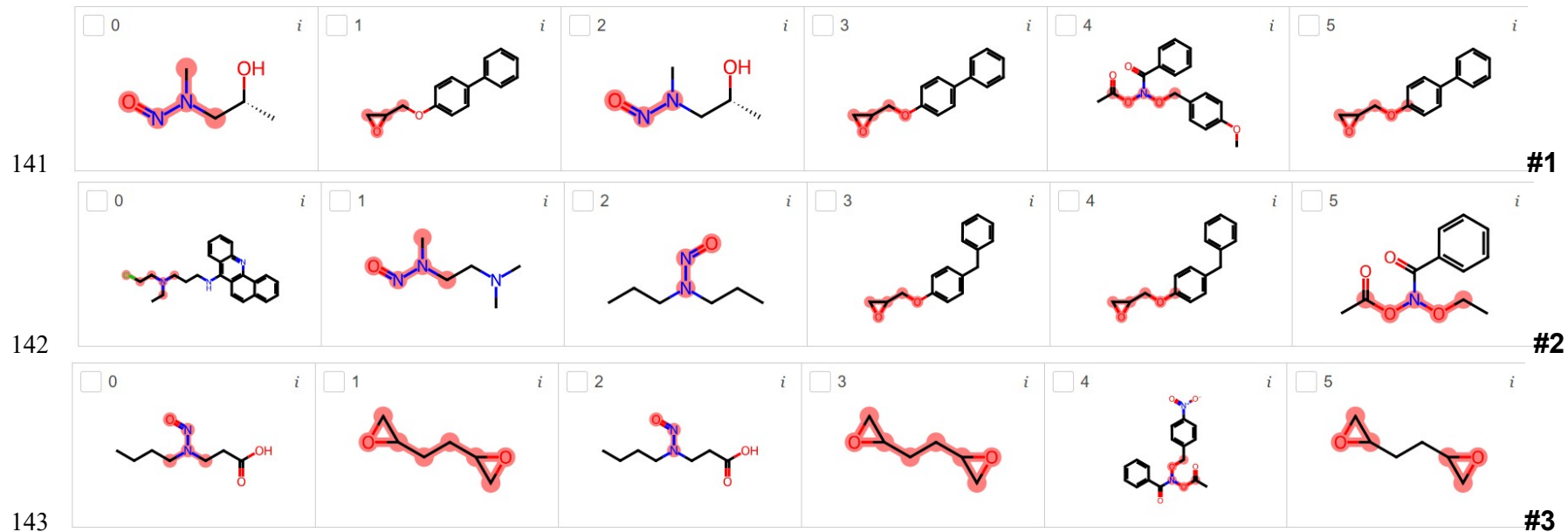
128 **Figure S13.** The most important substructures found by LIME for Bursi mutagenicity dataset using resampling method, 5-fold cross-
 129 validation, and the RF classification model for runs #1, #2, and #3.



133 **Figure S14.** The most important substructures found by LIME for Bursi mutagenicity dataset using resampling method, 5-fold cross-
 134 validation, and the ET classification model for runs #1, #2, and #3.
 135



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

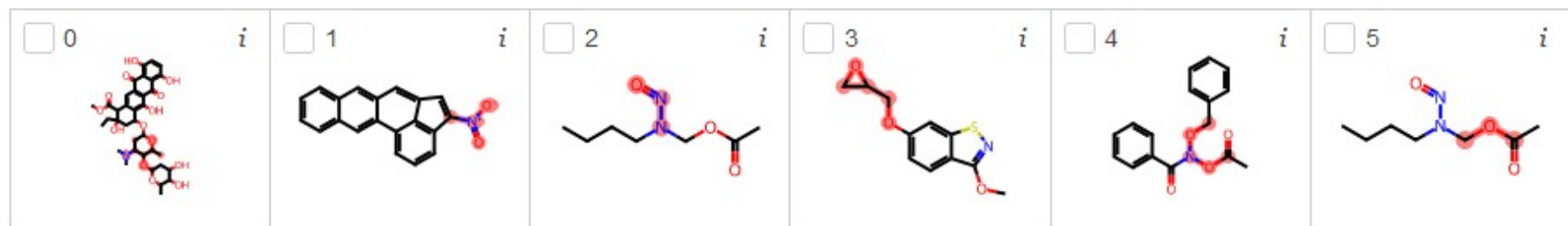


144 **Figure S16.** The most important substructures found by LIME for Hansen mutagenicity dataset using resampling method, 5-fold cross-
 145 validation, and the ET classification model for runs #1, #2, and #3.

146

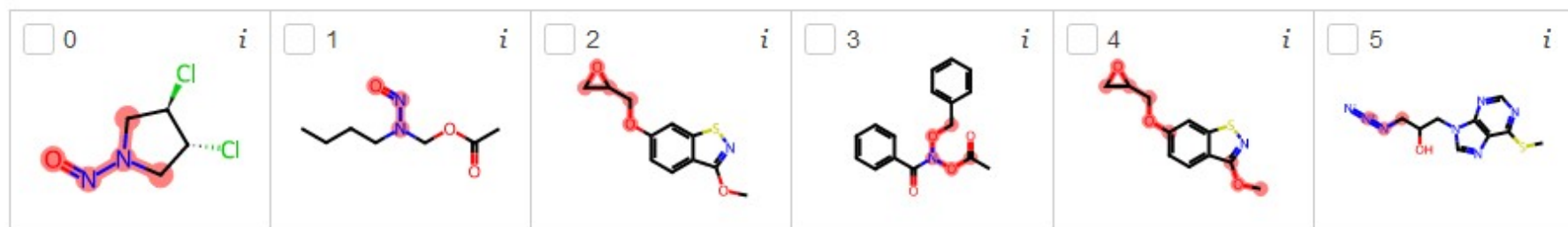
147

148
149



150

(A)



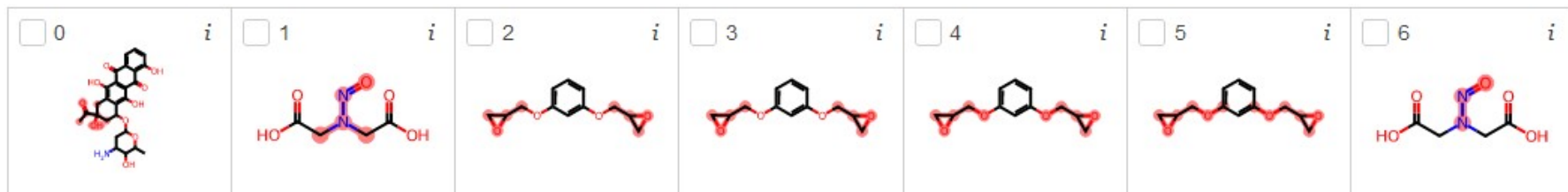
151
152

(B)

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

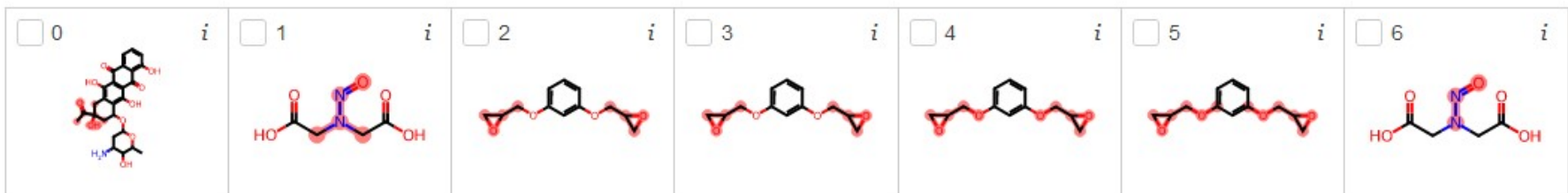
155
156

157
158



159

(A)



160

(B)

161

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

164

165