

# Supplemental: Taking a Respite from Representation Learning for Molecular Property Prediction

Jianyuan Deng<sup>1</sup>, Zhibo Yang<sup>2</sup>, Hehe Wang<sup>3</sup>, Iwao Ojima<sup>3</sup>, Dimitris Samaras<sup>2</sup>, and Fusheng Wang<sup>1,2,\*</sup>

<sup>1</sup>Stony Brook University, Department of Biomedical Informatics, Stony Brook, 11790, United States

<sup>2</sup>Stony Brook University, Department of Computer Science, Stony Brook, 11790, United States

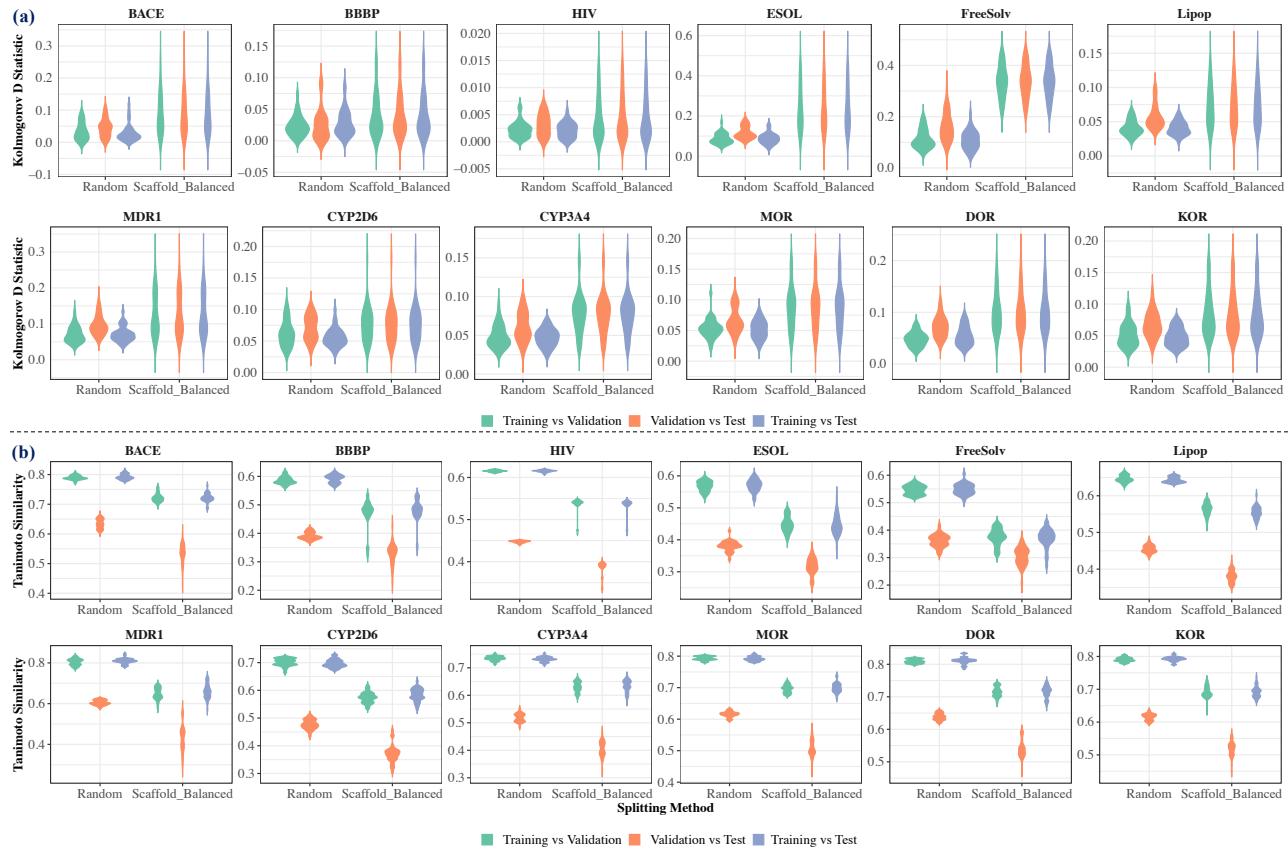
<sup>3</sup>Stony Brook University, Department of Chemistry, Stony Brook, 11790, United States

Dataset		BACE			BBBP			HIV		
Model	RF	MOLBERT	GROVER <sup>1</sup>	RF	MOLBERT	GROVER <sup>1</sup>	RF	MOLBERT	GROVER <sup>1</sup>	
AUROC	28	2	0	12	9	9	26	4	0	
AUPRC	26	3	1	8	13	9	30	0	0	
PPV	16	6	8	6	12	12	28	2	0	
NPV	20	7	3	17	8	5	13	17	0	
Dataset		ESOL			FreeSolv			Lipop		
Model	RF	MOLBERT	GROVER <sup>1</sup>	RF	MOLBERT	GROVER <sup>1</sup>	RF	MOLBERT	GROVER <sup>1</sup>	
RMSE	0	0	30	2	0	28	1	27	2	
MAE	0	0	30	3	0	27	2	25	3	
R2	0	0	30	2	0	28	1	27	2	
PEARSON_R	0	0	30	2	0	28	1	21	8	

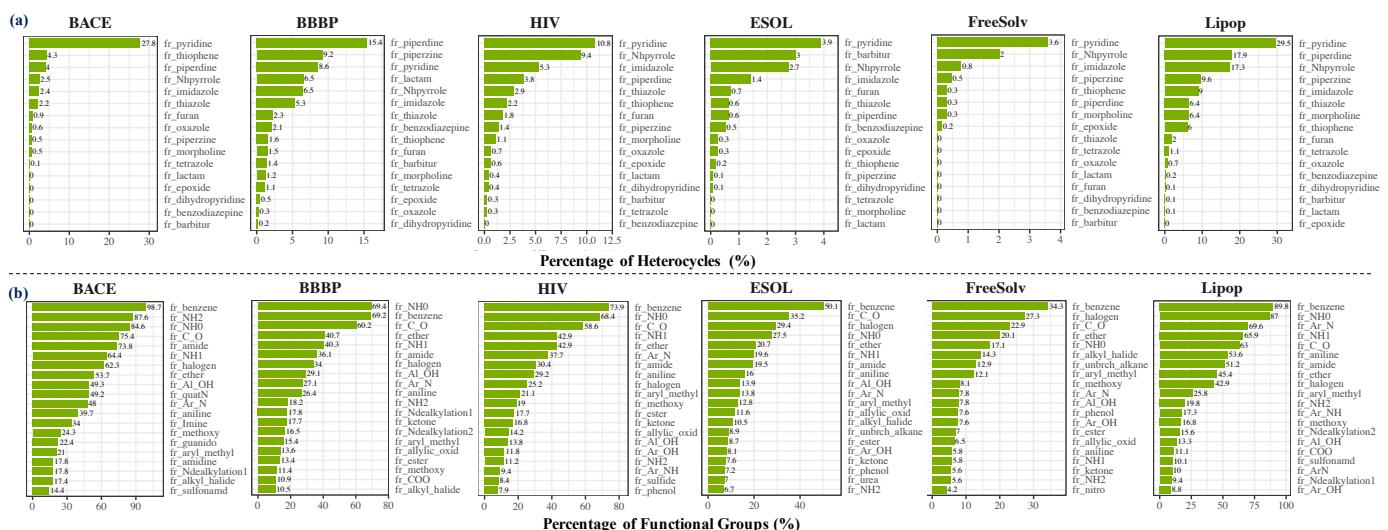
**Table 1.** Number of Single Fold where a Model Achieves the Best Performance using under Scaffold Split.

Dataset		BACE			BBBP			HIV		
Model	RF	MOLBERT	GROVER <sup>1</sup>	RF	MOLBERT	GROVER <sup>1</sup>	RF	MOLBERT	GROVER <sup>1</sup>	
AUROC	3,924	136	0	2,105	1,013	942	3,711	349	0	
AUPRC	3,797	206	57	1,370	1,934	756	4,060	0	0	
PPV	2,473	559	1,028	632	1,725	1,703	4,054	6	0	
NPV	3,504	373	183	2,980	653	427	1,857	2,203	0	
Dataset		ESOL			FreeSolv			Lipop		
Model	RF	MOLBERT	GROVER <sup>1</sup>	RF	MOLBERT	GROVER <sup>1</sup>	RF	MOLBERT	GROVER <sup>1</sup>	
RMSE	0	0	4,060	0	0	4,060	397	3,597	66	
MAE	0	0	4,060	14	0	4,046	840	3,096	124	
R2	0	0	4,060	17	0	4,043	387	3,673	0	
PEARSON_R	0	0	4,060	8	0	4,052	333	3,194	533	

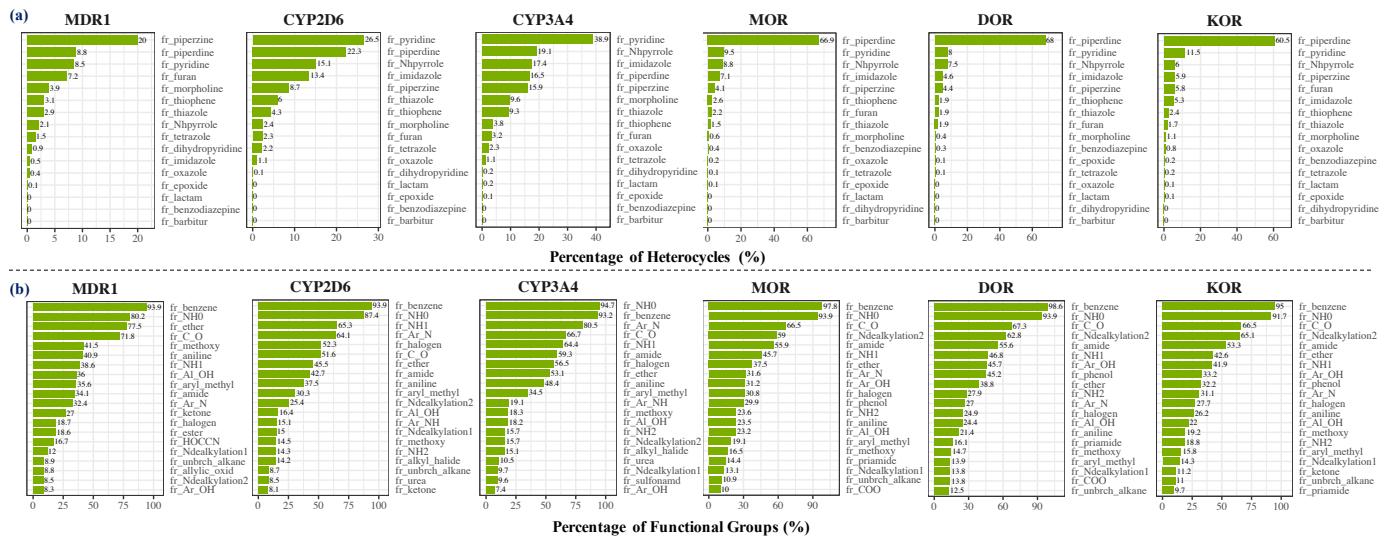
**Table 2.** Number of 3-Fold Combinations where a Model Achieves the Best Performance under Scaffold Split.



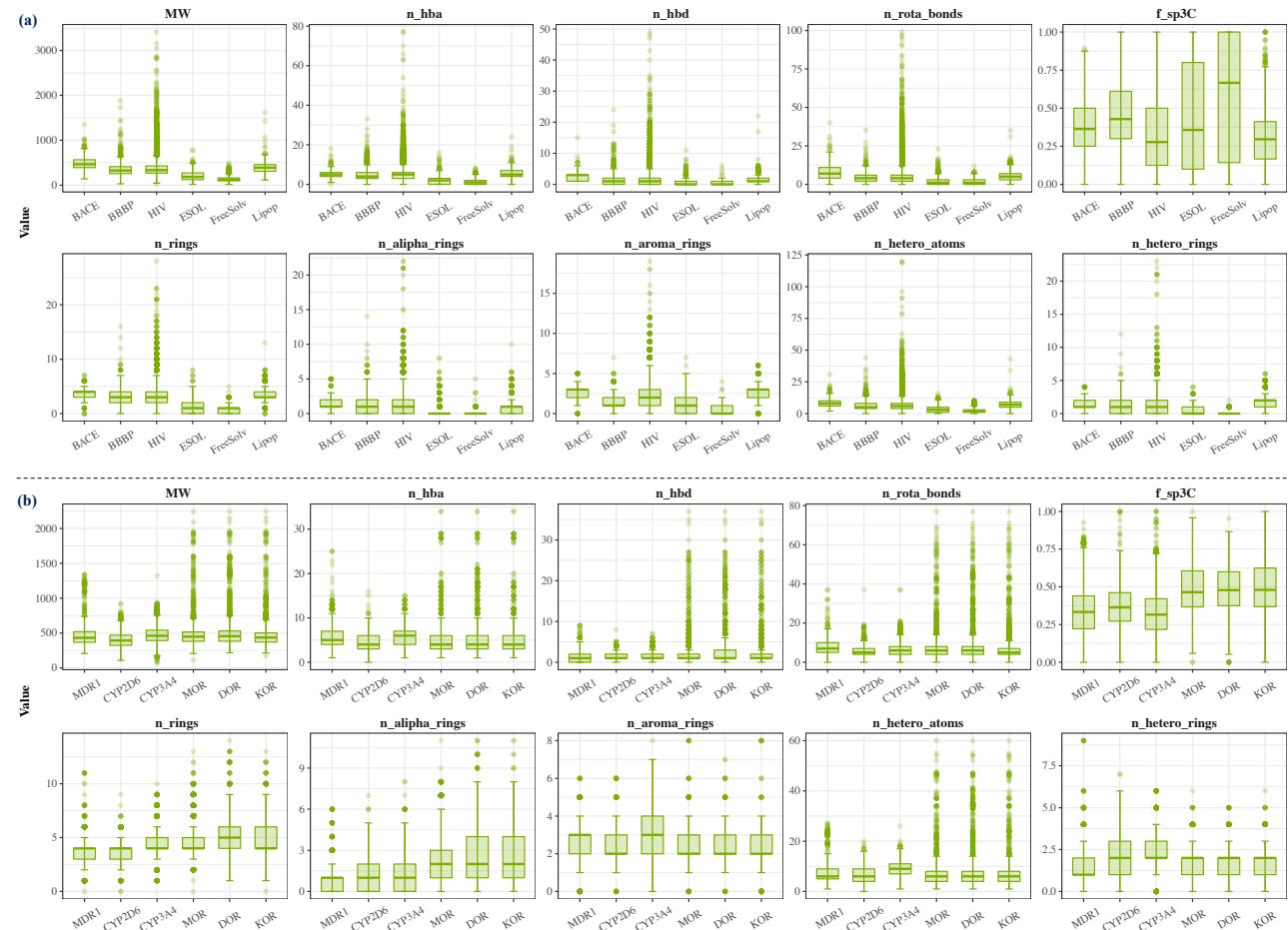
**Figure 1.** Distribution of the Kolmogorov  $D$  Statistic (a) and Tanimoto Similarity (b) among Training, Validation and Test Sets.



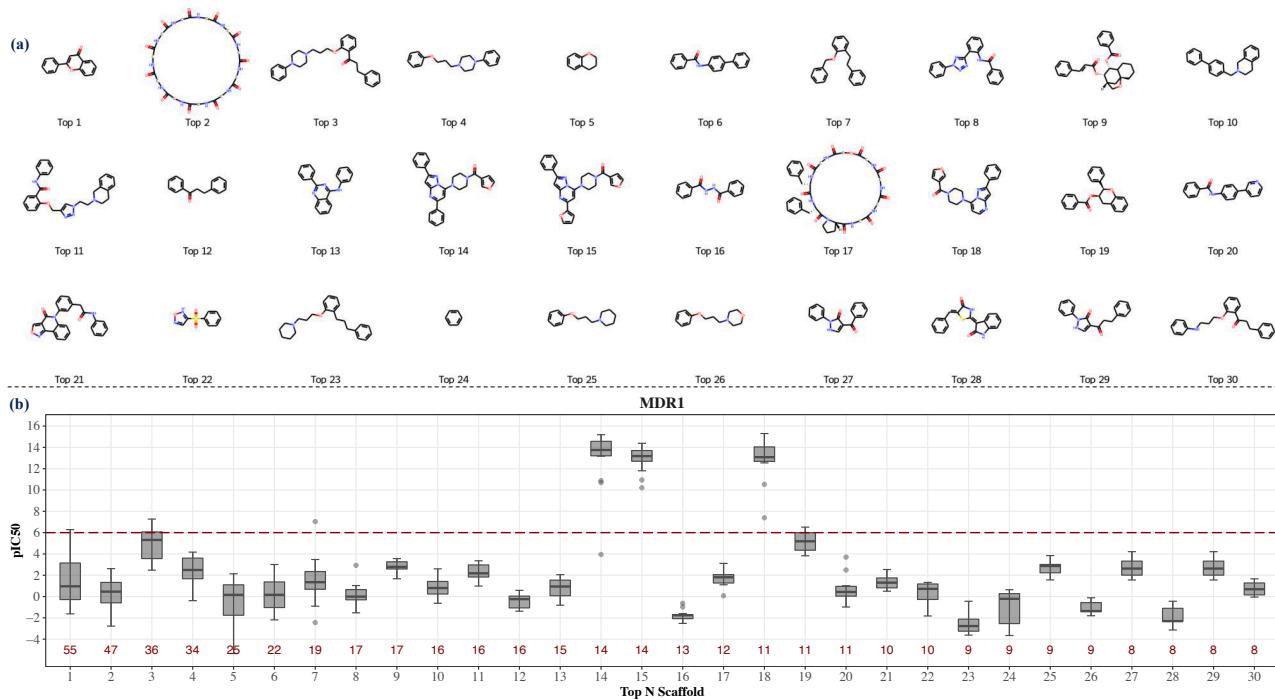
**Figure 2.** Top Fragments Prevalence for the Benchmark Datasets. (a). Top heterocycles (b). Top functional groups.



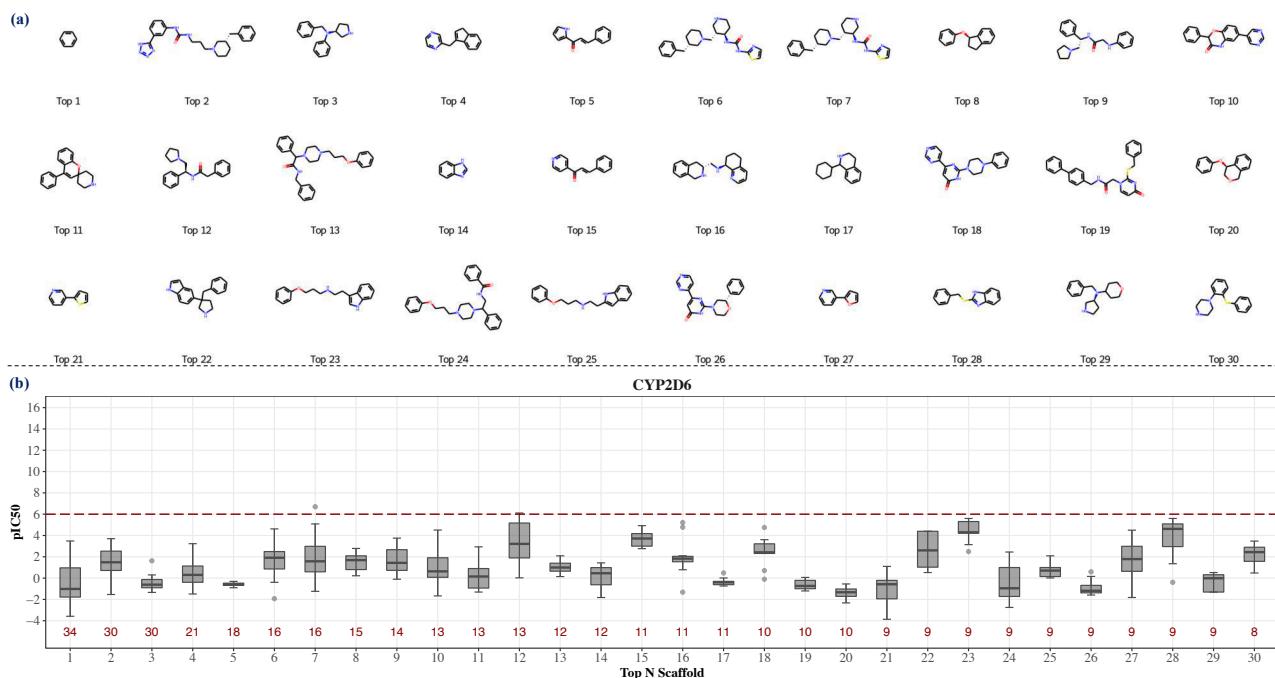
**Figure 3.** Top Fragments Prevalence for the Opioids-related Datasets. (a). Top heterocycles (b). Top functional groups.



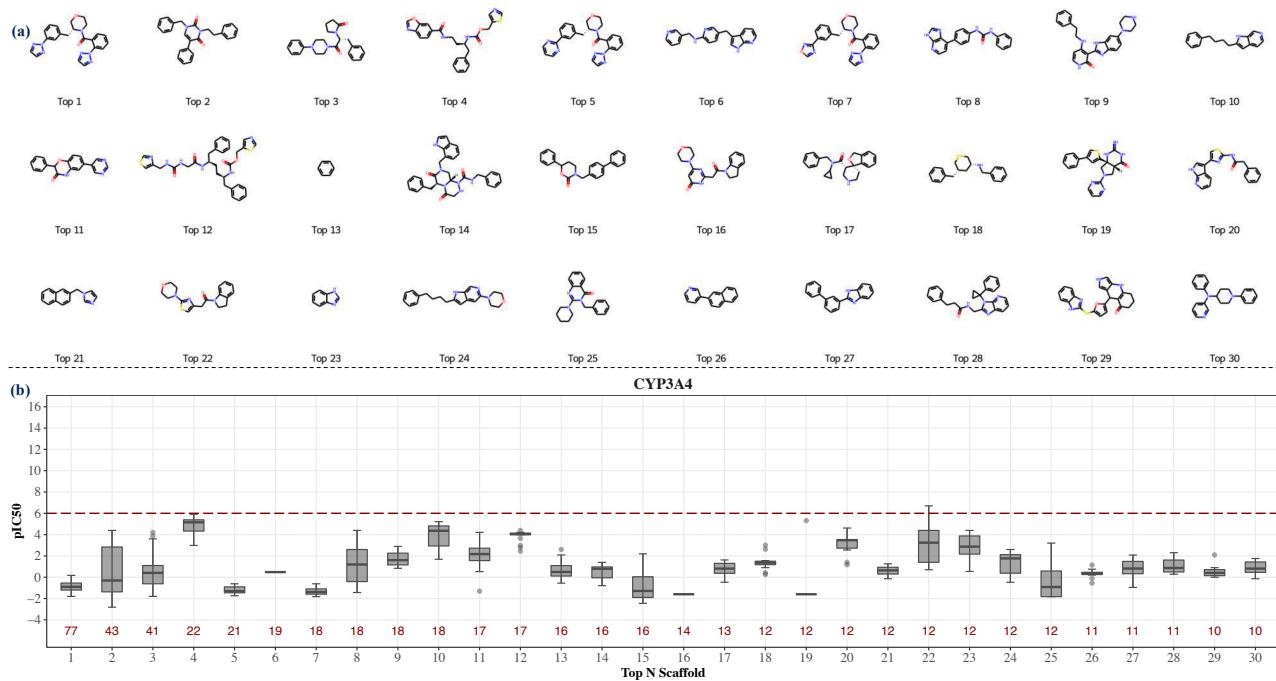
**Figure 4.** Distribution of Other Structural Traits for the Benchmark Datasets (a) and Opioids-related Datasets (b).



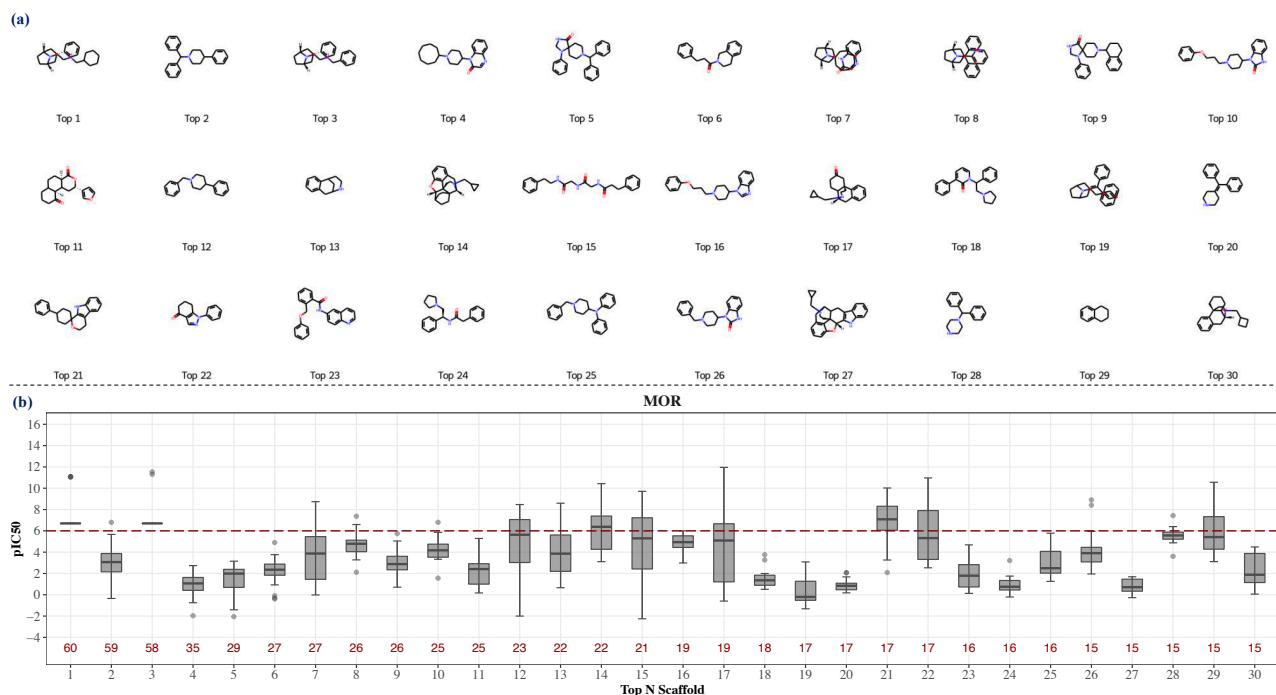
**Figure 5.** Scaffolds and Label Distribution in the MDR1 Dataset. (a). Top 30 scaffolds in the MDR1 dataset. (b). pIC50 distribution for molecules with the top30 scaffolds (red number shows how many molecules are equipped with the scaffold).



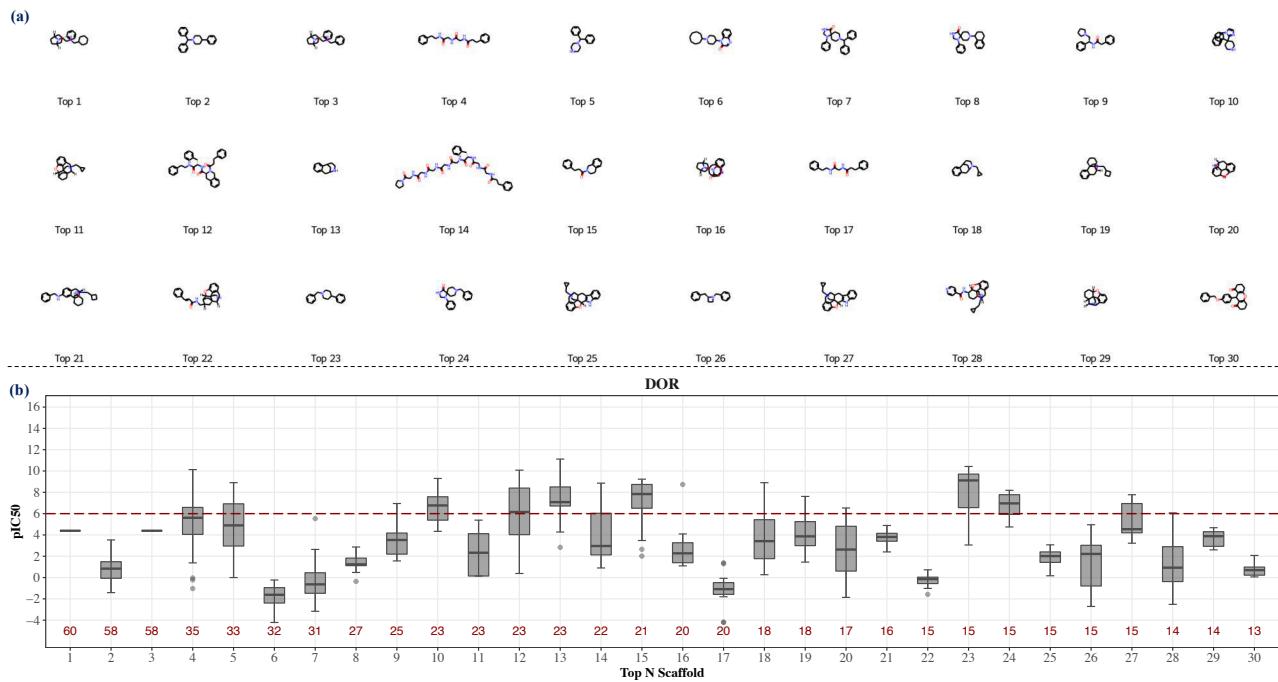
**Figure 6.** Scaffolds and Label Distribution in the CYP2D6 Dataset. (a). Top 30 scaffolds in the CYP2D6 dataset. (b). pIC50 distribution for molecules with the top30 scaffolds (red number shows how many molecules are equipped with the scaffold).



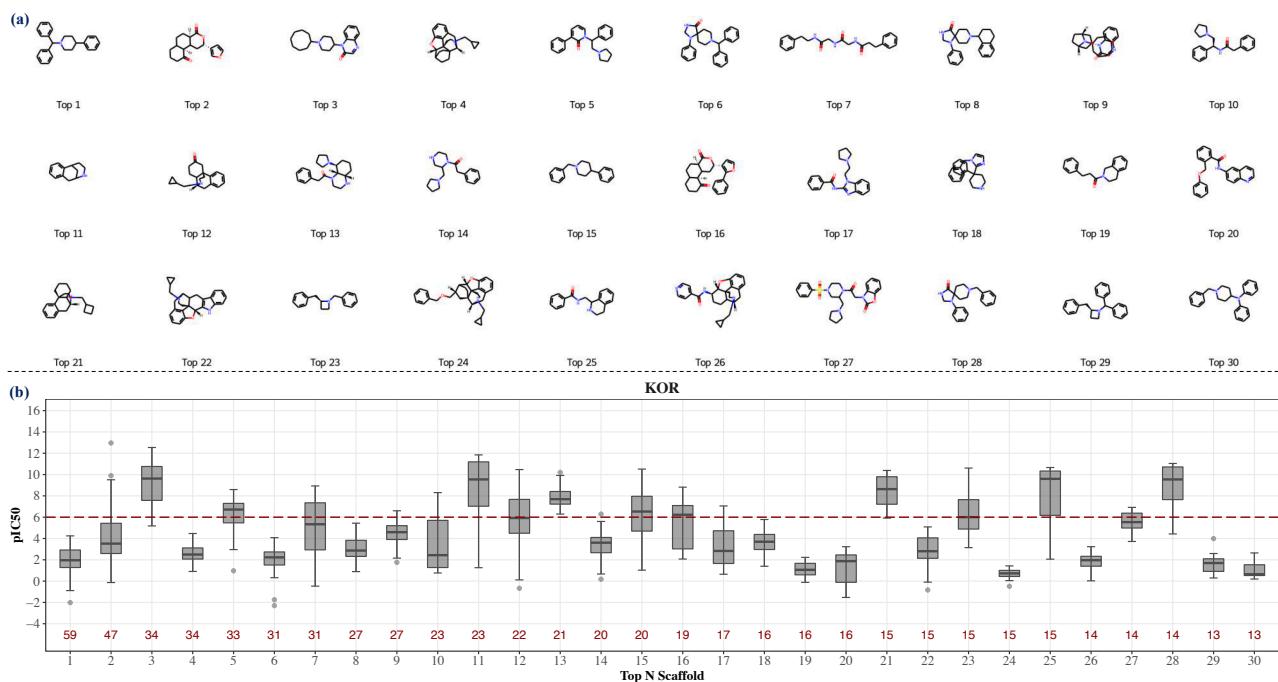
**Figure 7.** Scaffolds and Label Distribution in the CYP3A4 Dataset. (a). Top 30 scaffolds in the CYP3A4 dataset. (b). pIC50 distribution for molecules with the top30 scaffolds (red number shows how many molecules are equipped with the scaffold).



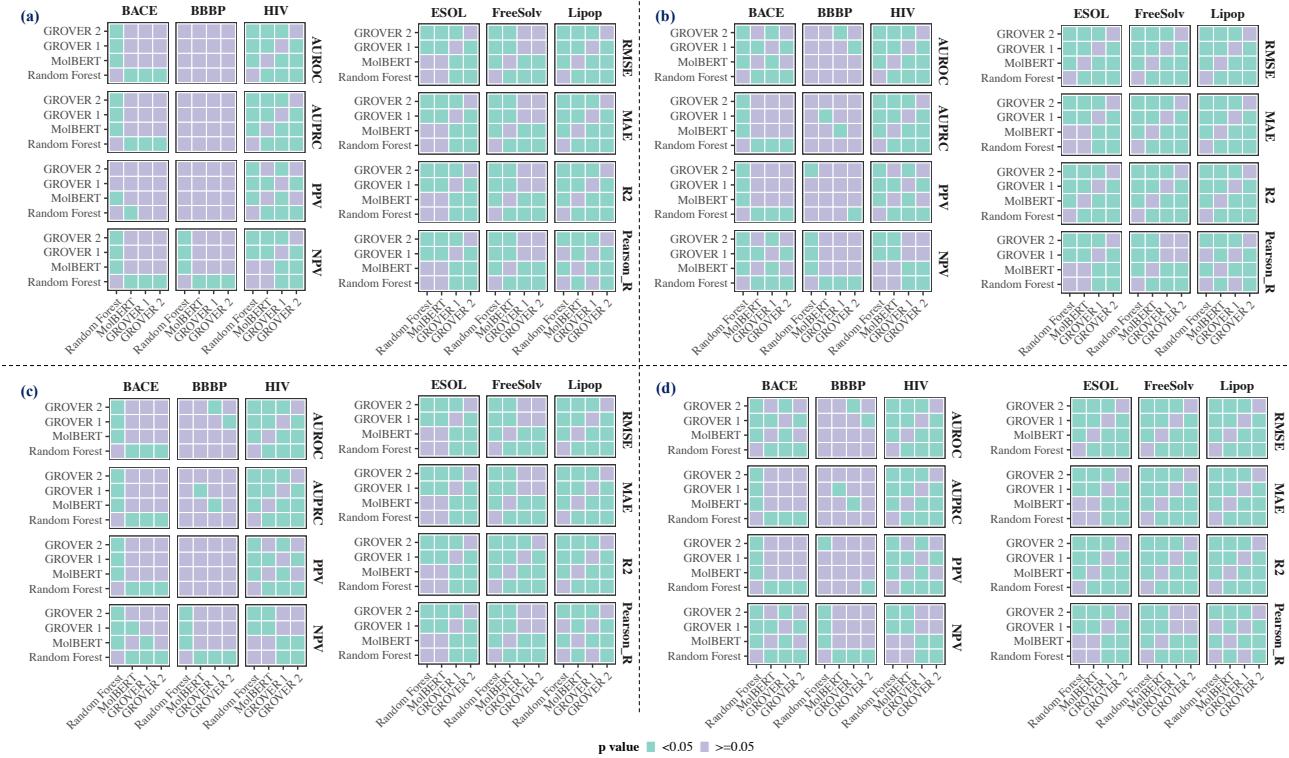
**Figure 8.** Scaffolds and Label Distribution in the MOR Dataset. (a). Top 30 scaffolds in the MOR dataset. (b). pIC50 distribution for molecules with the top30 scaffolds (red number shows how many molecules are equipped with the scaffold).



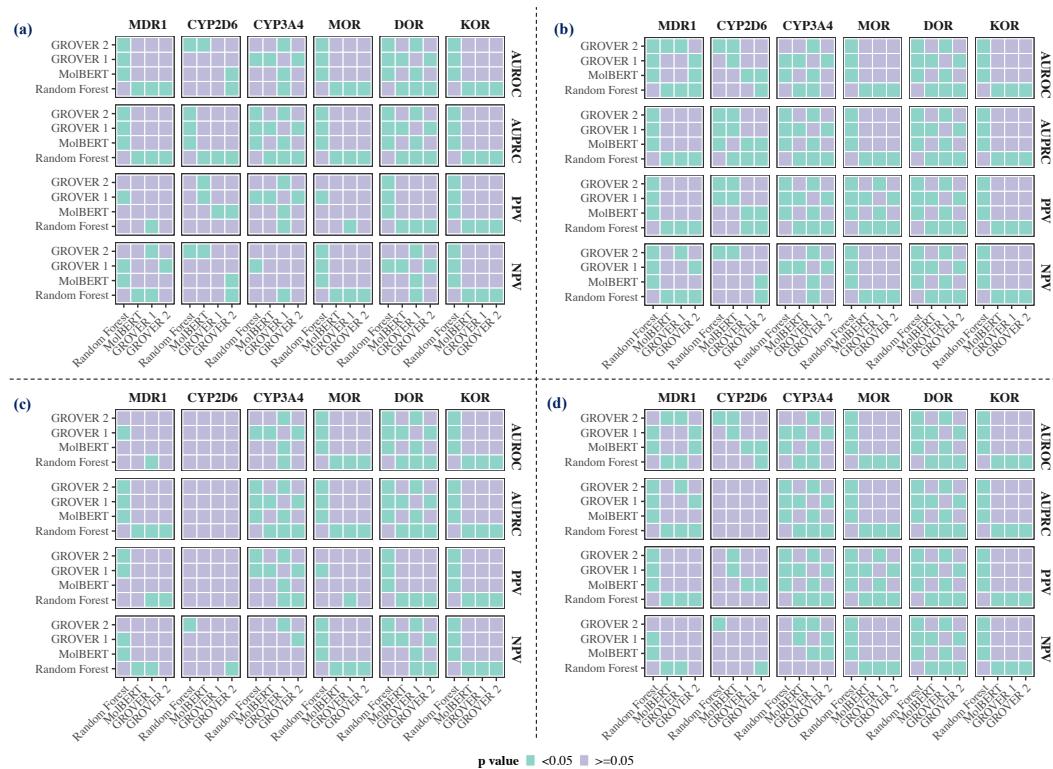
**Figure 9.** Scaffolds and Label Distribution in the DOR Dataset. (a). Top 30 scaffolds in the DOR dataset. (b). pIC50 distribution for molecules with the top30 scaffolds (red number shows how many molecules are equipped with the scaffold).



**Figure 10.** Scaffolds and Label Distribution in the KOR Dataset. (a). Top 30 scaffolds in the KOR dataset. (b). pIC50 distribution for molecules with the top30 scaffolds (red number shows how many molecules are equipped with the scaffold).



**Figure 11.** Pairwise Statistical Significance in Benchmark Prediction Performance Comparison under Scaffold Split with Different Tests. (a). Mann-Whitney  $U$  test (b). Wilcoxon signed-rank test (c). Unpaired  $t$  test (d). Paired  $t$  test



**Figure 12.** Pairwise Statistical Significance in Opioids Prediction Performance Comparison under Scaffold Split with Different Tests. (a). Mann-Whitney  $U$  test (b). Wilcoxon signed-rank test (c). Unpaired  $t$  test (d). Paired  $t$  test