Table S16: Results of statistical significance tests on the 5 CYP450 datasets with balanced scaffold split, where numbers represent the p-values (one-sided significance level) of the McNemar’ s test between the models in the table (CHEM-BERT, GROVER, MolCLRGIN, MolCLRGCN, RNN_LR, RNN_MLP, RNN_RF, TRFM_LR, TRFM_MLP and TRFM_RF) and ImageMol. The numbers in green background indicate statistically different models, using a significance threshold of 0.05.

	CHEM-BERT	GROVER	MolCLRGIN	MolCLRGCN	RNN_LR
CYP1A2	3.86E-57	0.003	0.151	0.001	2.29E-16
CYP2C9	8.63E-18	0.001	0.203	0.007	9.08E-09
CYP2C19	5.35E-34	0.002	0.056	0.791	8.32E-11
CYP2D6	0.001	0.065	0.135	0.736	0.002
CYP3A4	1.31E-27	0.019	0.069	0.001	6.14E-09
——— continue ———					
	RNN_MLP	RNN_RF	TRFM_LR	TRFM_MLP	TRFM_RF
CYP1A2	2.55E-29	9.14E-07	5.50E-05	8.58E-08	5.53E-06
CYP2C9	2.39E-19	3.24E-06	5.38E-08	8.24E-07	5.87E-05
CYP2C19	4.70E-19	1.91E-09	1.25E-06	4.45E-08	7.07E-05
CYP2D6	9.18E-05	0.043	0.021	2.80E-05	0.07
CYP3A4	9.02E-12	3.73E-09	0.001	7.39E-08	0.0002