Table S11: The performance of different methods on multi-labeled CYP450 dataset with balanced scaffold split, which evaluates the performance of the model on 5 tasks (CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4) simultaneously. SS represents the p-values (one-sided significance level) of the McNemar's test between ImageMol and other comparison methods. The numbers in green background indicate statistically different models, using a significance threshold of 0.05. 0 indicates statistical significance less than E-100.

multi-labeled CYP450								
	accuracy	**auc**	**aupr**	**f1**	**precision**	**recall**	**kappa**	**SS**
$MolCLR_{GIN}$	0.810±0.002	0.861±0.003	0.732±0.005	0.664±0.012	**0.725±0.018**	0.625±0.027	0.528±0.008	7.98E-05
$MolCLR_{GCN}$	0.806±0.002	0.843±0.003	0.712±0.001	0.657±0.007	0.730±0.007	0.608±0.015	0.519±0.007	2.30E-05
RNN_LR	0.737±0.000	0.702±0.000	0.473±0.000	0.434±0.001	0.492±0.001	0.396±0.001	0.259±0.000	0
TRFM_LR	0.783±0.000	0.775±0.000	0.595±0.000	0.544±0.000	0.607±0.000	0.502±0.000	0.399±0.000	2.16E-49
RNN_MLP	0.735±0.001	0.707±0.000	0.470±0.000	0.385±0.001	0.500±0.003	0.336±0.002	0.220±0.001	0
TRFM_MLP	0.711±0.016	0.769±0.001	0.553±0.002	0.470±0.005	0.317±0.004	**0.935±0.009**	0.204±0.010	0
RNN_RF	0.784±0.002	0.794±0.001	0.586±0.001	0.523±0.002	0.609±0.005	0.521±0.001	0.382±0.003	5.42E-44
TRFM_RF	0.816±0.000	0.820±0.000	0.656±0.001	0.553±0.001	0.669±0.003	0.533±0.001	0.442±0.001	7.29E-16
CHEM-BERT	0.816±0.003	0.831±0.004	0.672±0.010	0.616±0.014	0.673±0.030	0.597±0.047	0.493±0.008	7.35E-16
ImageMol	**0.843±0.003**	**0.866±0.007**	**0.736±0.009**	**0.681±0.007**	0.714±0.020	0.661±0.025	**0.575±0.005**	-