Table S14(a): Results of statistical significance tests between ImageMol and GROVER-10M, MPG-10M, X-MOL, MolCLR$_{GIN}$ on the 8 molecular property prediction datasets with scaffold split, where numbers represent the p-values (one-sided significance level) of the McNemar's test between the models in the table (GROVER-10M, MPG-10M, X-MOL and MolCLR$_{GIN}$) and ImageMol. The numbers in green background indicate statistically different models, using a significance threshold of 0.05.

	GROVER-10M	MPG-10M	X-MOL	MolCLR$_{GIN}$
BACE	0.473	0.061	0.523	1
BBBP	0.458	0.486	0.138	0.855
ClinTox	0.267	0.332	0.421	0.851
HIV	0.516	0.091	0.463	0.001
MUV	0.606	0.081	0.373	0.239
SIDER	0.379	0.010	3.59E-11	0.0000803
Tox21	0.041	0.006	0.615	0.046
ToxCast	0.301	0.299	3.31E-07	-

—— continue ——

Table S14(b): Results of statistical significance tests between ImageMol and GROVER-10M, MPG-10M on the 6 molecular property prediction datasets with random scaffold split, where numbers represent the p-values (one-sided significance level) of the McNemar's test between the models in the table (GROVER-10M, MPG-10M) and ImageMol. The numbers in green background indicate statistically different models, using a significance threshold of 0.05. 0 indicates significance level less than E-100.

	GROVER-10M	MPG-10M
BACE	0.404	0.182
BBBP	0.752	0.011
ClinTox	4.16E-06	0.683
SIDER	0.198	0.159
Tox21	0	0.136
ToxCast	0	0