Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supporting information
Are 2D fingerprints still valuable for drug discovery?

Kaifu Gao[1], Duc Duy Nguyen[1], Vishnu Sresht[2], Alan M. Mathiowetz[2], Meihua Tu[2], and Guo-Wei Wei[1,3,4,*]

[1] Department of Mathematics, Michigan State University, MI 48824, USA.

[2] Pfizer Medicine Design, 610 Main St, Cambridge, MA 02139, USA.

[3] Department of Electrical and Computer Engineering, Michigan State University, MI 48824, USA.

[4] Department of Biochemistry and Molecular Biology, Michigan State University, MI 48824, USA.

March 6, 2020

Figure S1 is the test of significance of four toxicity predictions.

Table S1 is the p values of four toxicity predictions.

Table S2-S5 are the function group performance analysis for toxicity datasets LD_{50}, LC_{50}, and LC_{50}-DM, as well as Log P dataset following the same strategy mentioned in the main text.

Tables S6-S24 detailedly showed the performances (R_2 or R, RMSE, Kendall tall) of our Gradient bootsting decision tree models, deep learning models, multitask deep learning models on all the data sets including toxicity, binding affinity, log P, and log S.

*Corresponding to Guo-Wei Wei. Email: wei@math.msu.edu

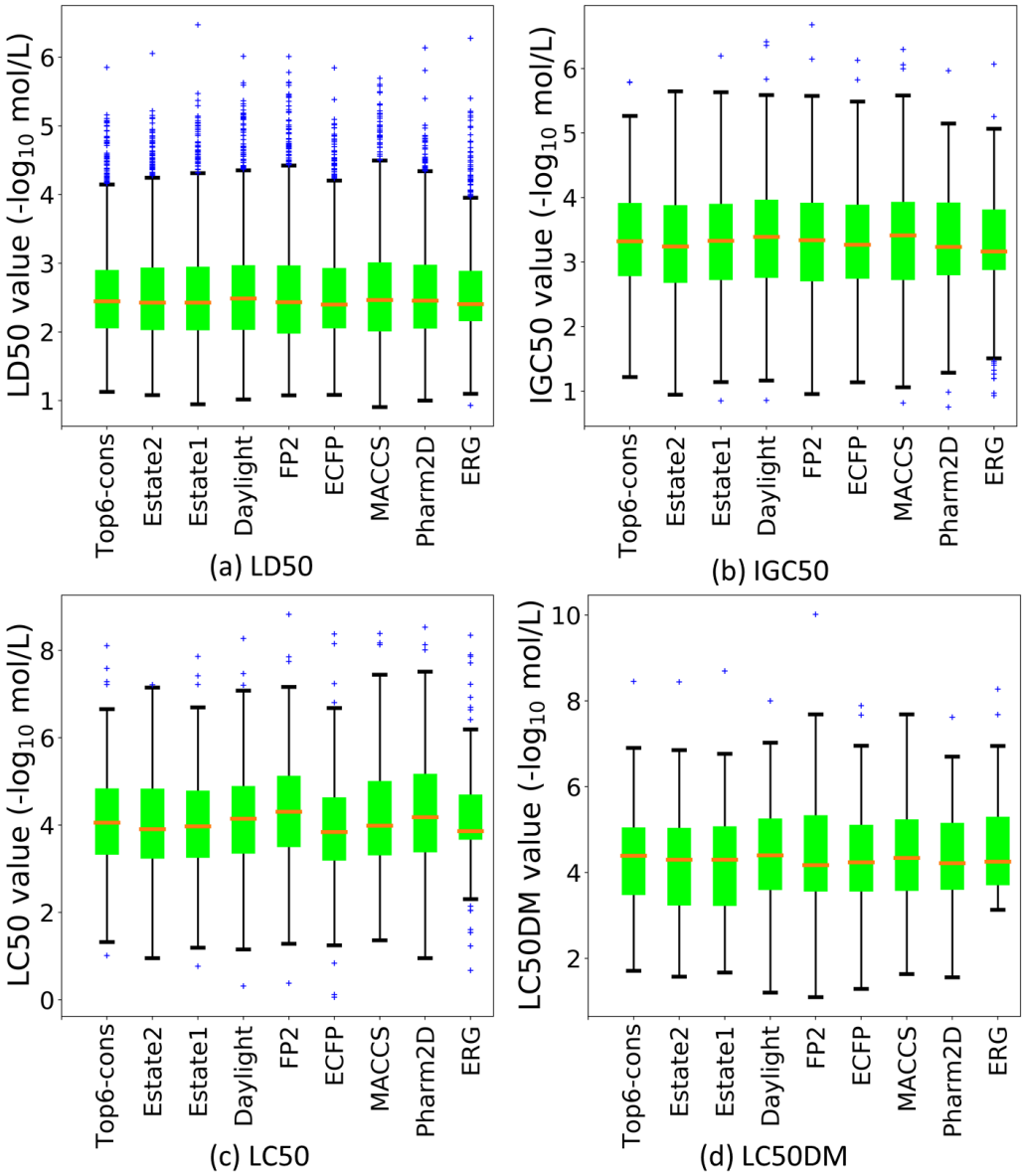
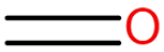
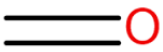
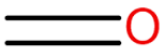
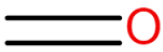
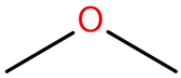
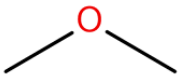
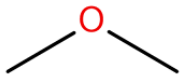
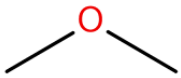
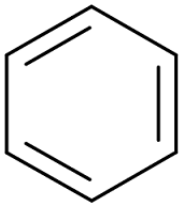
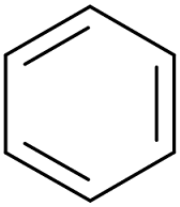
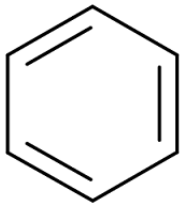
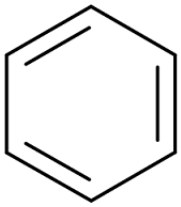
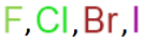
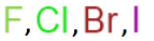
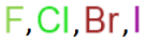
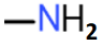
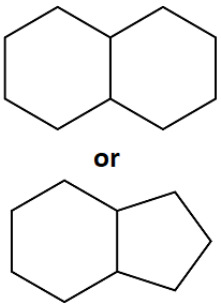
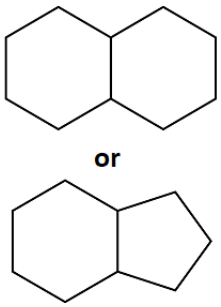
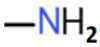
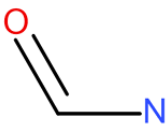
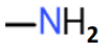
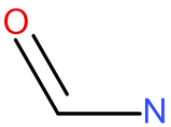


Figure S1: The distribution of predicted values by eight fingerprints and the top 6 consensus on four toxicity datasets: (a) LD50, (b) IGC50, (c) LC50, and (d) LC50DM.

	LD50	IGC50	LC50	LC50-DM
Top6-cons	0.00	6.48e-121	5.24e-46	2.02e-11
Estate2	3.35e-288	1.15e-106	4.86e-40	6.82e-12
Estate1	4.08e-301	1.34e-100	7.39e-39	1.86e-12
Daylight	4.44e-317	6.45e-93	4.08e-36	1.58e-08
Fp2	2.17e-322	2.10e-90	7.25e-35	4.72e-08
ECFP	1.95e-285	1.39e-82	9.90e-32	1.80e-10
MACCS	0.00	1.53e-81	1.00e-34	5.56e-10
Pharm2D	1.58e-190	2.50e-39	3.12e-28	3.10e-06
ERG	2.24e-162	1.46e-26	9.87e-17	1.44e-07

Table S1: The p values of predicted values by eight fingerprints and the top 6 consensus on toxicity datasets.

Ranking	FP2	Daylight	Estate1	Estate2
1	 carbonyl group: 125/271	 carbonyl group: 115/259	 carbonyl group: 141/138	 carbonyl group: 106/226
2	 ether: 116/271	 ether: 101/259	 ether: 127/338	 ether: 89/206
3	 unfused benzene ring: 86/271	 unfused benzene ring: 82/259	 unfused benzene ring: 111/338	 unfused benzene ring: 75/206
4	 hydroxyl: 76/271	 hydroxyl: 68/259	 hydroxyl: 97/338	 hydroxyl: 68/206
5	 64/271	 66/259	 89/338	 55/206
6	aliphatic chains with 8 or more members: 60/271	 amide: 49/259	 bicyclic compounds: 69/338	 bicyclic compounds: 45/206
7	 amide: 56/271	 carbonyl with Nitrogen: 46/259	 hydroxyl: 67/338	 amide: 43/206

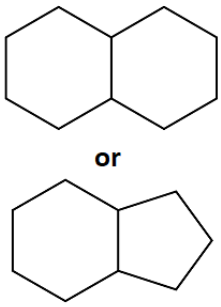
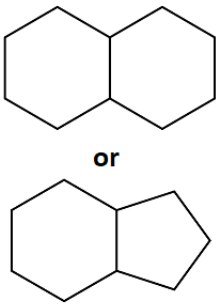
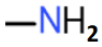
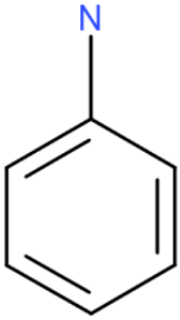
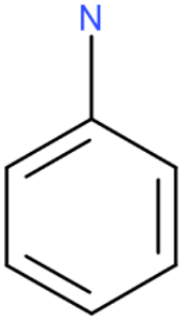
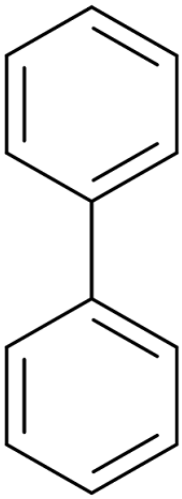
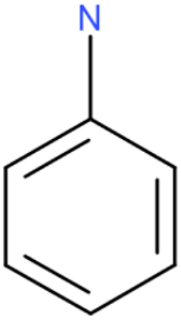
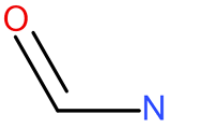
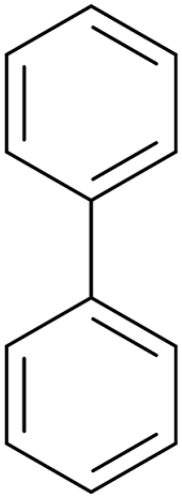
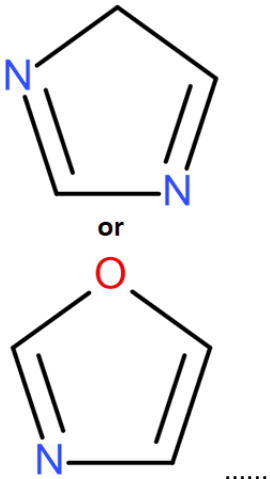
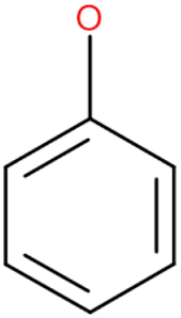
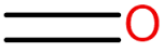
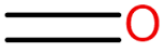
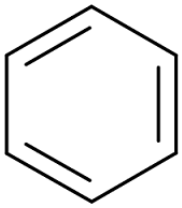
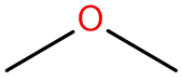
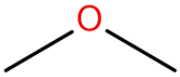
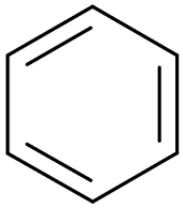
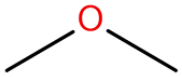
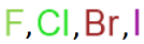
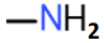
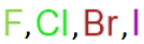
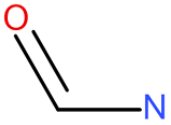
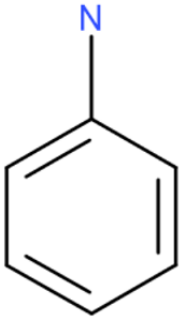
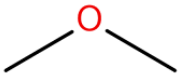
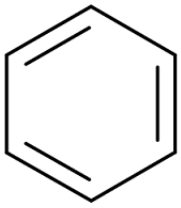
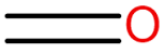
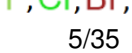
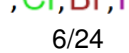
8	 ..... bicyclic compounds: 53/271	 ..... bicyclic compounds: 42/259	 carbonyl with Nitrogen: 50/338	 carboxylate amide: 39/206
9	 ..... aniline: 50/271	 37/259	 multiple non-fused benzene rings: 32/338	 azole: 38/206
10	 ..... carbonyl with Nitrogen: 43/271	 phenol: 21/259	 ..... azole: 23/338	 furan: 20/206

Table S2: The top 10 frequently occurred functional groups in LD50 toxicity set for each fingerprint. For each fingerprint, the occurrence frequency and the total number of molecules are also given.

Ranking	FP2	Estate1	Estate2	Daylight
1	 hydroxyl: 10/22	 carbonyl group: 12/35	 carbonyl group: 13/34	 hydroxyl: 11/24
2	 unfused benzene ring: 6/22	 hydroxyl: 11/35	 hydroxyl: 11/34	 carbonyl group: 11/24
3	 ether: 5/22	 ether: 8/35	 unfused benzene ring: 9/34	 ether: 9/24
4	 4/22	 amide: 7/35	 8/34	 carbonyl group with N: 7/24
5	aliphatic chains with 8 or more members: 4/22	 aniline: 5/35	 ether: 7/34	 unfused benzene ring: 7/24
6	 4/22	 5/35	aliphatic chains with 8 or more members: 6/34	 6/24

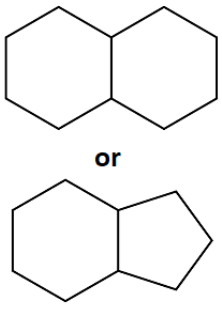
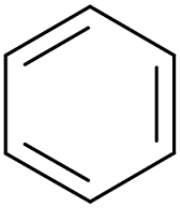
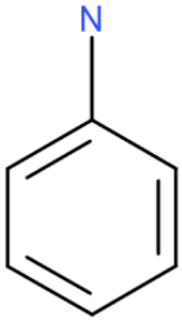
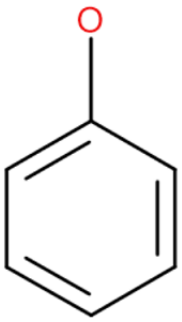
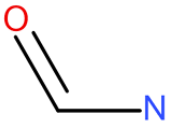
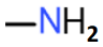
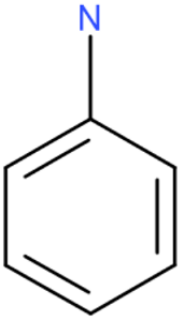
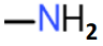
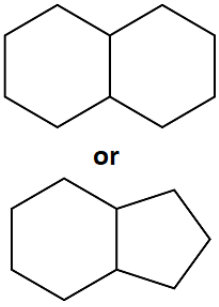
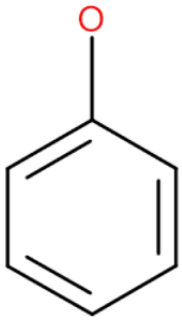
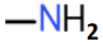
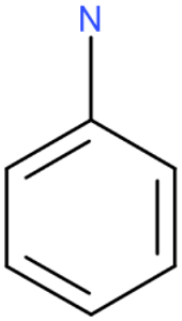
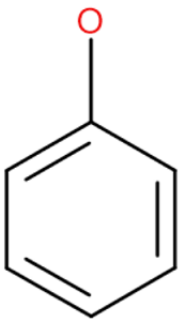
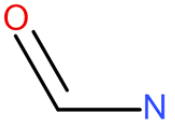
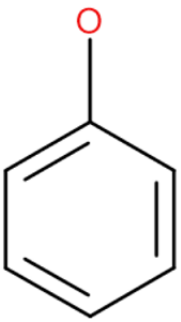
7	 or ..... bicyclic compounds: 2/22	 unfused benzene ring: 5/35	 aniline: 6/34	aliphatic chains with 8 or more members: 5/24
8	 phenol: 2/22	 carbonyl group with N: 4/35	 aniline: 6/34	 aniline: 4/24
9	 amide: 2/22	 or ..... bicyclic compounds: 4/35	 phenol: 4/34	 amide: 3/24
10	 aniline: 2/22	 phenol: 4/35	 carbonyl group with N: 4/34	 aniline: 2/24

Table S3: The top 10 frequently occurred functional groups in LC_{50} toxicity set for each fingerprint. For each fingerprint, the occurrence frequency and the total number of molecules are also given.

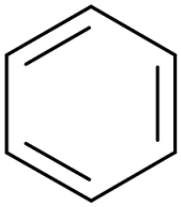
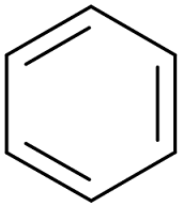
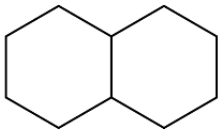
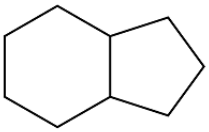
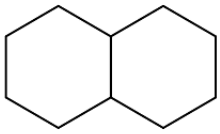
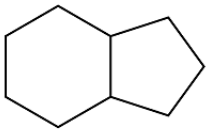
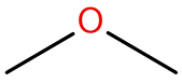
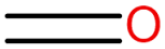
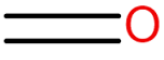
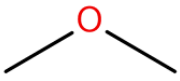
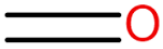
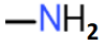
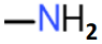
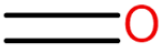
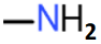
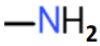
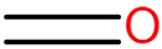
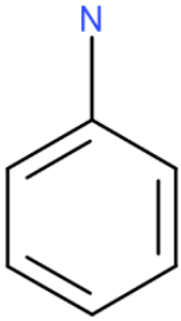
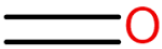
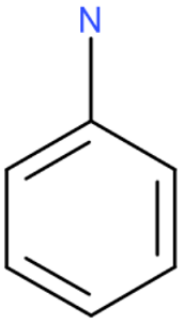
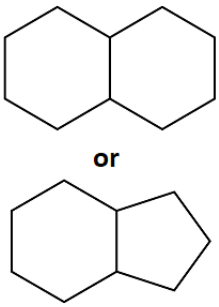
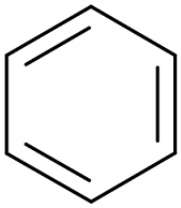
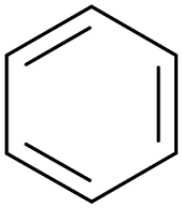
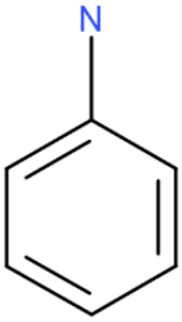
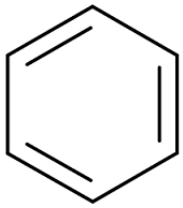
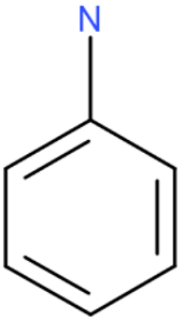
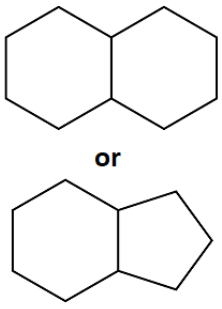
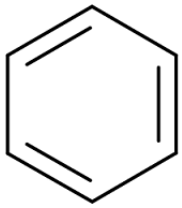
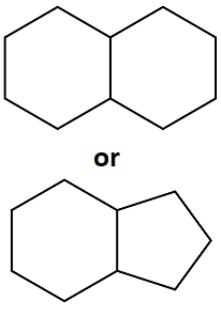
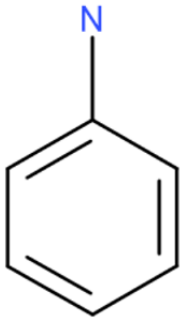
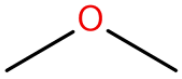
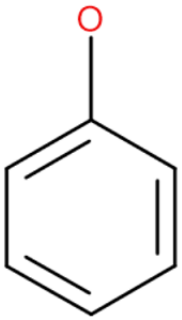
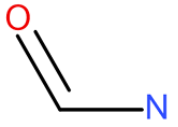
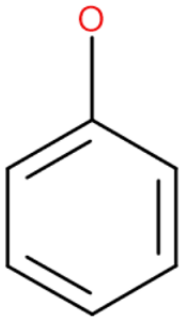
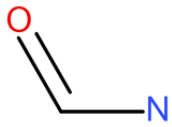
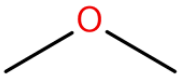
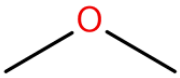
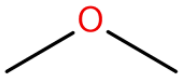
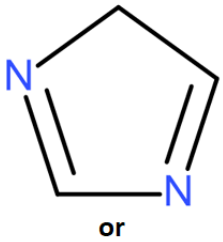
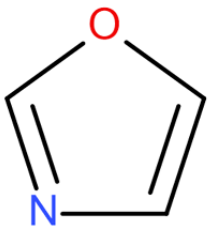
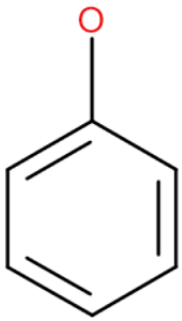
Ranking	Estate1	Estate2
1	 unfused benzene ring: 6/22	 unfused benzene ring: 6/21
2	 **or**  ..... bicyclic compounds: 6/22	 **or**  ..... bicyclic compounds: 5/21
3	 ether: 5/22	 hydroxyl: 5/22
4	 hydroxyl: 5/22	 carbonyl group: 5/21
5	 aniline: 5/22	 ether: 4/21

Table S4: The top 10 frequently occurred functional groups in LC_{50}-DM toxicity set for each fingerprint. For each fingerprint, the occurrence frequency and the total number of molecules are also given.

Ranking	ECFP	Estate1	Estate2	MACCS
1	 carbonyl group: 33/54	 hydroxyl: 25/38	 amide: 37/64	 hydroxyl: 29/45
2	 hydroxyl: 33/54	 amide: 22/38	 carbonyl group: 33/64	 amide: 11/24
3	 amide: 29/54	 carbonyl group: 21/38	 aniline: 31/64	 carbonyl group: 26/45
4	 aniline: 27/54	 or ..... bicyclic compounds: 18/38	 hydroxyl: 30/64	 unfused benzene ring: 19/45
5	 unfused benzene ring: 22/54	 aniline: 15/38	 unfused benzene ring: 27/64	 unfused benzene ring: 18/45

6	 or ..... bicyclic compounds: 22/54	 unfused benzene ring: 12/38	 or ..... bicyclic compounds: 27/64	 aniline: 18/45
7	 ether: 18/54	 phenol: 11/38	 aniline: 18/64	 phenol: 13/45
8	 carbonyl group with N: 16/54	 ether: 8/38	 ether: 15/64	 ether: 9/45
9	F, Cl, Br, I 13/54	 or  azole: 8/38	 phenol: 13/64	F, Cl, Br, I 8/45

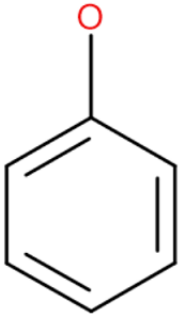
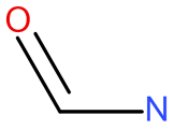
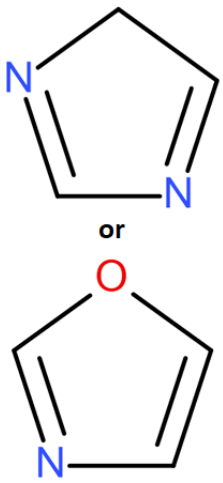
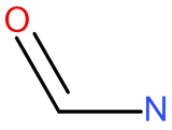
10	 phenol: 9/54	 carbonyl group with N: 8/38	 azole: 2/22	 carbonyl group with N: 8/45

Table S5: The top 10 frequently occurred functional groups in lop P set for each fingerprint. For each fingerprint, the occurrence frequency and the total number of molecules are also given.

	Top6-cons	Estate2	Estate1	daylight	Fp2	Ecfp	MACCS	Pharm2d	Erg
R_2	0.679	0.589	0.605	0.624	0.63	0.586	0.643	0.443	0.392
RMSE (-$\log_{10}$ mol/L)	0.549	0.619	0.607	0.593	0.59	0.62	0.58	0.725	0.756
Kendall tau	0.611	0.558	0.554	0.574	0.579	0.536	0.583	0.451	0.417

Table S6: Gradient boosting decision tree detailed results of the LD50 toxicity data set.

	Top6-cons	Estate2	Estate1	daylight	Fp2	ecfp	maccs	Pharm2d	erg
R_2	0.785	0.742	0.721	0.691	0.681	0.647	0.643	0.443	0.274
RMSE (-$\log_{10}$ mol/L)	0.457	0.5	0.522	0.561	0.562	0.592	0.58	0.725	0.877
Kendall tau	0.725	0.697	0.683	0.672	0.646	0.624	0.583	0.451	0.363

Table S7: Gradient boosting decision tree detailed results of the IGC50 toxicity data set.

	Top6-cons	Estate2	Estate1	daylight	Fp2	ecfp	maccs	Pharm2d	erg
R_2	0.715	0.662	0.651	0.623	0.609	0.573	0.608	0.528	0.348
RMSE (-$\log_{10}$ mol/L)	0.783	0.856	0.872	0.914	0.956	0.991	0.939	1.064	1.236
Kendall tau	0.662	0.62	0.627	0.572	0.553	0.579	0.58	0.507	0.394

Table S8: Gradient boosting decision tree detailed results of the LC50 toxicity data set.

	Top6-cons	Estate2	Estate1	daylight	Fp2	ecfp	maccs	Pharm2d	erg
R_2	0.486	0.502	0.52	0.377	0.357	0.452	0.434	0.275	0.336
RMSE (-$\log_{10}$ mol/L)	1.239	1.22	1.198	1.412	1.473	1.277	1.308	1.503	1.43
Kendall tau	0.512	0.53	0.559	0.485	0.443	0.467	0.438	0.361	0.353

Table S9: Gradient boosting decision tree detailed results of the LC50-DM toxicity data set.

	Cons-MT	Cons-ST	Estate2-MT	Estate2-ST	Estate1-MT	Estate1-ST	Daylight-MT	Daylight-ST
R_2	0.639	0.632	0.489	0.484	0.566	0.569	0.617	0.619
RMSE (-log$_{10}$ mol/L)	0.58	0.821	0.706	0.835	0.654	0.819	0.606	0.818
Kendall tau	0.577	0.572	0.483	0.474	0.534	0.528	0.568	0.58

Table S10: Deep learning detailed results of the LD50 toxicity data set.

	Cons-MT	Cons-ST	Estate2-MT	Estate2-ST	Estate1-MT	Estate1-ST	Daylight-MT	Daylight-ST
R_2	0.794	0.791	0.696	0.715	0.735	0.72	0.717	0.701
RMSE (-log$_{10}$ mol/L)	0.457	0.799	0.582	0.804	0.511	0.806	0.535	0.797
Kendall tau	0.729	0.722	0.663	0.652	0.686	0.683	0.674	0.672

Table S11: Deep learning detailed results of the IGC50 toxicity data set.

	Cons-MT	Cons-ST	Estate2-MT	Estate2-ST	Estate1-MT	Estate1-ST	Daylight-MT	Daylight-ST
R_2	0.765	0.687	0.66	0.569	0.694	0.65	0.724	0.57
RMSE (-log$_{10}$ mol/L)	0.718	1.224	0.876	1.243	0.796	1.217	0.806	1.232
Kendall tau	0.708	0.636	0.62	0.553	0.682	0.602	0.641	0.546

Table S12: Deep learning detailed results of the LC50 toxicity data set.

	Cons-MT	Cons-ST	Estate2-MT	Estate2-ST	Estate1-MT	Estate1-ST	Daylight-MT	Daylight-ST
R_2	0.725	0.523	0.623	0.433	0.684	0.601	0.694	0.346
RMSE (-log$_{10}$ mol/L)	0.935	1.558	1.076	1.559	0.969	1.525	0.981	1.612
Kendall tau	0.672	0.523	0.583	0.434	0.616	0.538	0.68	0.452

Table S13: Deep learning detailed results of the LC50-DM toxicity data set.

	Cons-6	ecfp	Fp2	Estate 2	Macss	Estate 1	Erg
R	0.716	0.697	0.696	0.694	0.69	0.683	0.678
RMSE(kcal/mol)	2.166	2.232	2.244	2.234	2.26	2.279	2.301
Kendall tau	0.513	0.504	0.494	0.493	0.492	0.479	0.472

Table S14: Cross validation results of CL1 in the S1322 data set.

	Cons-6	ecfp	Fp2	Estate 2	macss	Estate 1	erg
R	0.847	0.801	0.81	0.821	0.789	0.836	0.811
RMSE(kcal/mol)	1.538	1.741	1.69	1.639	1.77	1.581	1.693
Kendall tau	0.58	0.551	0.544	0.557	0.531	0.559	0.536

Table S15: Cross validation results of CL2 in the S1322 data set.

	Cons-6	ecfp	Fp2	Estate 2	macss	Estate 1	erg
R	0.708	0.644	0.684	0.671	0.686	0.636	0.677
RMSE(kcal/mol)	2.057	2.251	2.166	2.185	2.158	2.333	2.143
Kendall tau	0.528	0.484	0.513	0.489	0.5	0.458	0.505

Table S16: Cross validation results of CL3 in the S1322 data set.

	Cons-6	ecfp	Fp2	Estate 2	macss	Estate 1	erg
R	0.718	0.585	0.693	0.657	0.576	0.703	0.698
RMSE(kcal/mol)	1.716	2.001	1.768	1.853	2.022	1.742	1.752
Kendall tau	0.502	0.422	0.452	0.489	0.371	0.494	0.495

Table S17: Cross validation results of CL4 in the S1322 data set.

	Cons-6	ecfp	Fp2	Estate 2	macss	Estate 1	erg
R	0.831	0.75	0.764	0.773	0.773	0.799	0.742
RMSE(kcal/mol)	1.861	2.164	2.101	2.075	2.052	1.954	2.166
Kendall tau	0.612	0.568	0.538	0.541	0.517	0.56	0.48

Table S18: Cross validation results of CL5 in the S1322 data set.

	Cons-6	ecfp	Fp2	Estate 2	macss	Estate 1	erg
R	0.777	0.748	0.742	0.719	0.694	0.714	0.724
RMSE(kcal/mol)	1.892	2.011	1.984	2.057	2.143	2.087	2.061
Kendall tau	0.589	0.59	0.542	0.519	0.483	0.525	0.531

Table S19: Cross validation results of CL6 in the S1322 data set.

	Cons-6	ecfp	Fp2	Estate 2	macss	Estate 1	Erg
R	0.76	0.717	0.747	0.753	0.704	0.711	0.752
RMSE(kcal/mol)	1.572	1.707	1.631	1.592	1.737	1.723	1.606
Kendall tau	0.558	0.536	0.557	0.541	0.506	0.488	0.563

Table S20: Cross validation results of CL7 in the S1322 data set.

	Cons-6	ecfp	Fp2	Estate 2	macss	Estate 1	Erg
R	0.747	0.644	0.7	0.72	0.677	0.698	0.663
RMSE(kcal/mol)	2.016	2.267	2.125	2.059	2.181	2.121	2.218
Kendall tau	0.552	0.455	0.507	0.522	0.478	0.508	0.478

Table S21: Results of PDB-BIND v2016 data set.

	Top4-cons	Estate2	Estate1	daylight	Fp2	ecfp	maccs	Pharm2d	Erg
R_2	0.901	0.893	0.87	0.819	0.79	0.857	0.867	0.776	0.783
RMSE ($\log_{10}$ mol/L)	0.628	0.651	0.724	0.844	0.914	0.75	0.728	0.941	0.935
Kendall tau	0.845	0.839	0.82	0.774	0.759	0.802	0.817	0.723	0.724

Table S22: Results of FDA log P data set.

	Top4-cons	Estate2	Estate1	daylight	Fp2	ecfp	maccs	Pharm2d	Erg
R_2	0.857	0.823	0.799	0.729	0.79	0.823	0.776	0.629	0.584
RMSE ($\log_{10}$ mol/L)	0.625	0.692	0.745	0.857	0.911	0.688	0.781	1.004	1.094
Kendall tau	0.776	0.745	0.742	0.69	0.693	0.75	0.726	0.632	0.578

Table S23: Results of Star log P data set.

	Top4-cons	Estate2	Estate1	daylight	Fp2	ecfp	maccs	Pharm2d	Erg
R_2	0.741	0.764	0.699	0.729	0.588	0.656	0.658	0.487	0.542
RMSE ($\log_{10}$ mol/L)	1.233	1.168	1.287	0.857	1.434	1.396	1.436	1.634	1.625
Kendall tau	0.64	0.591	0.582	0.69	0.482	0.582	0.58	0.469	0.48

Table S24: Results of Nonstar log P data set.