Contents:

Table S9. Structural information for nine compounds with high blood brain barrier (BBB) permeability and eight compounds with low BBB permeability

Table S10. The meaning of descriptors selected by LASSO and random forests for the BBB data set

Figure S1. Comparison of the LASSO modeling and the best performance of GLM, RF, and SVM models based on variables selected by recursive random forests. Mean ± standard error is based on 20%-off bootstrapping (n = 10,000)

Table S2. Statistics of the best performance models developed using RF, SVM, LASSO, and GLM for four continuous data sets. Metrics include the determination of coefficient (R^2), R^2 between observed vs. predicted (R_0^2.xy) and predicted vs. observed (R_0^2.yx) through the origin[1], root mean squared errors (RMSE) and mean absolute error (MAE). The suffix sd means standard deviation of the corresponding metric

Data set	Model	R_0^2.xy	R_0^2.xy_sd	R_0^2.yx	R_0^2.yx_sd	R^2	R^2_sd	RMSE	RMSE_sd	MAE	MAE_sd
PPB	RF	0.462	0.050	0.594	0.023	0.623	0.021	0.208	0.007	0.143	0.006
	SVM	0.500	0.126	0.614	0.062	0.642	0.056	0.178	0.010	0.139	0.014
	LASSO	0.465	0.083	0.579	0.041	0.614	0.037	0.185	0.007	0.147	0.010
	GLM	0.426	0.140	0.543	0.076	0.589	0.064	0.220	0.020	0.149	0.015
BBB	RF	0.416	0.322	0.428	0.320	0.648	0.131	0.322	0.054	0.343	0.066
	SVM	0.288	0.350	0.289	0.384	0.603	0.127	0.479	0.109	0.375	0.082
	LASSO	-0.211	0.367	-0.321	0.520	0.446	0.095	0.379	0.088	0.433	0.040
	GLM	-0.214	0.219	-0.234	0.216	0.440	0.053	0.490	0.101	0.454	0.044
TP	RF	0.801	0.050	0.803	0.051	0.837	0.036	0.369	0.048	0.305	0.024
	SVM	0.824	0.053	0.824	0.053	0.852	0.039	0.363	0.049	0.271	0.027
	LASSO	0.793	0.063	0.794	0.064	0.830	0.044	0.434	0.055	0.319	0.031
	GLM	0.778	0.064	0.780	0.064	0.821	0.043	0.447	0.051	0.325	0.029
HDAC	RF	-0.212	0.924	0.564	0.208	0.584	0.212	0.367	0.114	0.580	0.010
	SVM	-0.140	0.968	0.549	0.165	0.589	0.172	0.383	0.041	0.539	0.084
	LASSO	-0.413	1.614	0.373	0.244	0.433	0.235	0.830	0.067	0.693	0.079
	GLM	-0.105	1.570	0.293	0.243	0.569	0.276	0.840	0.211	0.637	0.112

Table S3. Statistics of the best performance models developed using RF, SVM, LASSO, and GLM for four categorical data sets. Metrics include the sensitivity (sens.), specificity (spec.), correct classification rate (ccr), and accuracy. The suffix sd means standard deviation of the corresponding metric.

Data set	Model	sens.	sens_sd	spec.	spec_sd	ccr	ccr_sd	accuracy	accuracy_sd
BCPR	RF	0.753	0.086	0.826	0.118	0.790	0.049	0.798	0.050
	SVM	0.715	0.112	0.785	0.080	0.750	0.048	0.759	0.044
	LASSO	0.730	0.087	0.779	0.117	0.754	0.026	0.764	0.029
	GLM	0.733	0.069	0.787	0.059	0.760	0.043	0.764	0.041
MRP2	RF	0.900	0.105	0.858	0.080	0.879	0.081	0.875	0.079
	SVM	0.888	0.156	0.858	0.080	0.873	0.095	0.864	0.080
	LASSO	0.688	0.088	0.858	0.080	0.773	0.079	0.771	0.045
	GLM	0.763	0.093	0.876	0.079	0.819	0.077	0.813	0.056
PGP	RF	0.895	0.038	0.761	0.113	0.828	0.054	0.835	0.050
	SVM	0.884	0.026	0.696	0.109	0.790	0.043	0.789	0.041
	LASSO	0.693	0.113	0.683	0.095	0.688	0.079	0.680	0.079
	GLM	0.782	0.069	0.761	0.079	0.771	0.029	0.763	0.020
LT	RF	0.684	0.056	0.627	0.050	0.655	0.051	0.657	0.053
	SVM	0.700	0.115	0.630	0.073	0.665	0.040	0.671	0.045
	LASSO	0.613	0.071	0.530	0.133	0.571	0.060	0.579	0.053
	GLM	0.671	0.054	0.564	0.068	0.617	0.030	0.623	0.030

Table S4. The number of descriptors in the five optimal variable sets selected by LASSO under the context of 5-fold cross-validation. The number of descriptors of their intersection and union sets showed the consistencies among these variable sets

Data	M1	M2	M3	M4	M5	Intersection	Union
PPB	147	177	170	198	191	68	319
BBB	17	36	17	46	29	7	69
TP	89	87	83	81	93	59	105
HDAC	28	31	29	29	28	4	40
BCPR	81	65	68	59	54	25	133
MRP2	7	5	13	23	24	3	35
PGP	14	57	18	25	50	6	73
LT	20	51	43	56	52	7	81

Table S5. The performance of RF-LASSO models based on variables selected by LASSO for the four continuous data sets

Data	R_0^2.xy	R_0^2.xy_sd	R_0^2.yx	R_0^2.yx_sd	R^2	R^2_sd	RMSE	RMSE_sd	MAE	MAE_sd
PPB	0.459	0.081	0.596	0.040	0.624	0.036	0.179	0.006	0.142	0.012
BBB	0.124	0.491	0.125	0.473	0.556	0.144	0.314	0.038	0.366	0.065
TP	0.768	0.050	0.772	0.048	0.816	0.034	0.383	0.037	0.325	0.030
HDAC	-0.747	1.443	0.455	0.130	0.538	0.151	0.373	0.099	0.616	0.054

Table S6. The performance of RF-LASSO models based on variables selected by LASSO for the four categorical data sets

Data	sens.	sens_sd	spec.	spec_sd	ccr	ccr_sd	accuracy	accuracy_sd	auc	auc_sd
BCPR	0.675	0.101	0.841	0.081	0.758	0.045	0.772	0.046	0.870	0.033
MRP2	0.775	0.105	0.858	0.080	0.816	0.082	0.812	0.061	0.864	0.118
PGP	0.810	0.085	0.701	0.121	0.756	0.040	0.742	0.049	0.846	0.037
LT	0.632	0.075	0.548	0.140	0.590	0.064	0.596	0.062	0.605	0.069

Table S7. The highest and second highest pairwise correlation rate (HPCR1 and HPCR2) of the optimal variable sets selected by RF and LASSO

Model		PPB	BBB	TP	HDAC	BCPR	MRP2	PGP	LT
RF	HPCR1	0.51	0.50	0.64	0.46	0.77	0.25	0.14	0.00
	HPCR2	0.30	0.33	0.62	0.42	0.80	0.16	0.26	0.54
LASSO	HPCR1	0.28	0.14	0.66	0.00	0.090	0.00	0.24	0.30
	HPCR2	0.28	0.10	0.66	0.060	0.26	0.070	0.21	0.27

Table S8. The background value of highest and second highest pairwise correlation rate (HPCR1 and HPCR2) for all eight data sets. Mean ± standard error is based on 20%-off bootstrapping (n = 10,000)

Model		PPB	BBB	TP	HDAC	BCPR	MRP2	PGP	LT
RF	HPCR1	0.049	0.043	0.38	0.27	0.53	0.041	0.052	0.10
	HPCR2	0.048	0.043	0.38	0.26	0.53	0.040	0.052	0.10
LASSO	HPCR1	0.21	0.11	0.75	0.20	0.19	0.049	0.28	0.36
	HPCR2	0.21	0.11	0.75	0.20	0.19	0.048	0.28	0.36

Table S9. Structural information for nine compounds with high blood brain barrier (BBB) permeability and eight compounds with low BBB permeability

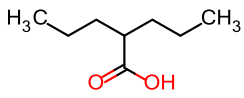
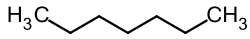
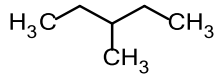
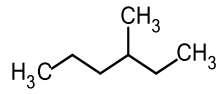
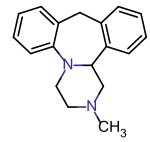
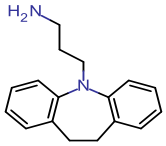
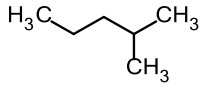
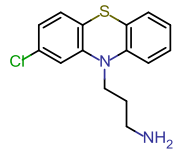
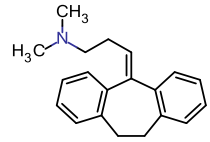
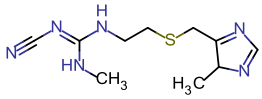
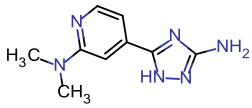
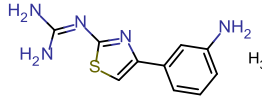
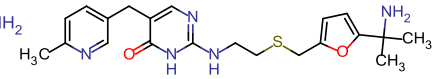
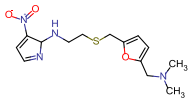
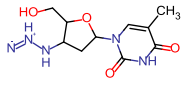
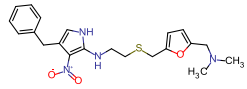
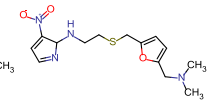
Hexane $logBB_{exp}$ = 0.80 $logBB_{pred}$ = 0.77 	Heptane $logBB_{exp}$ = 0.81 $logBB_{pred}$ = 0.78 	3-Methylpentane $logBB_{exp}$ = 1.01 $logBB_{pred}$ = 0.84 	3-Methylhexane $logBB_{exp}$ = 0.9 $logBB_{pred}$ = 0.86 	Mianserin $logBB_{exp}$ = 0.99 $logBB_{pred}$ =0.86 
Desmethyl-desipramine $logBB_{exp}$ = 1.06 $logBB_{pred}$ = 0.88 	2-Methylpentane $logBB_{exp}$ = 0.97 $logBB_{pred}$ = 0.91 	Chlorpromazine $logBB_{exp}$ = 1.06 $logBB_{pred}$ = 0.94 	36 $logBB_{exp}$ = 0.89 $logBB_{pred}$ = 0.99 	
Cimetidine $logBB_{exp}$ = -1.42 $logBB_{pred}$ = -1.23 	11 $logBB_{exp}$ = -1.17 $logBB_{pred}$ = -1.16 	17 $logBB_{exp}$ = -1.15 $logBB_{pred}$ = -1.12 	Lupitidine $logBB_{exp}$ = -1.06 $logBB_{pred}$ = -1.09 	
Zidovudine $logBB_{exp}$ = -0.87 $logBB_{pred}$ = -0.79 	Domperidone $logBB_{exp}$ = -0.78 $logBB_{pred}$ = -0.76 	Compound 36 $logBB_{exp}$ = -0.73 $logBB_{pred}$ = - 0.68 	Compound 35 $logBB_{exp}$ = -1.12 $logBB_{pred}$ = -0.93 	

Table S10. The meaning of descriptors selected by LASSO and random forests for the BBB data set

Descriptor	Model	Meaning
MATS3p	LASSO	Moran autocorrelation of lag 3 weighted by polarizability
MATS7p	LASSO	Moran autocorrelation of lag 7 weighted by polarizability
RBF	LASSO	rotatable bond fraction
MATS2m	LASSO	Moran autocorrelation of lag 2 weighted by mass
X3Av	LASSO	average valence connectivity index of order 3
nCb-	LASSO	number of substituted benzene C(sp2)
nArNR2	LASSO	number of tertiary amines (aromatic)
D/Dr11	LASSO	distance/detour ring index of order 11
nO	LASSO	number of Oxygen atoms
JGI10	LASSO	mean topological charge index of order 10
nC(=N)N2	LASSO	number of guanidine derivatives
C-006	LASSO	CH2RX
nCconj	LASSO	number of X on exo-conjugated C Functional group count
SEigm	LASSO	Barysz matrix weighted by mass (Dz(m))
GATS6e	LASSO	Geary autocorrelation of lag 5 weighted by Sanderson electronegativity
GATS5m	LASSO	Geary autocorrelation of lag 5 weighted by mass
nR05	LASSO	number of 5-membered rings
nR06	LASSO	number of 6-membered rings
nN	RF/LASSO	number of Nitrogen atoms
T(O..O)	RF/LASSO	sum of topological distances between O..O
T(N..N)	RF	sum of topological distances between N..N
Ms	RF	mean electrotopological state
MAXDN	RF	maximal electrotopological negative variation
GGI10	RF	topological charge index of order 10
T(N..O)	RF	sum of topological distances between N..O
D/Dr05	RF	distance/detour ring index of order 5
ESpm15u	RF	Edge adjacency indices
ESpm15r	RF	Edge adjacency indices

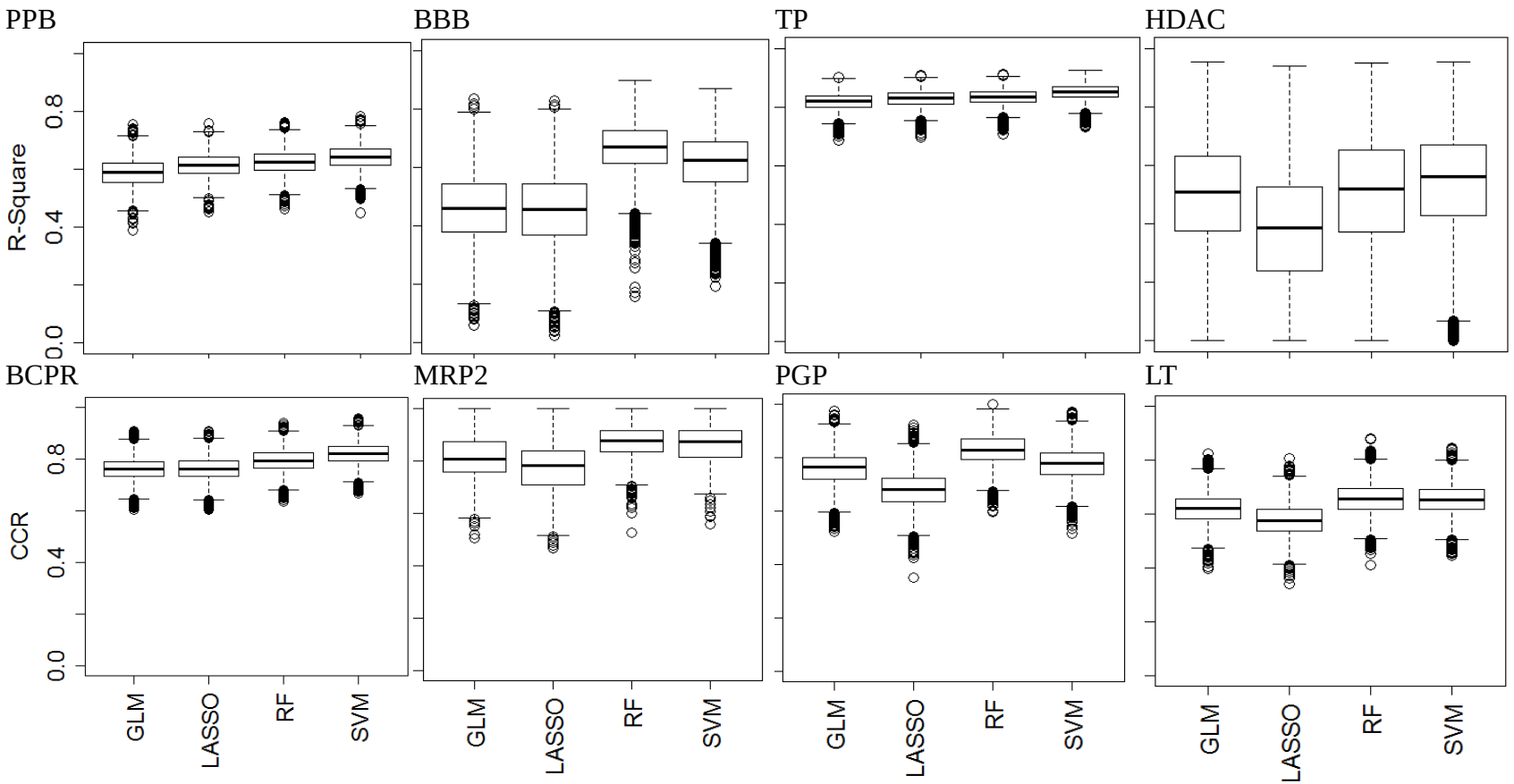


Figure S1. Comparison of the LASSO modeling and the best performance of GLM, RF, and SVM models based on variables selected by recursive random forests. Mean ± standard error is based on 20%-off bootstrapping (n = 10,000)

Reference

(1) Golbraikh, A.; Tropsha, A. Beware of q(2)! *J. Mol. Graph. Model.* **2002**, *20*, 269–276.