

SUPPLEMENTARY INFORMATION

LGBM-BBB: A Web-Server for Blood-Brain Barrier Permeability Prediction using Winsorized Feature Engineering and Mathew Correlation Coefficient Threshold Optimisation

Vivek Yadav¹, Gauri Mishra², Jatin Jangra², Rajnish Kumar^{2*}

*¹Department of Computer Science, Indira Gandhi National Tribal University, Amarkantak,
Madhya Pradesh-484887, India*

*²Department of Pharmaceutical Engineering and Technology, Indian Institute of Technology
(Banaras Hindu University), Varanasi, Uttar Pradesh-221005, India*

*Corresponding author

Rajnish Kumar, M. Pharm., Ph. D.

Associate Professor

Department of Pharmaceutical Engineering & Technology

Indian Institute of Technology (Banaras Hindu University)

Varanasi-221005 (U.P.) India

Email: rajnish.phe@iitbhu.ac.in

ORCID: 0000-0002-7021-6996

Table S1: List of top 59 features selected after feature selection process to build the LGBM-BBB model

Sr.no.	Features
1	TopoPSA
2	nHBDon
3	PEOE_VSA10
4	ATSC1se
5	AATS5s
6	n4HRing
7	PEOE_VSA1
8	SdssC
9	AATS7dv
10	nAcid
11	AATS8s
12	MDEC-33
13	SlogP_VSA2
14	Estate_VSA10
15	AATSC0v
16	ATSC1s
17	SlogP
18	ATS8dv
19	AATS7Z
20	VSA_Estate3

21	SsssCH
22	C1SP2
23	IC1
24	AATS7d
25	Xch-5d
26	Kier2
27	AATS1i
28	Lipinski
29	VSA_Estate5
30	GhoseFilter
31	AETA_eta_F
32	nN
33	VSA_Estate2
34	SsNH2
35	MATS1c
36	nS
37	ZMIC3
38	NssS
39	PEOE_VSA5
40	nRot
41	ATSC4c
42	GATS2i
43	PEOE_VSA13
44	BCUTi-1I

45	n7FAHRing
46	ATSC2s
47	SlogP_VSA3
48	NdssC
49	BCUTs-1I
50	GATS8d
51	MDEC-22
52	ATSC7c
53	Xch-6d
54	VSA_EState7
55	C4SP3
56	GATS3d
57	IC2
58	GATS1c
59	GATS2p

BBB Permeability Distribution in B3DB dataset

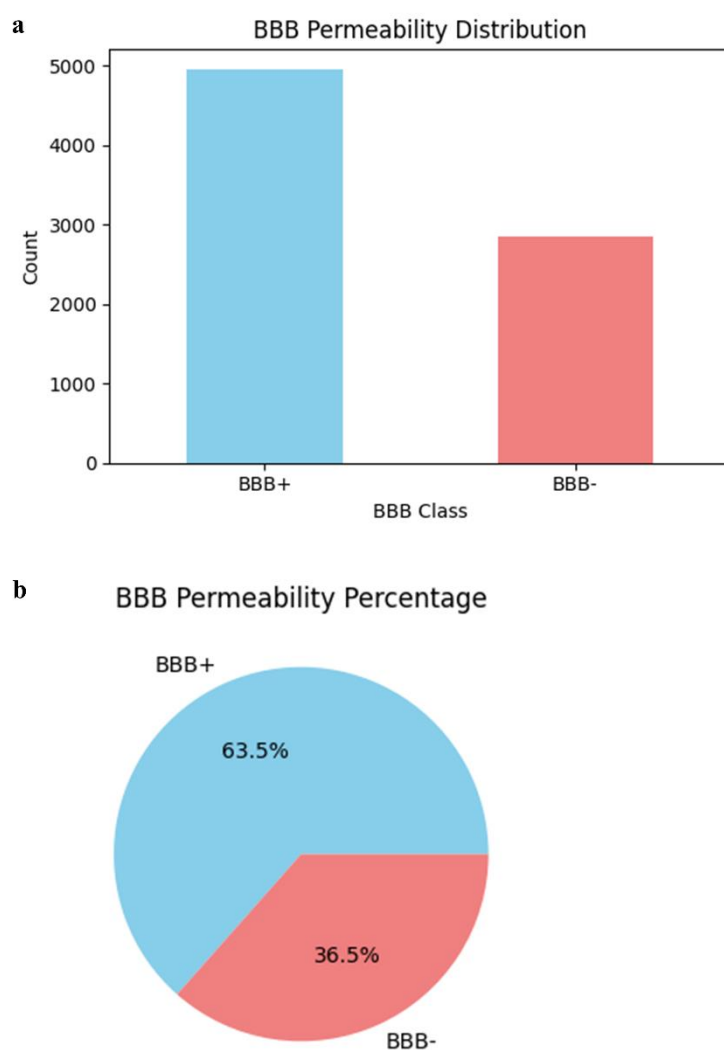


Fig. S1 The count and percentage distribution analysis of BBBP+ and BBBP- compounds across the B3DB dataset

The distribution of various descriptors across the B3DB dataset

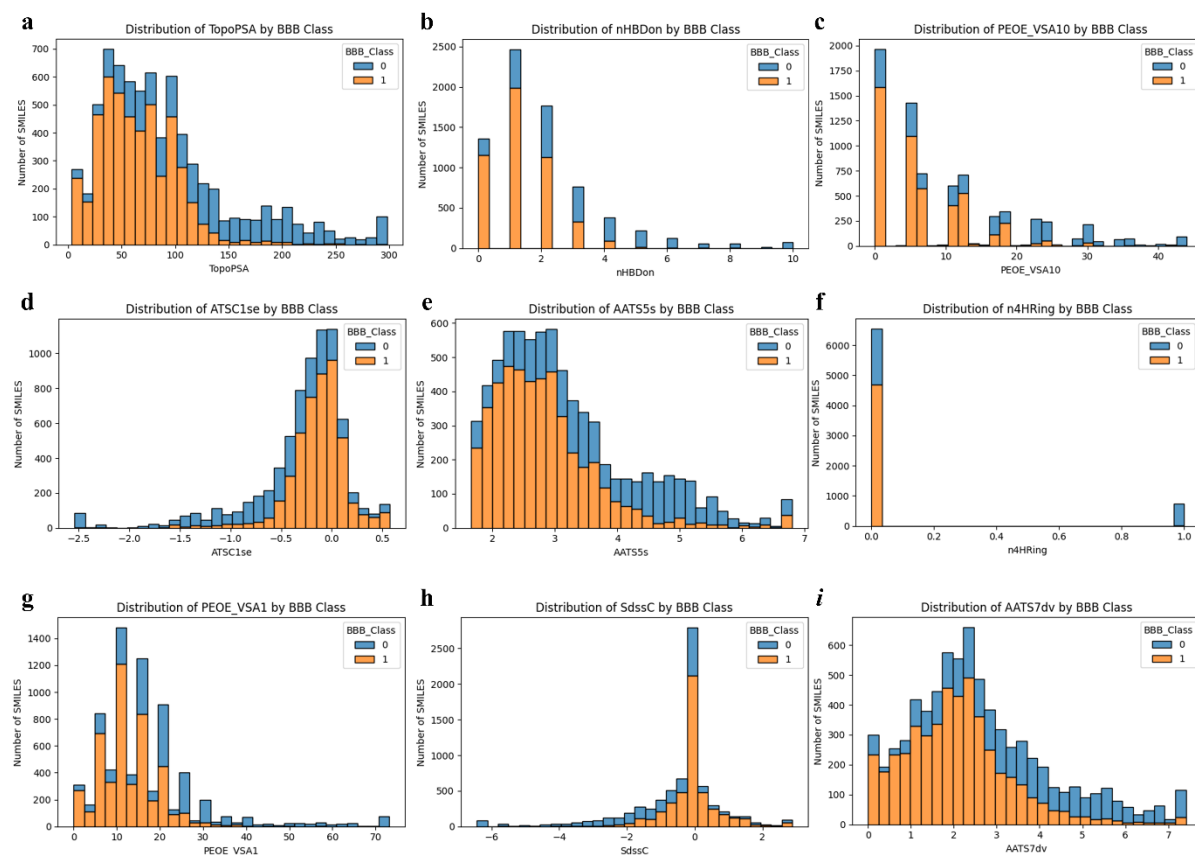


Fig. S2 (a-i) The distribution of (a) TopoPSA, (b) nHBDOn, (c) PEOE_VSA10, (d) ATSC1se, (e) AATS5s, (f) n4HRing, (g) PEOE_VSA1, (h) SdssC, (i) AATS7dv descriptors across BBBP+ and BBBP- compounds across the B3DB dataset

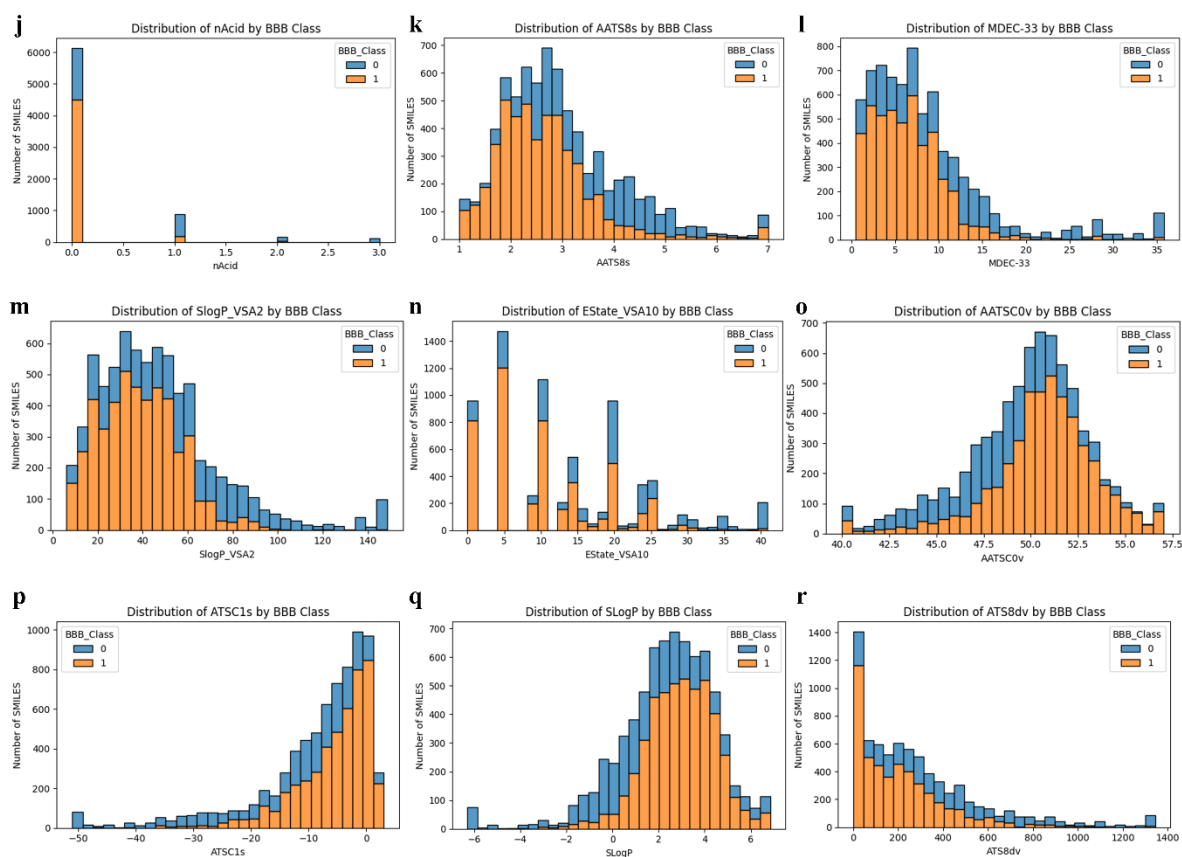


Fig. S2 (j-r) The distribution of (j) nAcid, (k) AATS8s, (l) MDEC-33, (m) SlogP_VSA2, (n) Estate_VSA10, (o) AATSC0v, (p) ATSC1s, (q) SlogP, (r) ATSC8dv descriptors across BBBP+ and BBBP- compounds across the B3DB dataset

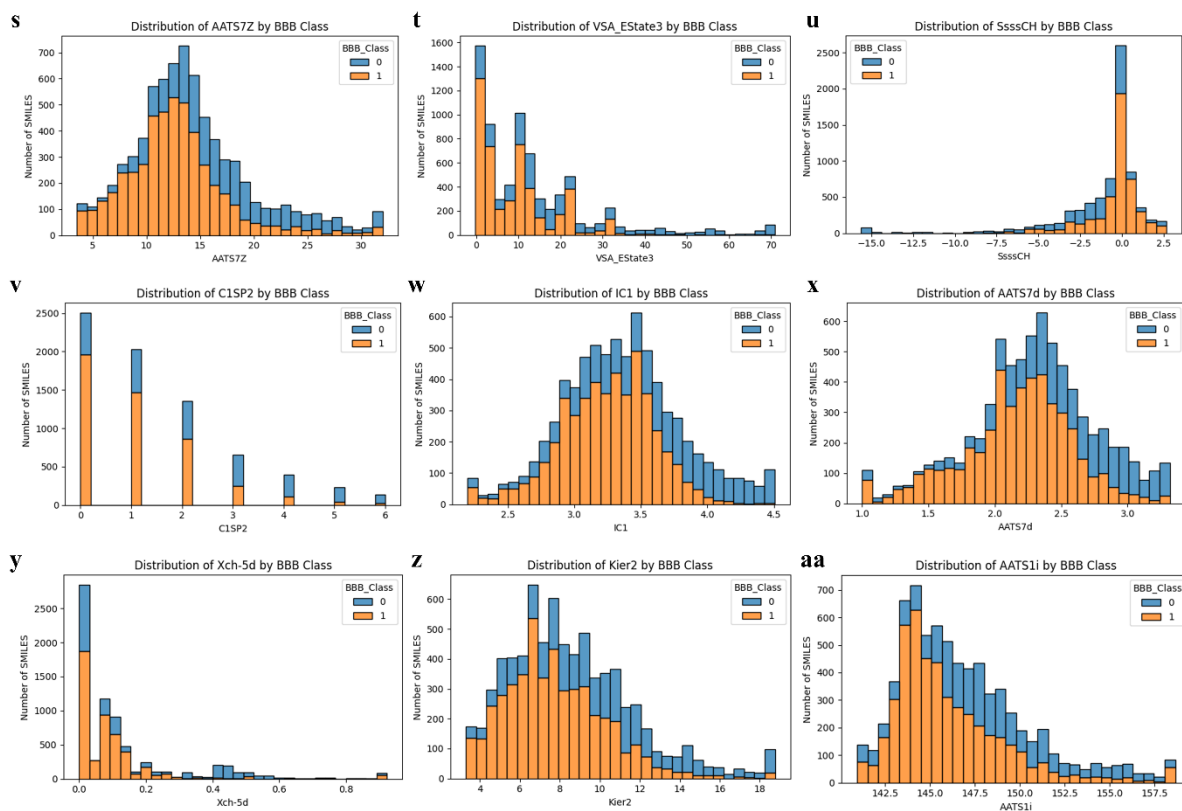


Fig. S2 (s-aa) The distribution of (s) AATS7Z, (t) VSA_Estate3, (u) SsssCH, (v) C1SP2, (w) IC1, (x) AATS7d, (y) Xch-5d, (z) Kier2, (aa) AATS1i descriptors across BBBP+ and BBBP- compounds across the B3DB dataset

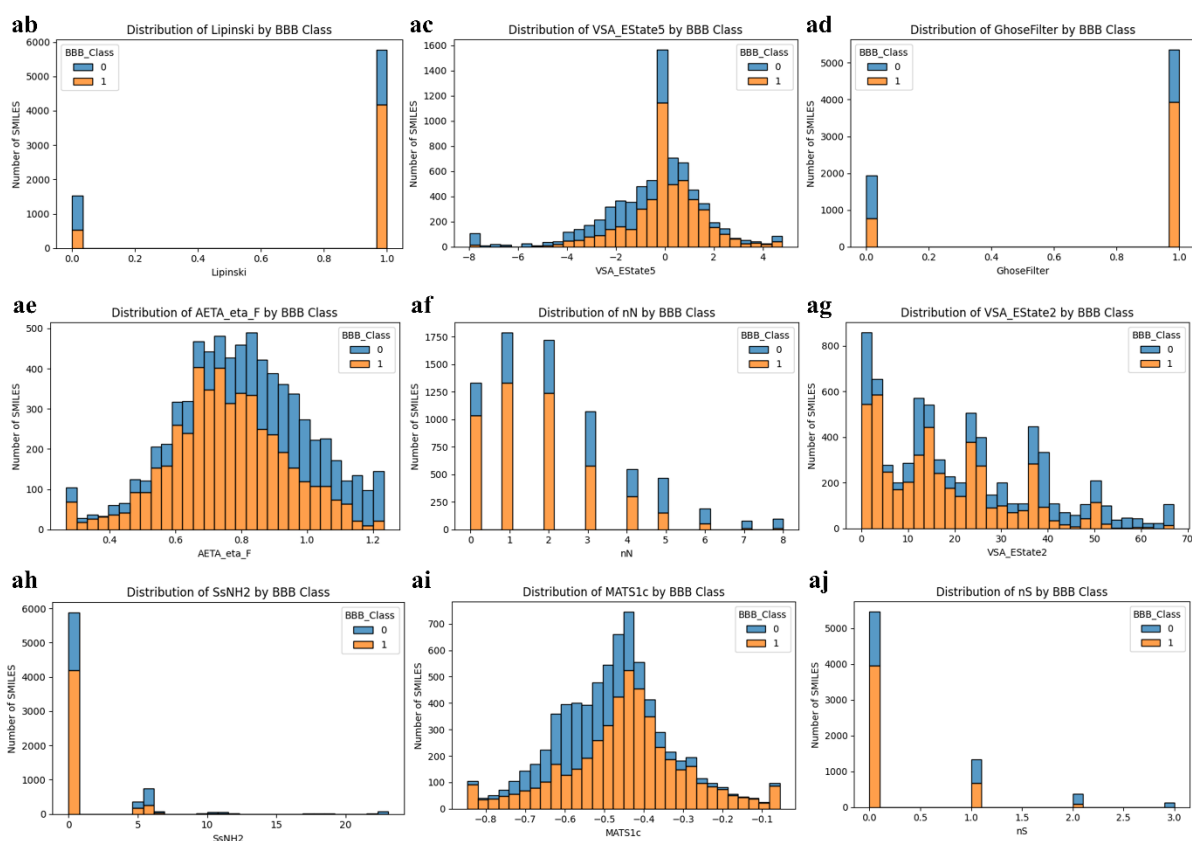


Fig. S2 (ab-aj) The distribution of (ab) Lipinski, (ac) VSA_Estate5, (ad) GhoseFilter, (ae) AETA_eta_F, (af) nN, (ag) VSA_Estate2, (ah) SsNH2, (ai) MATS1c, (aj) nS descriptors across BBBP+ and BBBP- compounds across the B3DB dataset

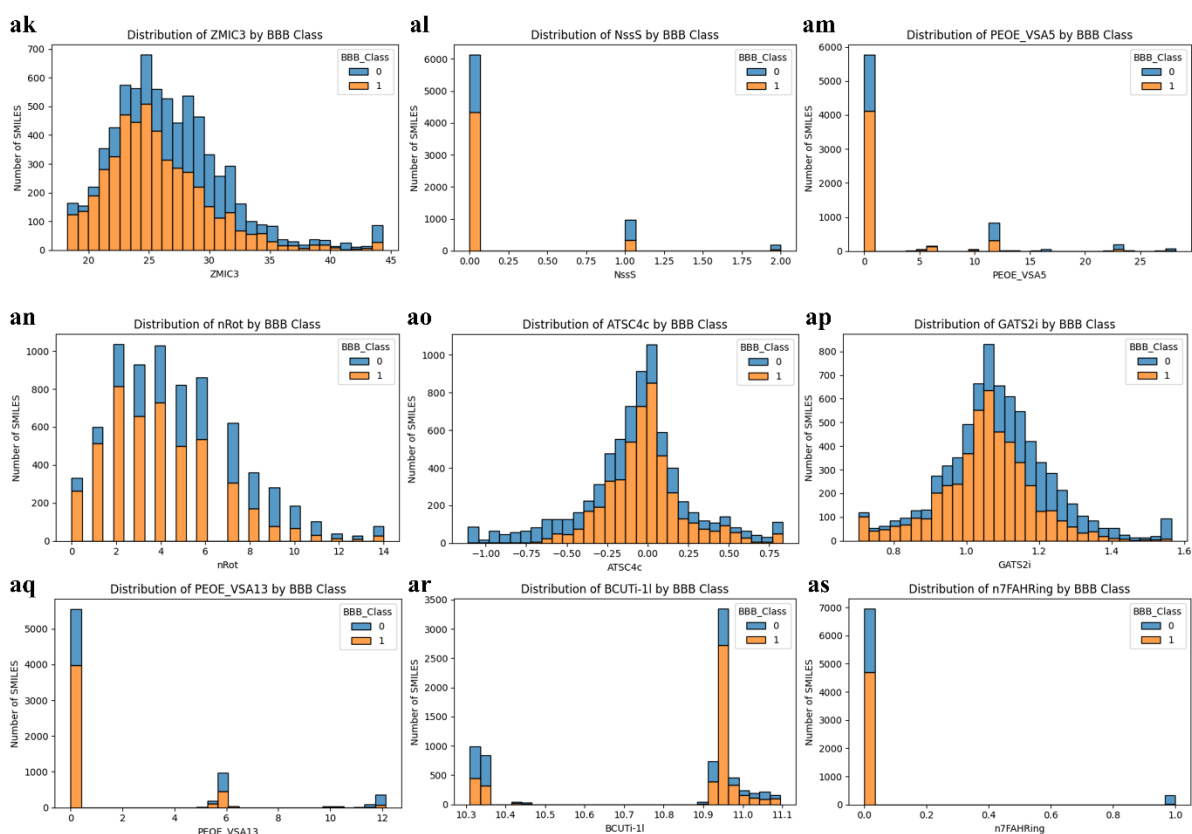


Fig. S2 (ak-as) The distribution of (ak) ZMIC3, (al) NssS, (am) PEOE_VSA5, (an) nRot, (ao) ATSC4c, (ap) GATS2i, (aq) PEOE_VSA13, (ar) BCUTi-1I, (as) n7FAHRing descriptors across BBBP+ and BBBP- compounds across the B3DB dataset

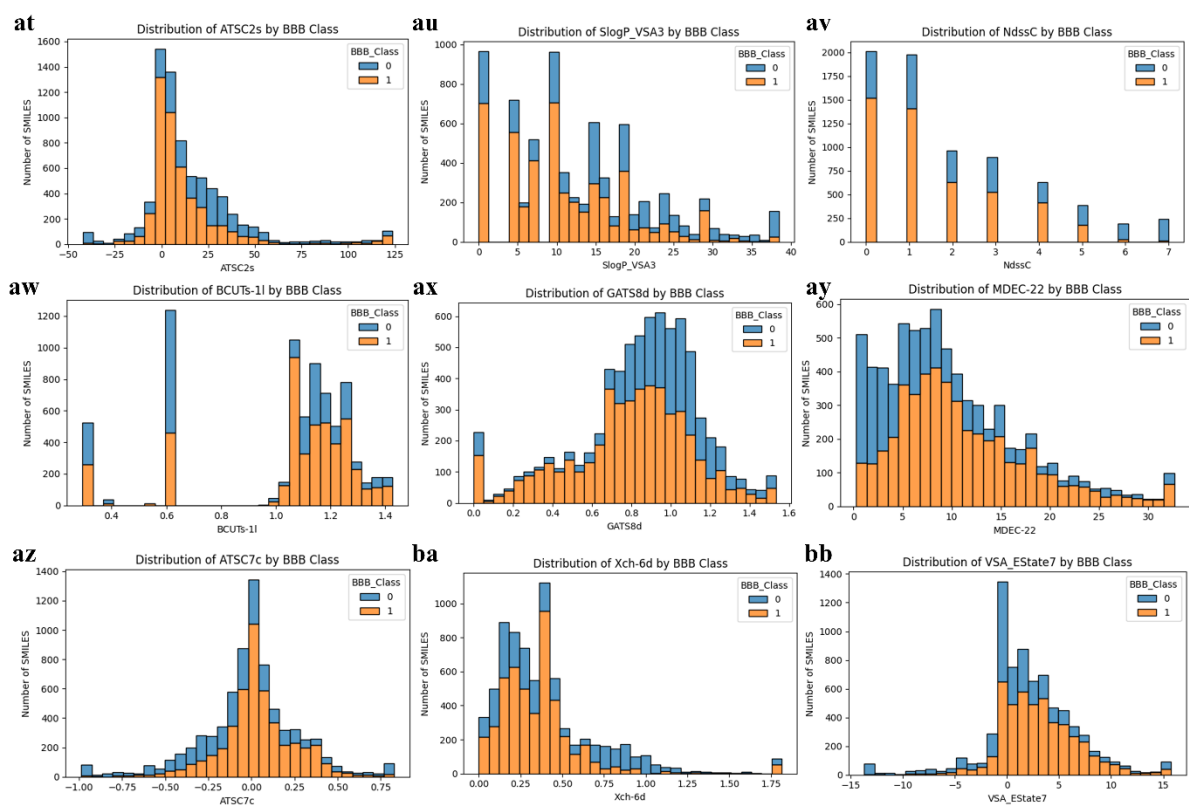


Fig. S2 (at-bb) The distribution of (at) ATSC2s, (au) SlogP_VSA3, (av) NdssC, (aw) BCUTs-1l, (ax) GATS8d, (ay) MDEC-22, (az) ATSC7c, (ba) Xch-6d, (bb) VSA_EState7 descriptors across BBBP+ and BBBP- compounds across the B3DB dataset

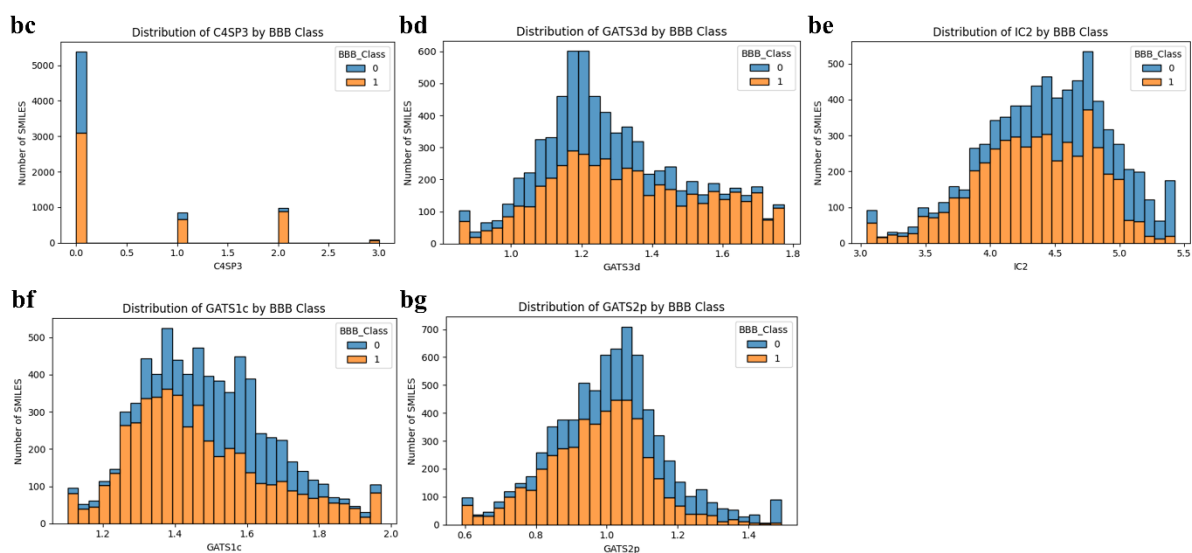


Fig. S2 (bc-bg) The distribution of (bc) C4SP3, (bd) GATS3d, (be) IC2, (bf) GATS1c, (bg) GATS2p descriptors across BBBP+ and BBBP- compounds across the B3DB dataset