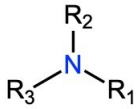
Supplementary information

Figures

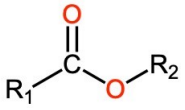
R = alkyl, alkenyl, aryl; X = F, Cl, Br, I

[CX4,CX3,c;!$([#6]=[OX1])][F,Cl,Br,I]



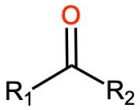
R1 = alkyl, aryl; R2, R3 = H, alkyl, aryl

[NX3]([#1,#6&!$([CX3]=[O,S,N,C])&!$([#6]([NX3])[N,O,S,P])&!$(C([NX3])#N)])
([#1,#6&!$([CX3]=[O,S,N,C])&!$([#6]([NX3])[N,O,S,P])&!$(C([NX3])#N)])[#1,#6&!$
([CX3]=[O,S,N,C])&!$([#6]([NX3])[N,O,S,P])&!$(C([NX3])#N)]



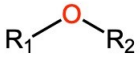
R1 = H, alkyl, aryl; R2 = alkyl, aryl

[#1,#6][CX3](=[OX1])[OX2][#6&!$([CX3]=[OX1,SX1,NX2])]



R1, R2 = alkyl, aryl

[#6&!$(C#N)][CX3](=[OX1])[#6&!$(C#N)]

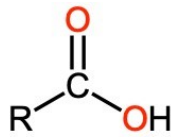


R1, R2 = alkyl, aryl

[#6&!$([CX4]([OX2])([#7,O,S,F,Cl,Br,I,P]))&!$([CX3]([OX2])=[OX1,SX1,NX2,C])]
[OX2][#6&!$([CX4]([OX2])([#7,O,S,F,Cl,Br,I,P]))&!$([CX3]
([OX2])=[OX1,SX1,NX2,C])]

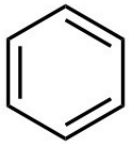
R = alkyl, aryl

[#6&!$([CX4]([OH1])[#8,#16,#7,#15])&!$([CX3]([OH1])=[OX1,SX1,NH2,C])][OH1]

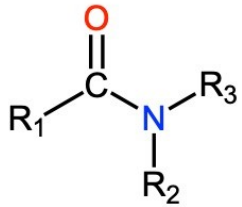


R = H, alkyl, aryl

[#1,#6][CX3](=[OX1])[OX2H1]

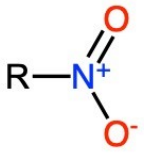


c1ccccc1



R1, R2, R3 = H, alkyl, aryl

[#1,#6][CX3](=[OX1])[NX3]([#1,#6&!$([CX3]=[OX1,SX1,NX1])])[#1,#6&!$([CX3]=[OX1,SX1,NX1])]



R = alkyl, aryl

[#6&!$([CX3]=[OX1,SX1])][$([NX3](=[OX1])=[OX1]),$([NX3+](=[OX1])[O-])]

Figure S1: SMARTS patterns used for functional group identification

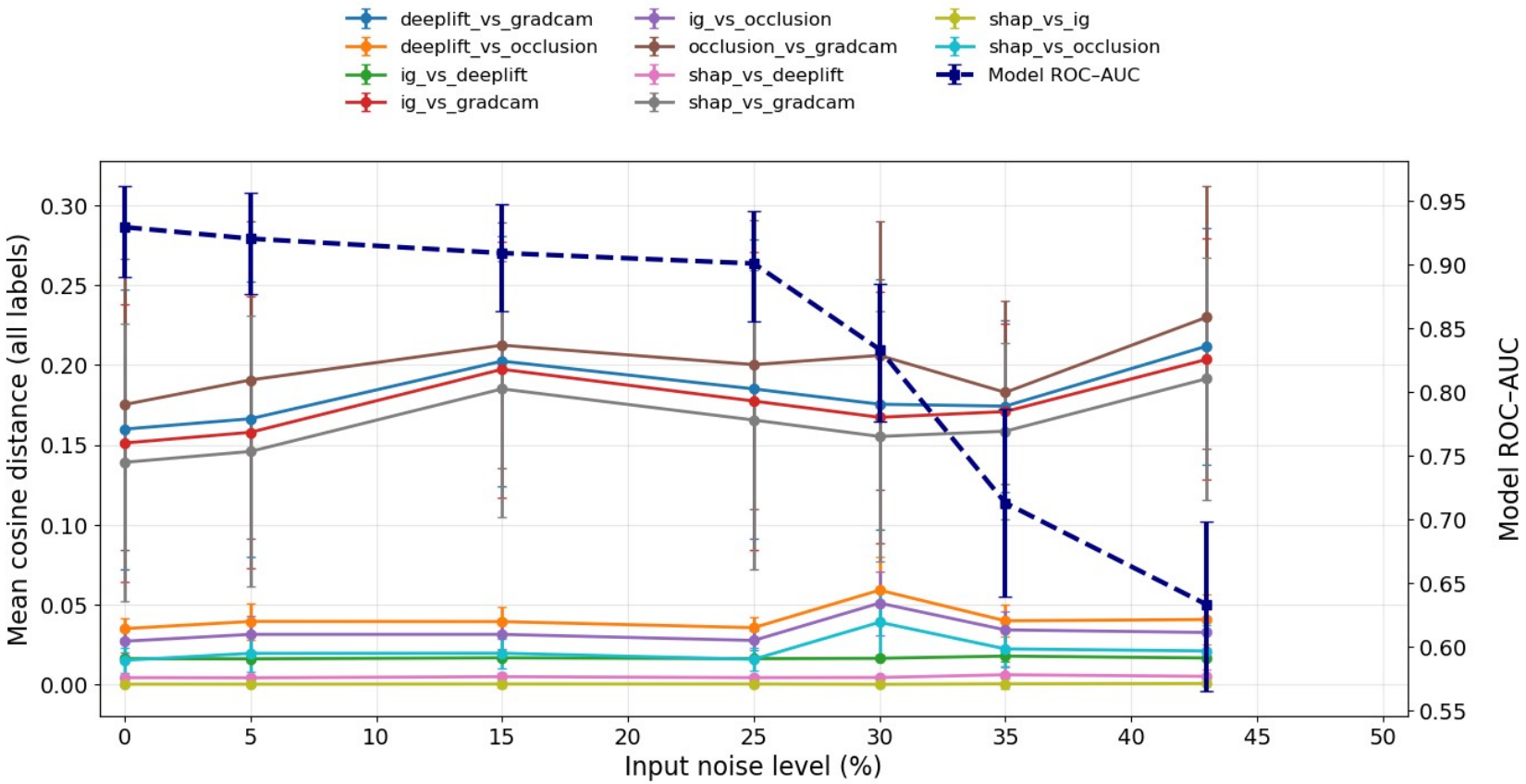


Figure S2: Cosine distance comparisons for all XAI method pairs at different input noise levels for BBB dataset

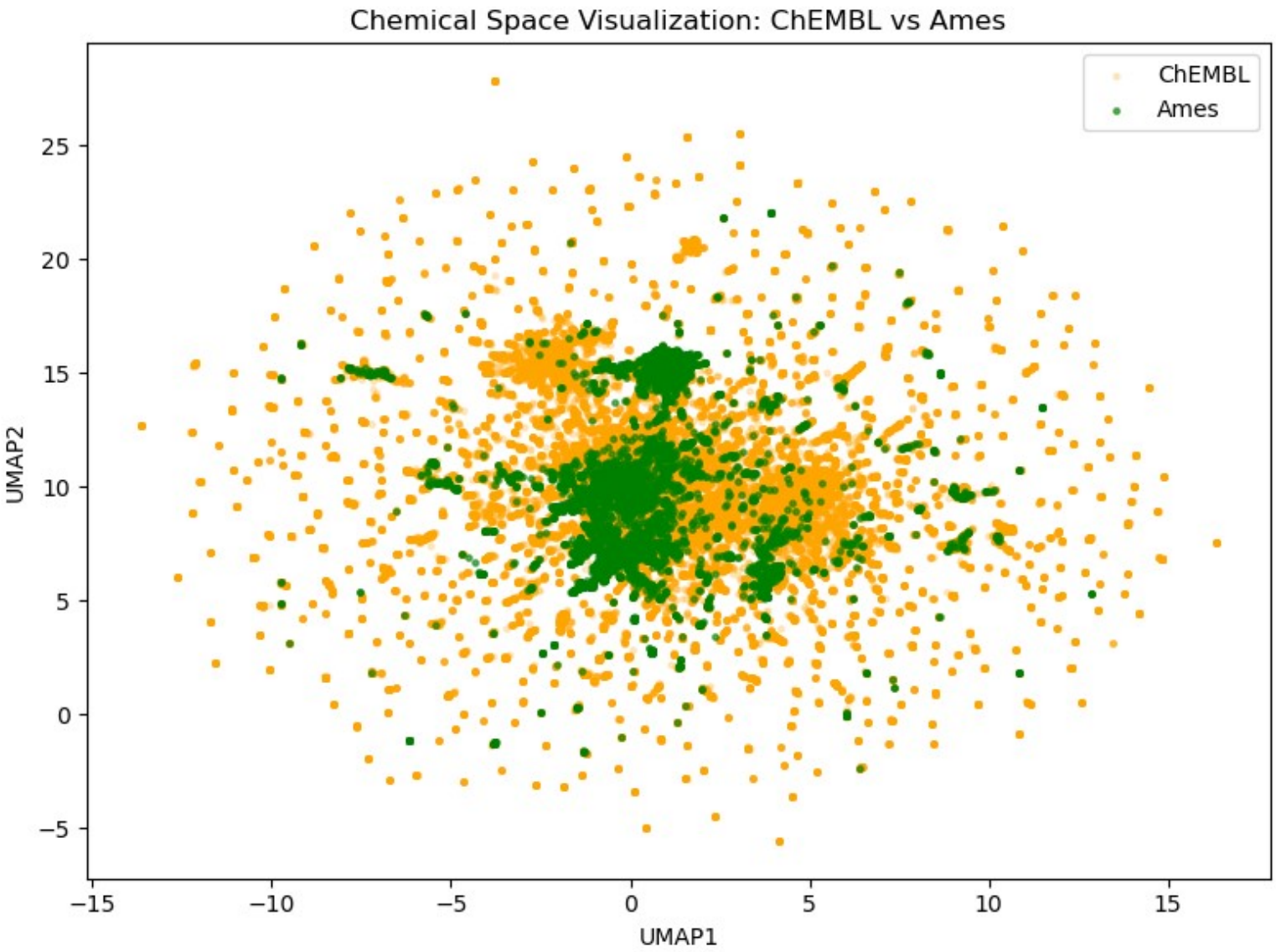


Figure S3: UMAP projection of molecular embedding space: ChEMBL vs Ames dataset

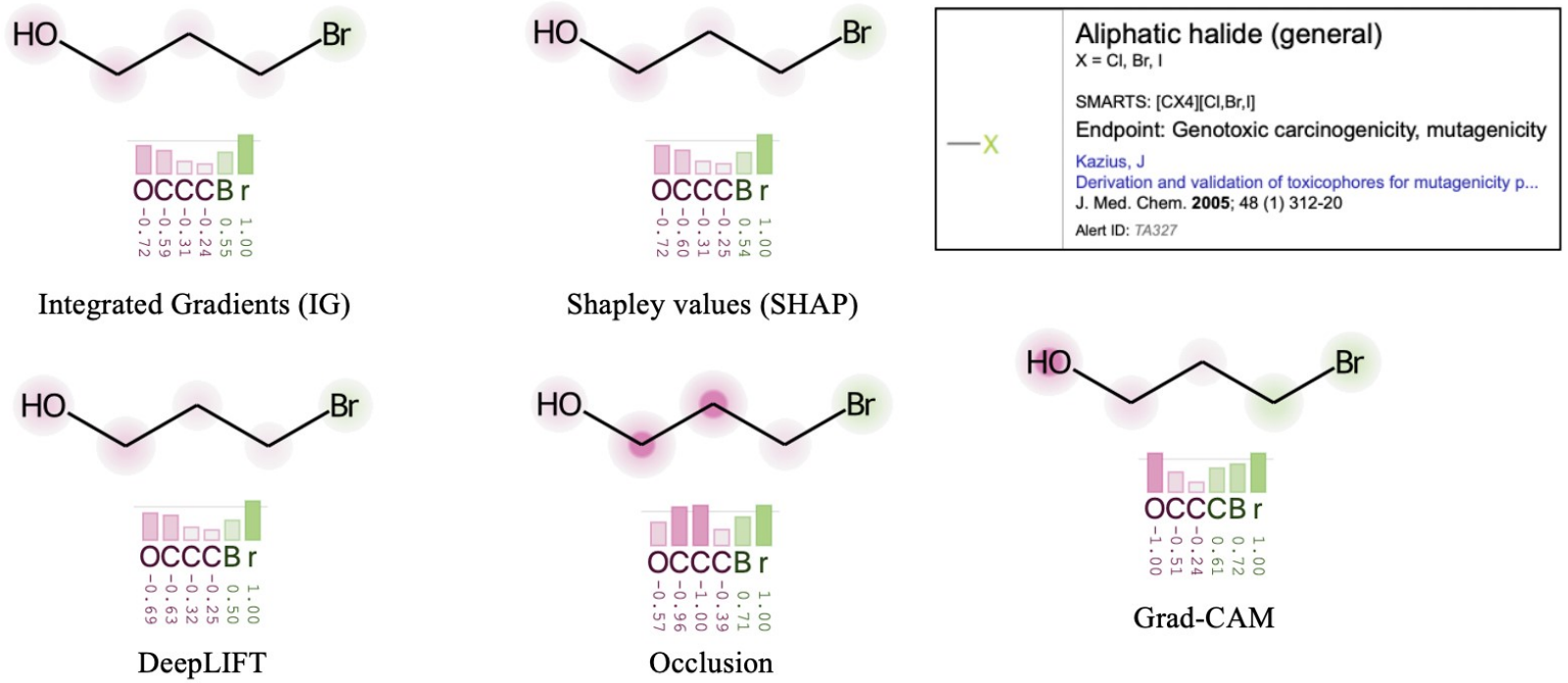


Figure S4: Comparison of atom-level attribution maps for the aliphatic halide toxicophore using five explainability methods: Integrated Gradients (IG), SHAP, DeepLIFT, Occlusion, and Grad-CAM (ROC–AUC = 0.85). In this example, all methods correctly highlight Br as part of the functional group, while Grad-CAM additionally emphasizes the adjacent aliphatic carbon. The color scale encodes relative importance, with magenta indicating low-importance tokens and green indicating high-importance tokens; attribution values were normalized to [-1, 1], such that colors reflect relative magnitude rather than the direction of contribution.

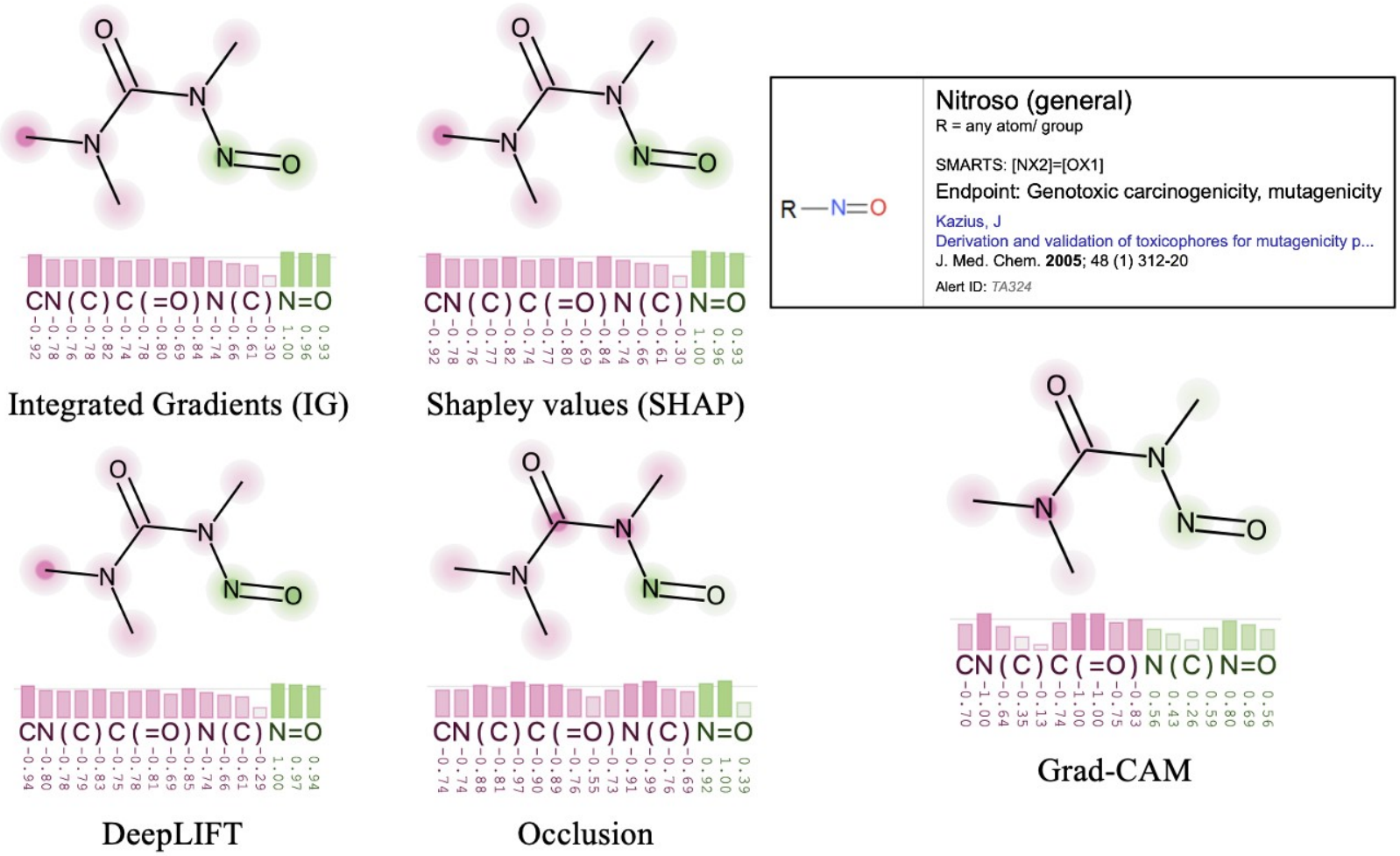


Figure S5: Comparison of atom-level attribution maps for the nitroso toxicophore using five explainability methods: Integrated Gradients (IG), SHAP, DeepLIFT, Occlusion, and Grad-CAM (ROC–AUC = 0.85). In this example, all methods highlight the nitroso group (N=O) as the key structural feature contributing to the prediction, while Grad-CAM additionally extends the attribution to the adjacent N–CH_3 substituent. The color

scale encodes relative importance, with magenta indicating low-importance tokens and green indicating high-importance tokens; attribution values were normalized to [-1, 1], such that colors reflect relative magnitude rather than the direction of contribution.

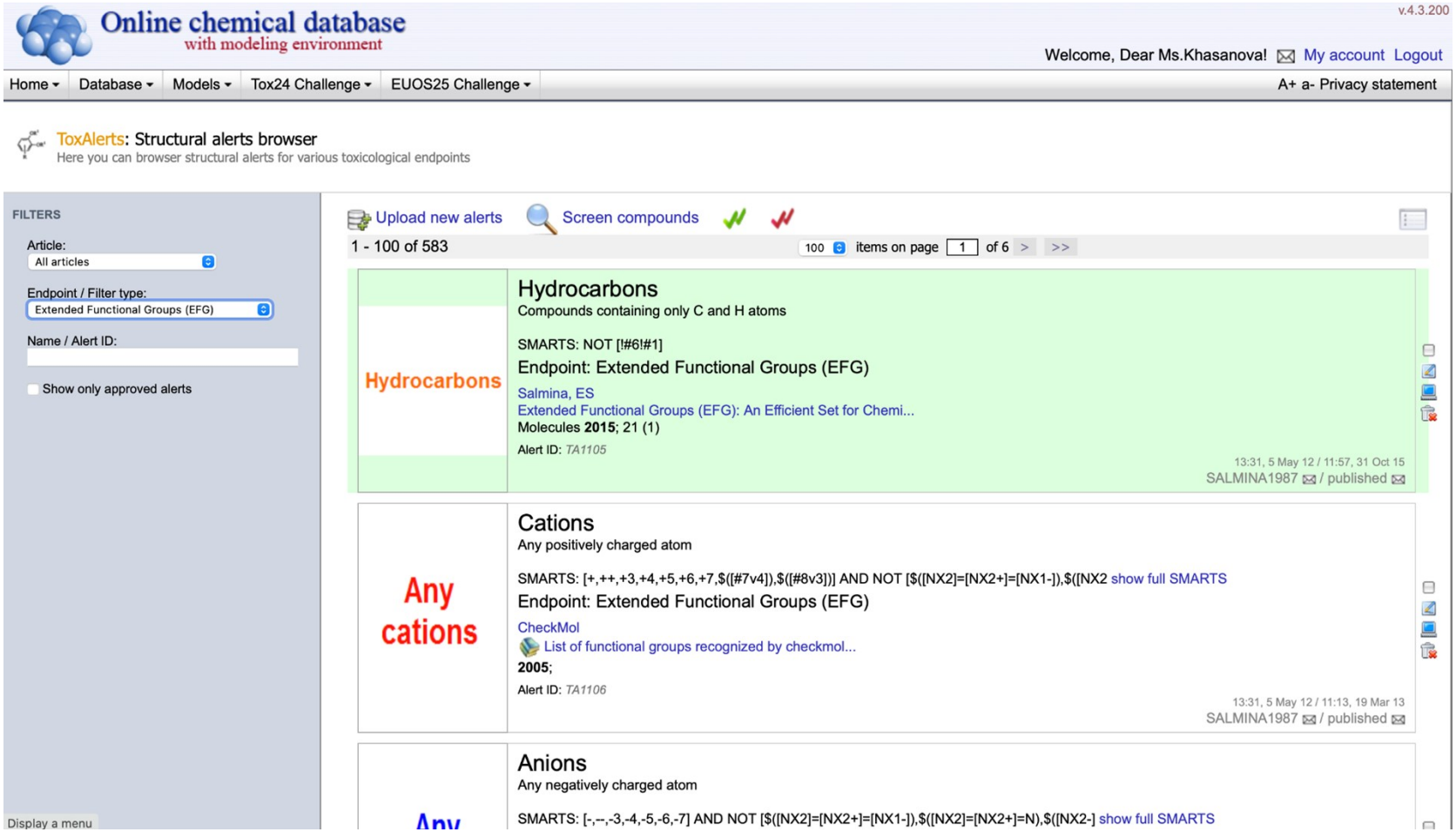


Figure S6: Accessing SMARTS patterns in the OCHEM ToxAlerts database

SMARTS patterns used in this study can be accessed through the OCHEM platform via the ToxAlerts database. The navigation path is as follows: OCHEM → Database → ToxAlerts → View alerts. Within the ToxAlerts interface, users can browse structural alerts corresponding to different toxicological endpoints as well as general chemical functional groups. Each alert entry includes the associated SMARTS pattern, which can be viewed and copied directly. Alternatively, the "Screen compounds against alerts" option allows users to apply these SMARTS patterns within the OCHEM environment. For convenience, all SMARTS patterns used in this study are also provided in a machine-readable format in the accompanying GitHub repository.

Tables

Table S1: Distribution of functional group counts in the dataset.

	0	1	2	3	4	5	6	7	8	9	10	11	12	15	19
halide	5839	669	380	174	87	37	41	8	9	2	4	1	2	1	1
amine	5624	1184	408	29	7	2	1	1	0	0	0	0	0	0	0
carboxylic_esters	6521	561	144	20	9	0	0	0	0	0	0	0	0	0	0
ketone	6550	412	260	22	10	0	0	0	0	0	0	0	0	0	0
ether	6025	896	241	48	25	8	5	4	2	0	1	0	0	0	0
carboxylic_acid	6653	518	72	4	8	0	0	0	0	0	0	0	0	0	0
arene	2558	2202	1445	588	331	105	21	2	2	1	0	0	0	0	0

carboxylic_amides	6669	487	85	7	7	0	0	0	0	0	0	0	0	0	0
nitro	6237	812	167	30	8	0	1	0	0	0	0	0	0	0	0
hydroxyl	5502	956	488	149	83	42	14	9	5	4	3	0	0	0	0

Table S2: Cosine distances between pairs of XAI methods for different functional groups

	halide positive	halide all	amine positive	amine all	carboxylic_esters positive	carboxylic_esters all
shap/ig	0.0000 ± 0.0002	0.0001 ± 0.0010	0.0001 ± 0.0009	0.0001 ± 0.0015	0.0001 ± 0.0003	0.0001 ± 0.0004
shap/deeplift	0.0010 ± 0.0022	0.0034 ± 0.0069	0.0021 ± 0.0032	0.0023 ± 0.0034	0.0056 ± 0.0060	0.0044 ± 0.0115
shap/occlusion	0.0556 ± 0.0562	0.0265 ± 0.0330	0.0307 ± 0.0402	0.0196 ± 0.0272	0.0314 ± 0.0303	0.0267 ± 0.0304
shap/gradcam	0.4060 ± 0.1782	0.2297 ± 0.1576	0.3881 ± 0.1850	0.2308 ± 0.1595	0.3316 ± 0.1670	0.2004 ± 0.1262
ig/deeplift	0.0010 ± 0.0020	0.0032 ± 0.0066	0.0019 ± 0.0028	0.0022 ± 0.0032	0.0053 ± 0.0055	0.0043 ± 0.0114
ig/occlusion	0.0557 ± 0.0564	0.0264 ± 0.0332	0.0306 ± 0.0401	0.0196 ± 0.0271	0.0313 ± 0.0301	0.0266 ± 0.0303
ig/gradcam	0.4061 ± 0.1785	0.2298 ± 0.1576	0.3881 ± 0.1848	0.2310 ± 0.1594	0.3321 ± 0.1676	0.2004 ± 0.1263
deeplift/occlusion	0.0525 ± 0.0533	0.0283 ± 0.0373	0.0335 ± 0.0394	0.0223 ± 0.0290	0.0327 ± 0.0349	0.0252 ± 0.0285

deeplift/gradcam	0.3938 ± 0.1788	0.2316 ± 0.1545	0.3698 ± 0.1836	0.2289 ± 0.1543	0.3357 ± 0.1723	0.2031 ± 0.1256
occlusion/gradcam	0.3406 ± 0.2107	0.2194 ± 0.1514	0.4059 ± 0.2057	0.2394 ± 0.1677	0.3994 ± 0.1810	0.2149 ± 0.1344

	ketone positive	ketone all	ether positive	ether all	carboxylic_acid positive	carboxylic_acid all
shap/ig	0.0002 ± 0.0013	0.0001 ± 0.0005	0.0001 ± 0.0003	0.0001 ± 0.0007	0.0002 ± 0.0009	0.0001 ± 0.0006
shap/deeplift	0.0013 ± 0.0022	0.0027 ± 0.0112	0.0021 ± 0.0036	0.0048 ± 0.0145	0.0023 ± 0.0035	0.0021 ± 0.0045
shap/occlusion	0.0170 ± 0.0192	0.0197 ± 0.0242	0.0313 ± 0.0382	0.0265 ± 0.0388	0.0279 ± 0.0255	0.0195 ± 0.0229
shap/gradcam	0.2845 ± 0.1727	0.1860 ± 0.1174	0.3746 ± 0.1799	0.2193 ± 0.1476	0.3081 ± 0.1432	0.1842 ± 0.1213
ig/deeplift	0.0012 ± 0.0019	0.0027 ± 0.0112	0.0020 ± 0.0036	0.0046 ± 0.0130	0.0022 ± 0.0033	0.0020 ± 0.0044
ig/occlusion	0.0166 ± 0.0187	0.0196 ± 0.0241	0.0312 ± 0.0382	0.0265 ± 0.0386	0.0282 ± 0.0259	0.0195 ± 0.0230
ig/gradcam	0.2848 ± 0.1729	0.1860 ± 0.1175	0.3743 ± 0.1803	0.2193 ± 0.1477	0.3079 ± 0.1439	0.1842 ± 0.1214
deeplift/occlusion	0.0198 ± 0.0246	0.0194 ± 0.0257	0.0333 ± 0.0384	0.0257 ± 0.0375	0.0362 ± 0.0326	0.0206 ± 0.0244
deeplift/gradcam	0.2730 ± 0.1703	0.1890 ± 0.1163	0.3575 ± 0.1803	0.2202 ± 0.1437	0.2921 ± 0.1418	0.1831 ± 0.1192
occlusion/gradcam	0.3193 ± 0.1843	0.1905 ± 0.1219	0.4047 ± 0.2080	0.2236 ± 0.1555	0.3991 ± 0.1703	0.1980 ± 0.1349

	arene positive	arene all	carboxylic_amides positive	carboxylic_amides all	nitro positive	nitro all
shap/ig	0.0002 ± 0.0018	0.0002 ± 0.0015	0.0001 ± 0.0003	0.0002 ± 0.0009	0.0000 ± 0.0001	0.0001 ± 0.0008
shap/deeplift	0.0014 ± 0.0025	0.0026 ± 0.0048	0.0102 ± 0.0118	0.0070 ± 0.0124	0.0010 ± 0.0012	0.0034 ± 0.0067
shap/occlusion	0.0121 ± 0.0113	0.0196 ± 0.0300	0.0362 ± 0.0276	0.0374 ± 0.0613	0.0112 ± 0.0132	0.0182 ± 0.0242
shap/gradcam	0.2486 ± 0.1418	0.2393 ± 0.1421	0.3275 ± 0.1688	0.1960 ± 0.1250	0.2038 ± 0.1347	0.1802 ± 0.1132
ig/deeplift	0.0011 ± 0.0017	0.0024 ± 0.0047	0.0098 ± 0.0107	0.0067 ± 0.0115	0.0010 ± 0.0013	0.0033 ± 0.0062
ig/occlusion	0.0118 ± 0.0108	0.0194 ± 0.0300	0.0365 ± 0.0280	0.0374 ± 0.0614	0.0112 ± 0.0131	0.0182 ± 0.0243
ig/gradcam	0.2484 ± 0.1421	0.2392 ± 0.1421	0.3272 ± 0.1685	0.1959 ± 0.1251	0.2039 ± 0.1346	0.1803 ± 0.1132
deeplift/occlusion	0.0122 ± 0.0113	0.0205 ± 0.0325	0.0474 ± 0.0418	0.0430 ± 0.0701	0.0129 ± 0.0153	0.0189 ± 0.0286
deeplift/gradcam	0.2416 ± 0.1401	0.2357 ± 0.1403	0.3203 ± 0.1696	0.1964 ± 0.1245	0.2005 ± 0.1329	0.1826 ± 0.1148
occlusion/gradcam	0.2585 ± 0.1401	0.2452 ± 0.1386	0.4103 ± 0.2102	0.2139 ± 0.1430	0.2353 ± 0.1481	0.1921 ± 0.1188

	hydroxyl positive	hydroxyl all
shap/ig	0.0002 ± 0.0014	0.0001 ± 0.0007
shap/deeplift	0.0011 ± 0.0021	0.0054 ± 0.0079
shap/occlusion	0.0457 ± 0.0572	0.0329 ± 0.0376
shap/gradcam	0.4473 ± 0.1916	0.2539 ± 0.1744
ig/deeplift	0.0010 ± 0.0018	0.0053 ± 0.0078
ig/occlusion	0.0453 ± 0.0570	0.0327 ± 0.0374
ig/gradcam	0.4471 ± 0.1918	0.2538 ± 0.1743
deeplift/occlusion	0.0431 ± 0.0525	0.0358 ± 0.0383
deeplift/gradcam	0.4319 ± 0.1939	0.2599 ± 0.1683
occlusion/gradcam	0.4048 ± 0.2276	0.2364 ± 0.1742