

Coverage bias in small molecule machine learning

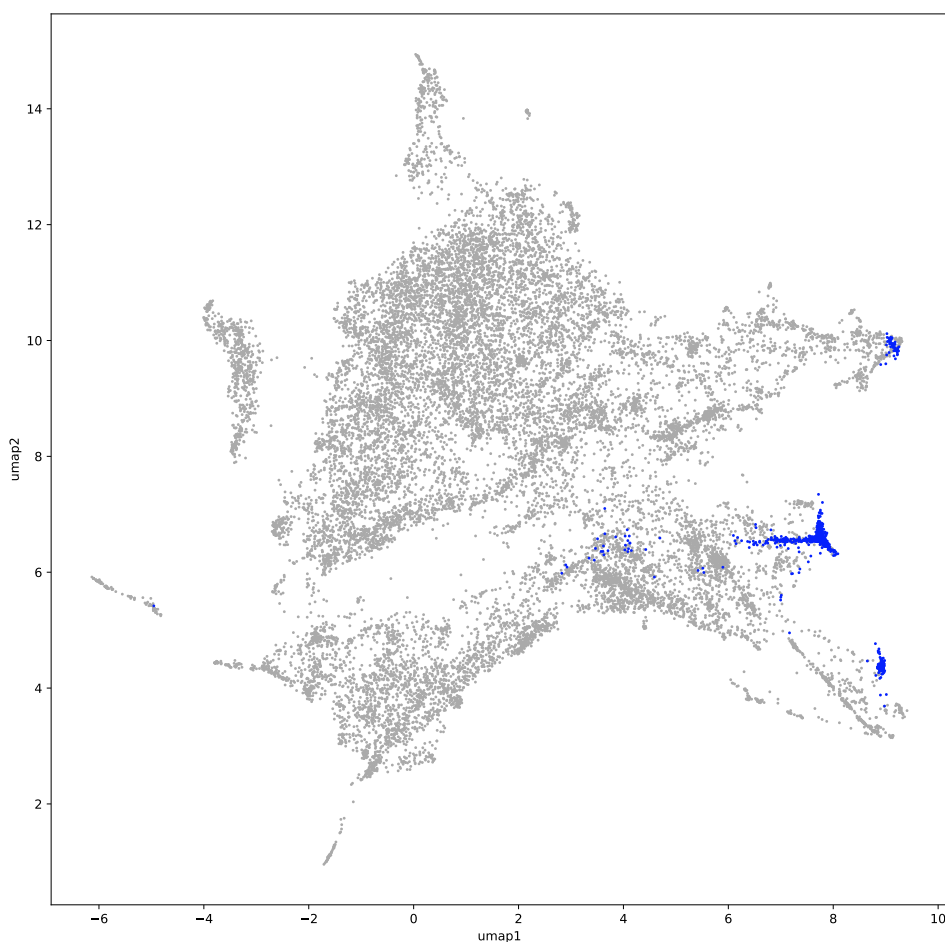
Supplementary material

Fleming Kretschmer¹, Jan Seipp², Marcus Ludwig^{1,3}, Gunnar W. Klau², and Sebastian Böcker¹

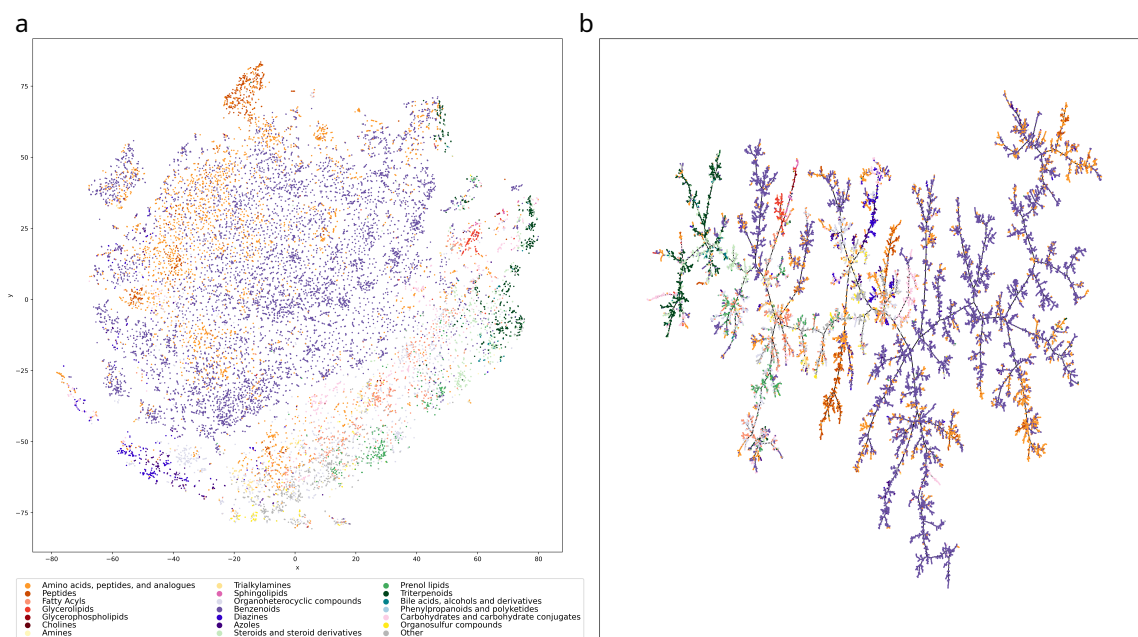
¹ Chair for Bioinformatics, Institute for Computer Science, Friedrich Schiller University Jena, Jena, Germany

² Algorithmic Bioinformatics, Institute for Computer Science, Heinrich Heine University Düsseldorf, Germany

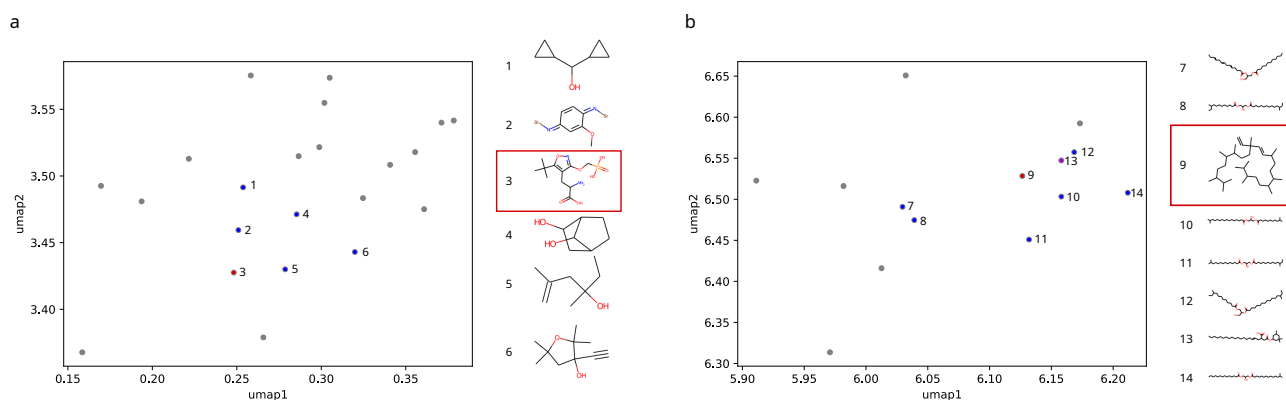
³ Currently at Bright Giant, Jena, Germany



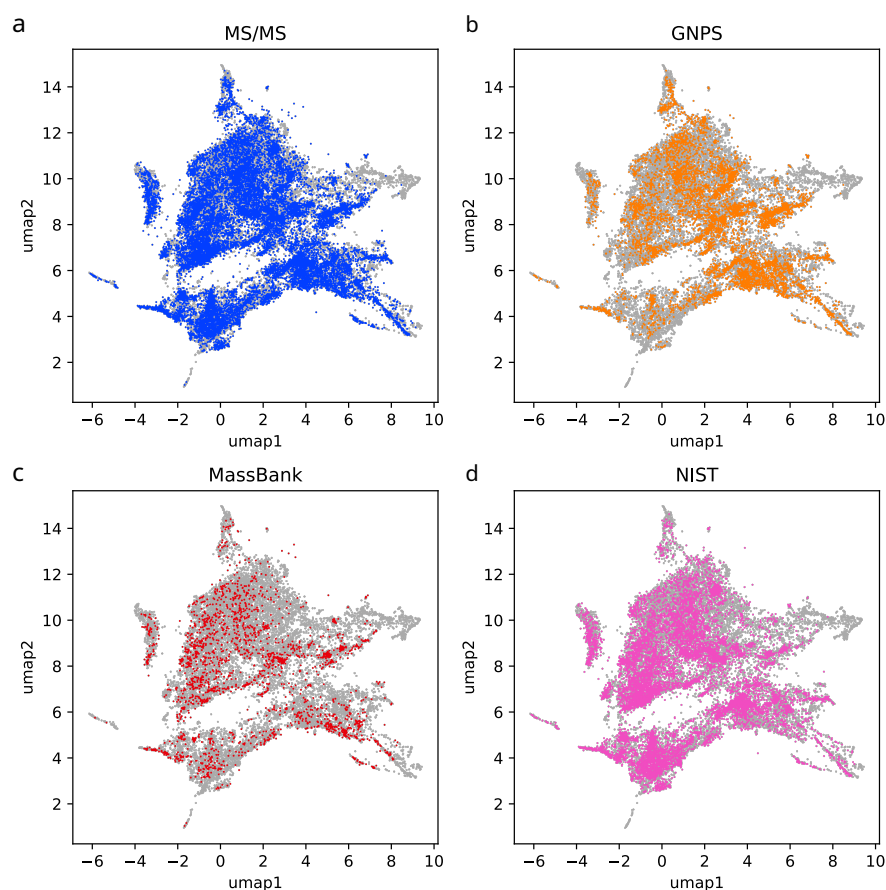
Supplementary Fig. 1. Full map of biomolecular structures. We did not recompute the UMAP embedding for this figure. Rather, the same UMAP embedding as in Fig. 3 was used. UMAP embedding computed from the 18,096 molecular structures. Shown in this plot are all 19,994 molecular structures subsampled from the biomolecular structures. Highlighted are the 1,898 molecular structures not shown in Fig. 3. Compare to Fig. 1.



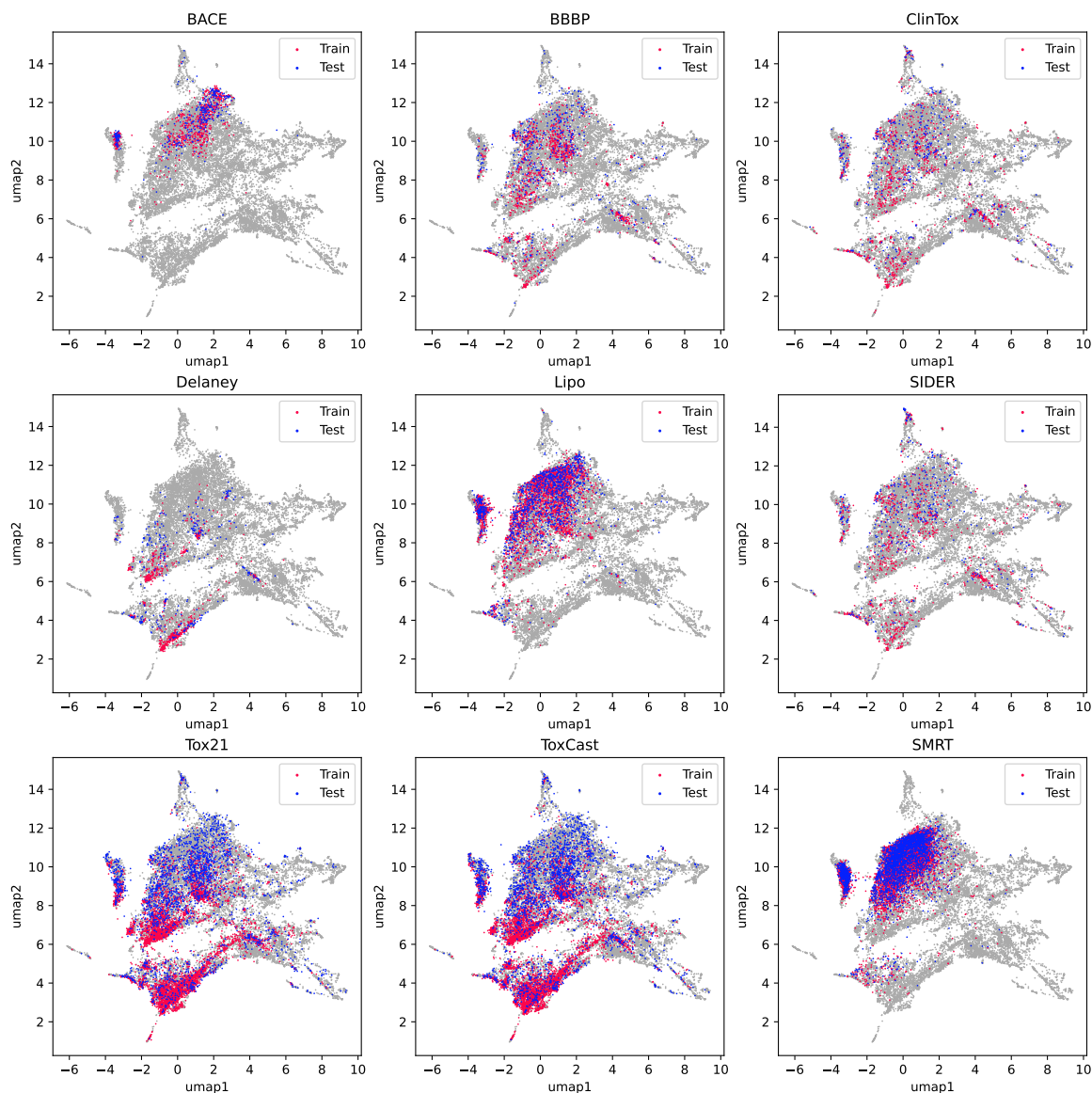
Supplementary Fig. 2. t-SNE and Minimum Spanning Tree visualizations. As alternative visualization methods to UMAP, t-SNE (a) and Minimum Spanning Tree computation (b) are applied to the pairwise myopic MCES distances of the 18,096 biomolecular structures (outlier lipid clusters excluded). The color-coding of compound classes from Fig. 1 is used. t-SNE computations are performed with scikit-learn³⁷, using the default settings (version 1.1.3). The Minimum Spanning Tree is computed using SciPy³⁸ version 1.10.1, based on Kruskal’s algorithm³⁹.



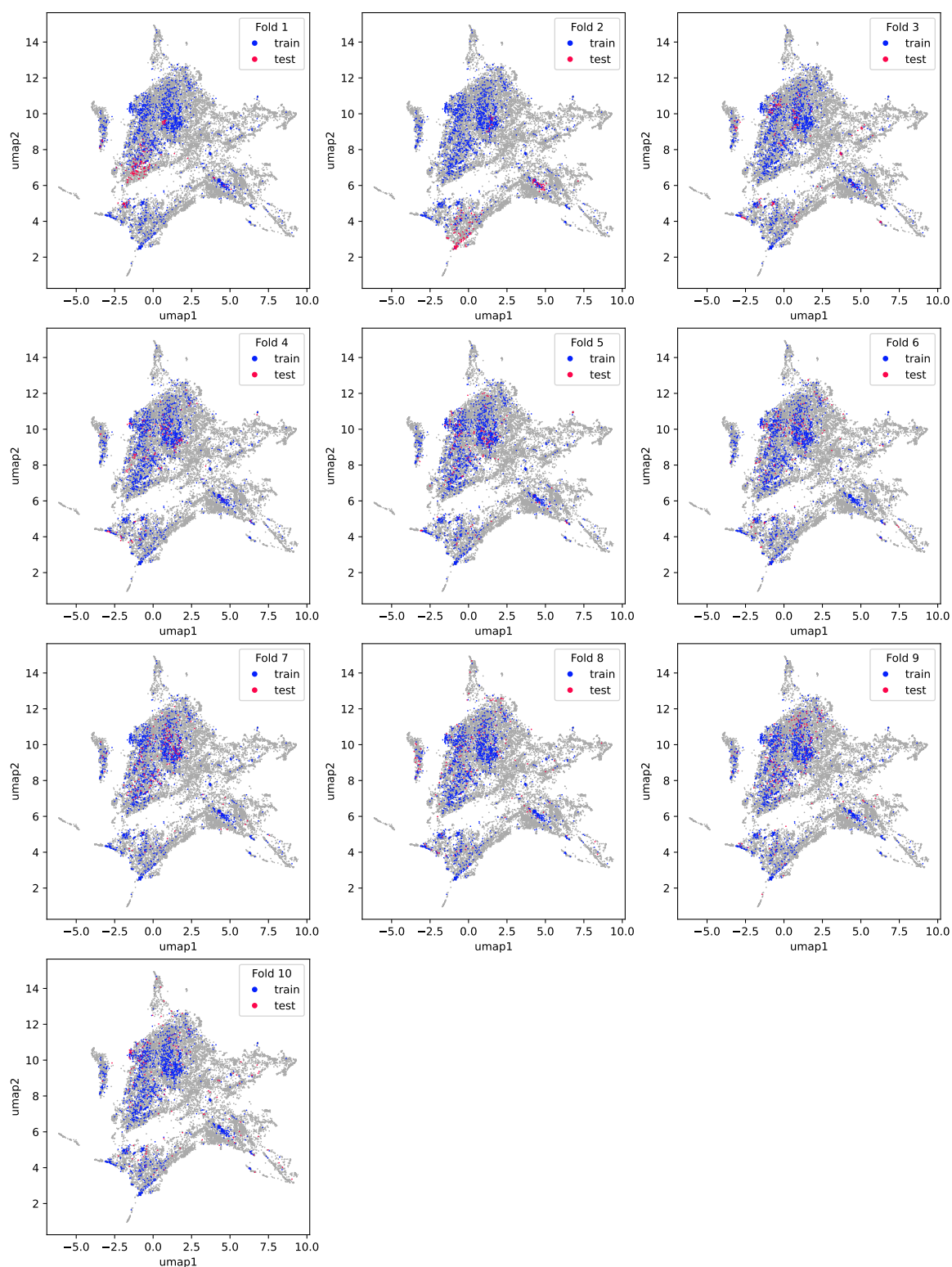
Supplementary Fig. 3. UMAP embeddings are clearly far from perfect. Both examples show a section of the MCES-based UMAP embedding of biomolecular structures (compare to Fig. 3). (a) This section of the embedding contains structures that do not share much similarity with one another, but with seemingly no better possible placement by UMAP. For instance, the structure labeled “3” is a member of the SMRT dataset mapped onto the UMAP space, with large myopic MCES distances (from 17.5 to 20.5) to all other labeled structures. (b) In this example, the structure labeled “9” clearly does not belong in this cluster of fatty acyls. A possible explanation for the strange placement might be, that the “outlier”-structure has approximately the same myopic MCES distance to all other labeled compounds (18 or 19 except for structure “13”, where the distance is 15); this is similar for the structure labeled “13” (albeit with a smaller distance: 10), which intuitively seems to be structurally much more closely related to the cluster. Numerous similar examples to those two cases can be found throughout the embedding.



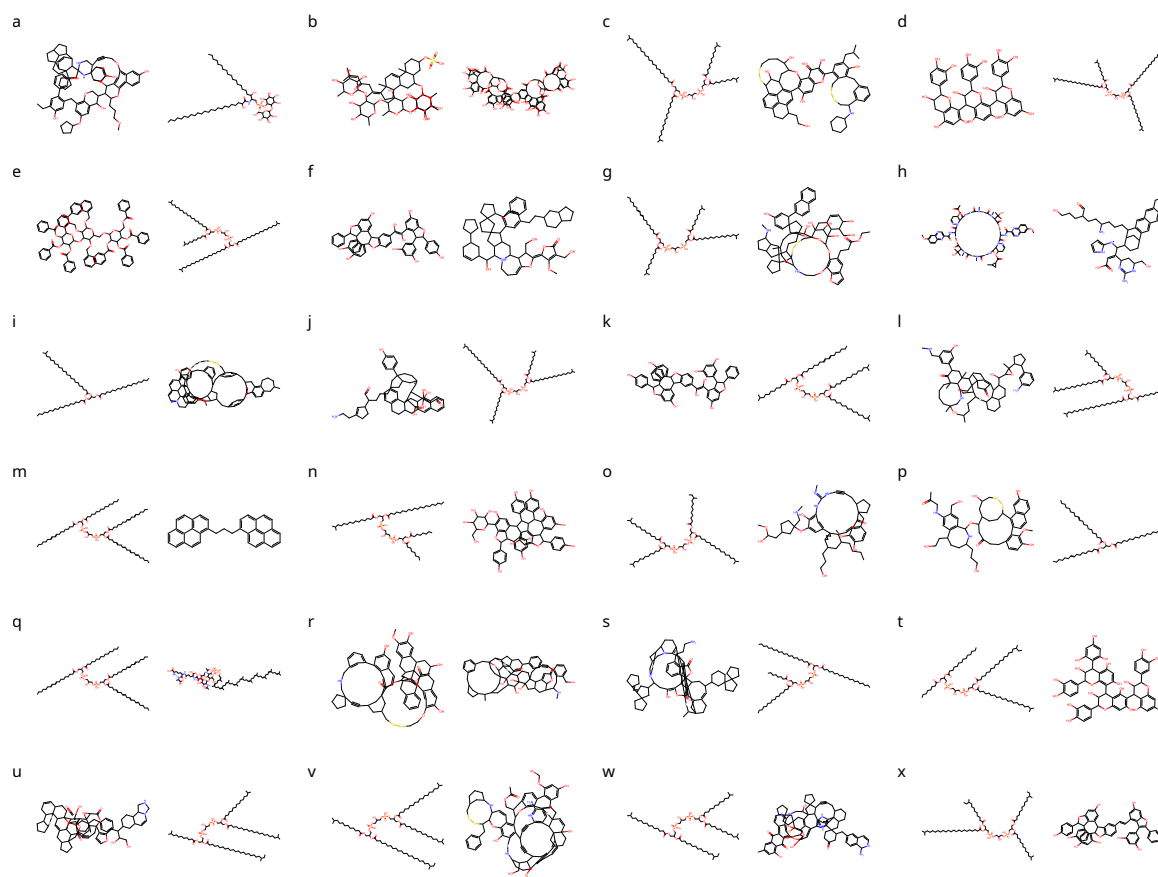
Supplementary Fig. 4. Maps of molecular structures in the MS/MS training dataset. Shown are both the total 18,848 molecular structures in the dataset (a) as well as those of the individual spectral libraries GNPS (b), MassBank (c), and NIST (d). A total of 36,944 molecular structures are shown in all plots combined. We stress that the assignment of molecular structures to the three libraries is not perfect: A molecular structure might be contained in more than one library but is displayed only for one. Hence, the plots *must not be overinterpreted to compare the coverage of the three libraries*. What we can learn from the three library plots is that each library covers the universe of biomolecular structures rather well. We use the same UMAP embedding as in Fig. 3. No molecular structures fall outside of this plot. Biomolecular structures are shown in light gray.



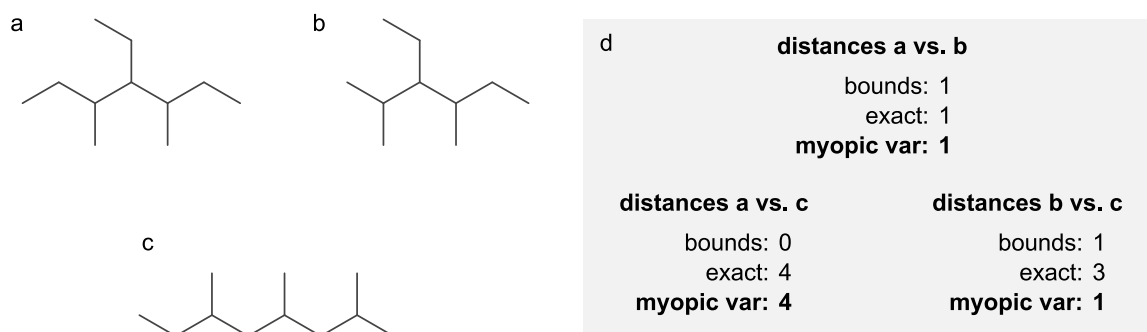
Supplementary Fig. 5. Map of nine scaffold splits. Molecular structures from the nine machine learning datasets from Fig. 5 were split according to Bemis-Murcko scaffolds. We used the `splits.ScaffoldSplitter.train_test_split` function from DeepChem (version 2.6.1) to generate a 70/30 split. For the datasets Delaney, SIDER, Tox21 and ToxCast, we observe that splits are clearly non-uniform. However, we do not observe a clear global separation between splits, for any datasets; instead, molecular structures of both train and test splits are spread across the whole covered structure space.



Supplementary Fig. 6. Map of the BBBP dataset with a 10-fold scaffold split. We generated a ten-fold scaffold split of the training dataset BBBP, compare to Supplementary Fig. 5. Here, we used the `splits.ScaffoldSplitter.k_fold_split` function from DeepChem. We observe that the split is clearly not uniform, in particular for the first two folds. Notably, this implementation of a cross-validation scaffold split sorts the scaffolds and, consequently, also splits based on size; this explains the increasing uniformity of the folds. We can observe small clusters where structurally similar molecular structures were sorted into the same fold.

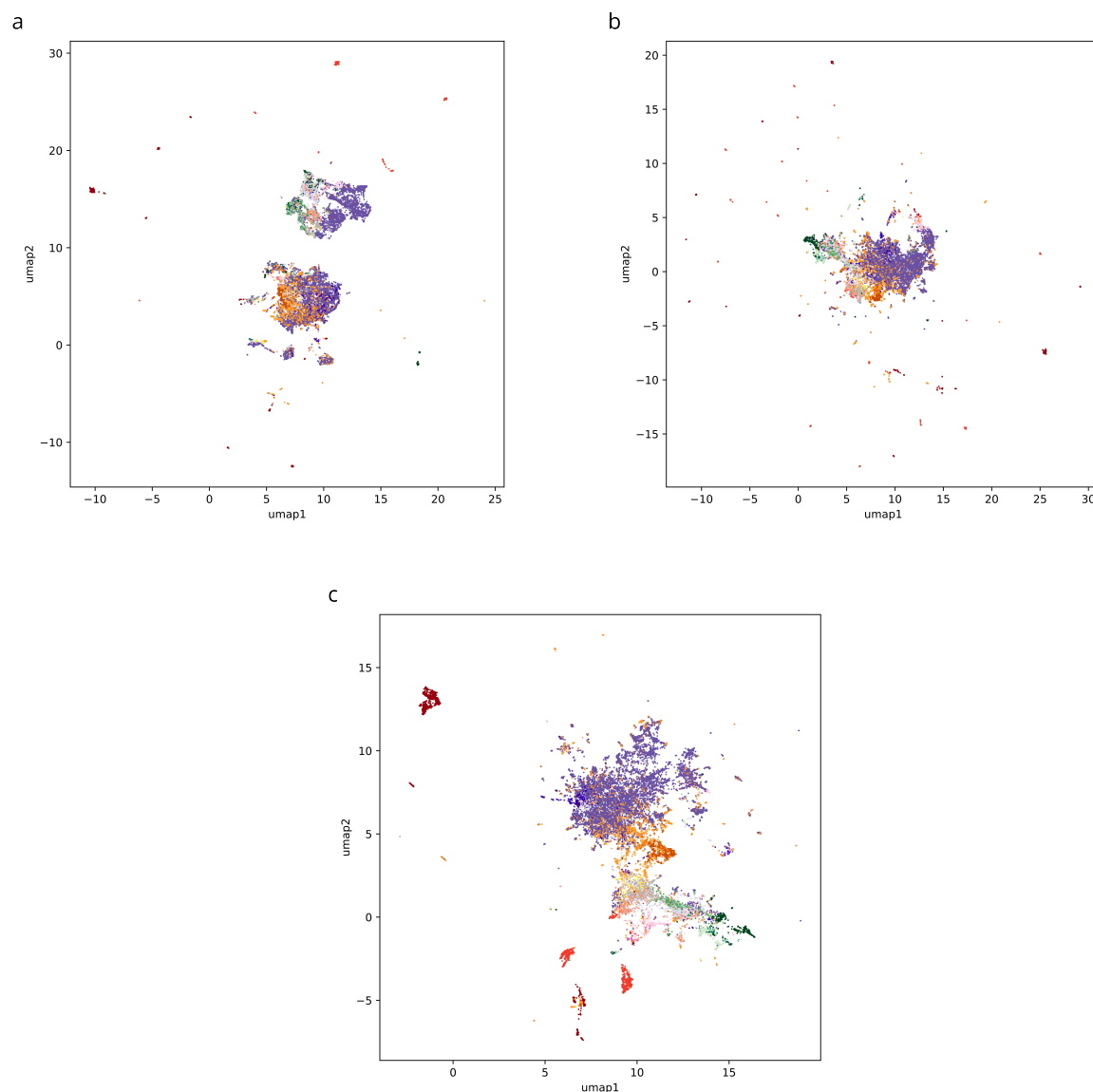


Supplementary Fig. 7. For 24 instances, exact MCES computations via Integer Linear Programming took more than four days. Forcing exact computation of the MCES distance for the subsampled 20,000 instances resulted in huge running times for a few instances, with 24 instances exceeding four days of wall clock time. Molecular structures of these instances are shown here (a–x). In all cases, structures are highly dissimilar and large, with myopic MCES distances (bounds only) ranging from 24.5 to 108. We may safely assume that for none of these instances, computing the exact MCES distance provides any useful information beyond that of the myopic MCES distance, which can be computed in fractions of a second.

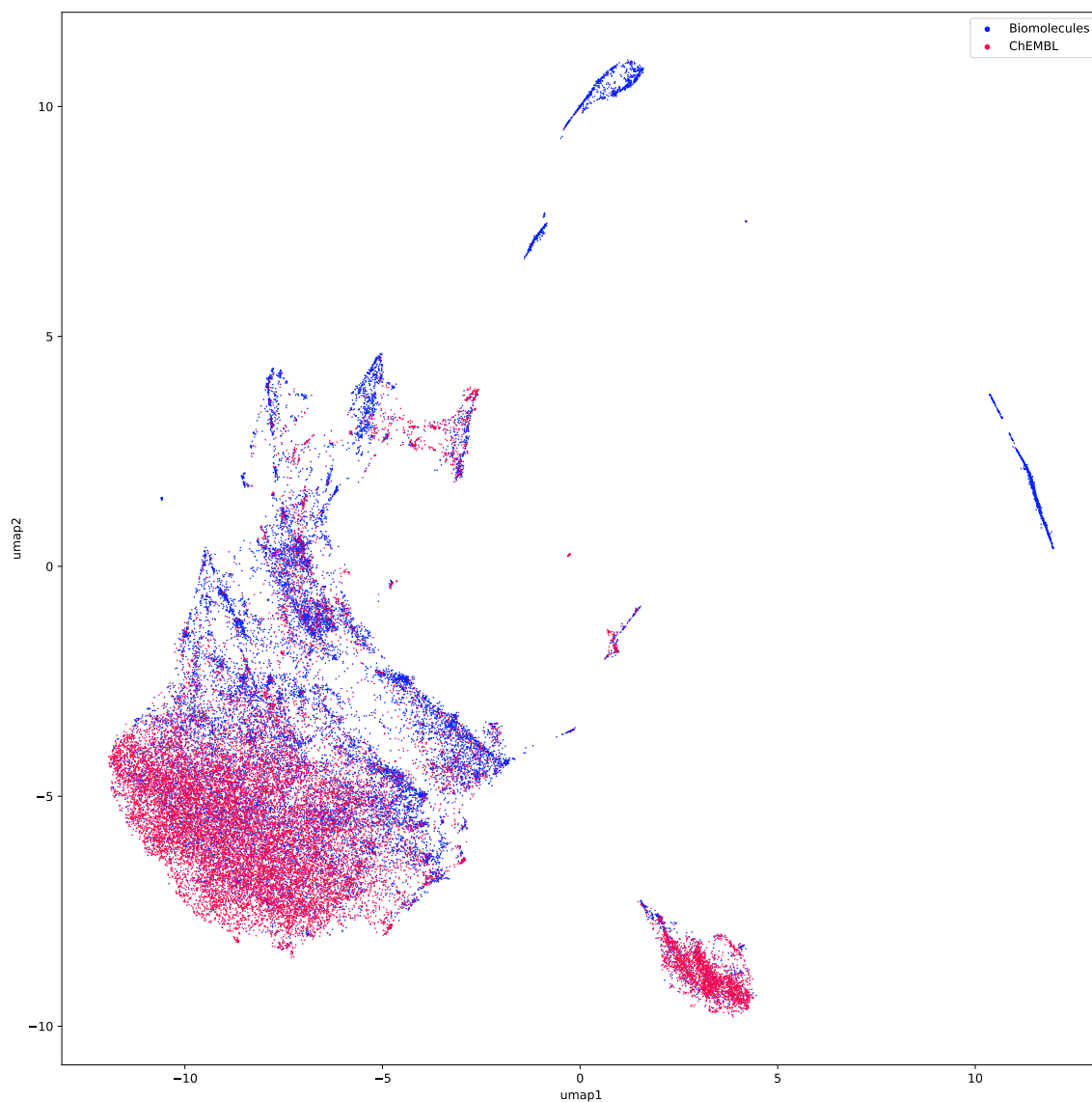


Supplementary Fig. 8. Double thresholding is needed to enforce the triangle inequality. Different from the myopic MCES distance used in the rest of the paper, we now consider a variant (“myopic var”) that does not employ double thresholding: We again first compute the MCES bounds, then decide whether the distance bound is below T . In this case, we compute the exact MCES distance. Different from the rest of the paper, we then use this exact MCES distance, *even if it is larger than T* . As shown in (d), the resulting distance violates the triangle inequality for molecular structures (a,b,c): For $T = 1$ the “myopic var” distance between a and c is 4, which is strictly larger than the sum of “myopic var” distances from a to b and from b to c. The problem is resolved for the myopic MCES distance as there, the distance between a and c is only $\min\{4, T\} = 1$.

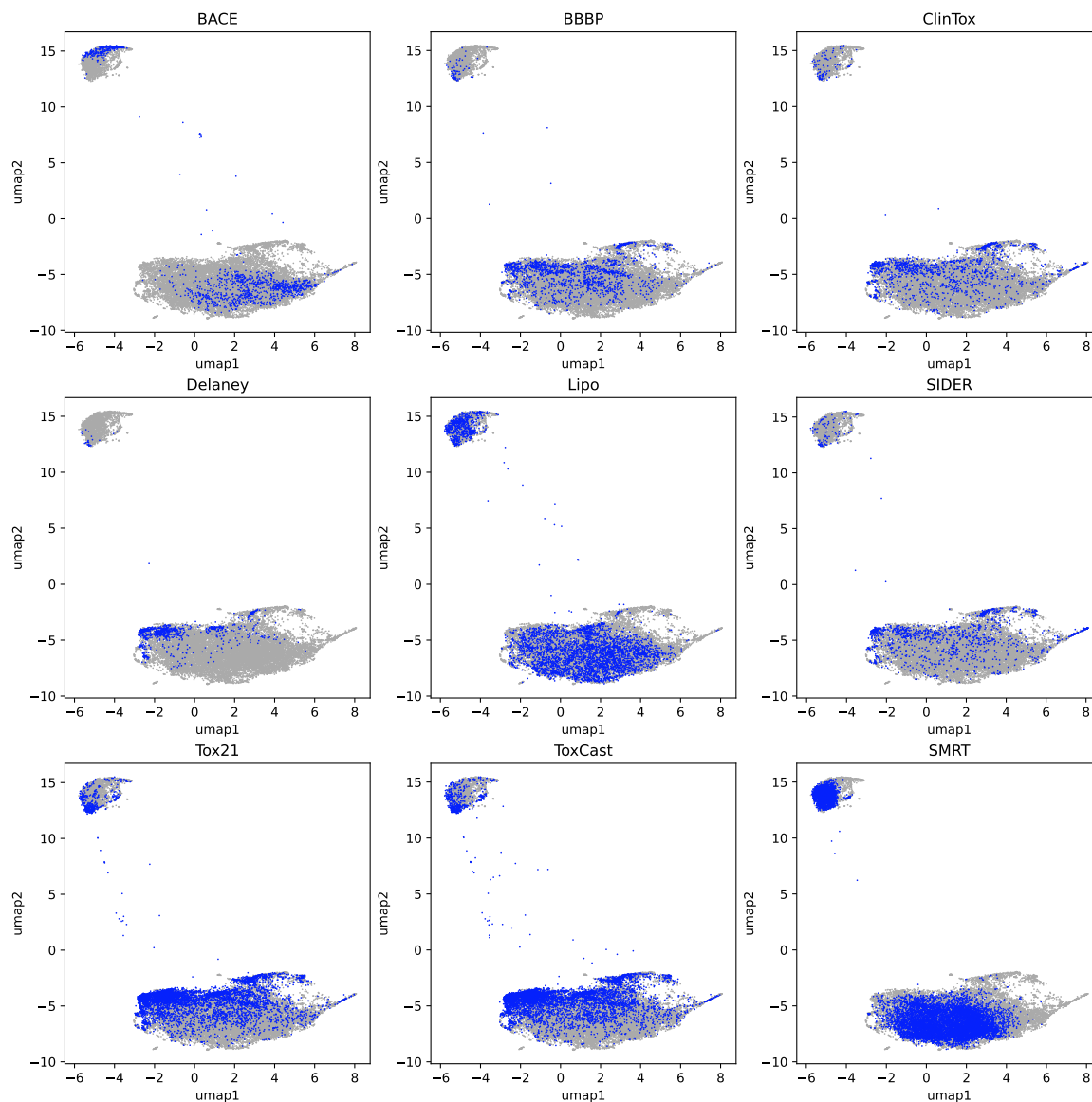
	biomol	BACE	BBBP	ClinTox	Delaney	Lipo	SIDER	Tox21	ToxCast	SMRT	M5MS		biomol	BACE	BBBP	ClinTox	Delaney	Lipo	SIDER	Tox21	ToxCast	SMRT	M5MS									
Fatty acid esters	14.7%	1	23	54	36	24	48	287	291	2,706	1,377	Benzene sulfonfyl compounds	2.0%	12	28	46	12	421	45	258	281	5,850	308									
	Carboxylic acids	12.3%	25	189	265	0	448	313	911	1,126	3,362		4,144	Aminopyrimidines and derivatives	1.9%	35	43	63	13	720	70	127	6,917	510								
	Polyols	12.0%	1	65	73	23	49	89	221	253	88		2,867		Organic amines	1.9%	2	32	63	16	45	48	30	348	11	631						
	Prenol lipids	11.3%	4	13	36	23	18	31	245	256	223		1,833			N-alkylpyrrolizanes	1.9%	4	107	56	0	223	45	132	143	4,488	269					
	Alcetals	10.7%	14	108	80	18	47	86	255	273	2,060		2,410				Beta hydroxy acids and derivatives	1.9%	1	11	20	0	8	29	38	67	8	489				
	Carbohydrates and carbohydrate derivatives	10.2%	0	64	34	18	33	91	207	233	1,447		1,601					Primary carboxylic acids	1.9%	8	33	45	5	184	60	120	113	66	142			
	Lactones	9.7%	2	40	57	15	38	47	184	197	355		2,162						Organic zwitterions	1.9%	6	33	25	15	34	56	150	359	454	343		
	Pyrans	9.4%	1	12	17	10	71	19	85	90	462		2,297							N-acyl-alpha amino acids	1.9%	8	6	26	0	25	48	57	88	99		
	Benzopyrans	9.2%	47	21	20	11	79	16	102	113	571		2,339								Quaternary ammonium salts	1.9%	3	36	68	0	17	46	269	291	7	418
	Triarylamines	9.0%	51	550	291	1	995	281	838	900	16,206		14,288									1,2-diols	1.8%	1	9	13	7	36	19	48	65	80
Aralkylamines	8.8%	595	371	257	7	767	270	637	793	5,191	1,601	Phenothiazines	1.8%										2	33	42	15	69	68	161	161	406	115
Cyclic alcohols and derivatives	8.8%	1	178	139	39	33	150	264	288	229	2,002		Oxoesters	1.8%									0	122	84	41	12	90	152	157	42	499
1-benzopyrans	8.8%	47	21	20	11	79	16	100	110	548	2,255			Carbamoyl acids and derivatives	1.8%								8	57	45	26	113	49	122	123	576	235
Glycosyl compounds	8.4%	0	51	55	11	26	62	141	155	77	2,164				Hemiacetals	1.8%							0	8	20	4	1	20	40	35	8	242
Dicarboxylic acids and derivatives	8.2%	6	118	118	42	134	114	111	237	508	1,470					Carbanion esters	1.8%						3	37	57	26	13	60	120	121	576	235
Tertiary alcohols	7.5%	5	170	118	42	92	117	259	318	379	556						Carboxylic acid imides	1.7%					0	68	33	54	55	29	104	108	2,105	267
Pyranones and derivatives	7.2%	0	10	12	9	70	9	70	78	379	1,946							Pyrazoles	1.7%				7	36	23	13	6	351	17	65	7,997	263
1-hydroxy-4-unsubstituted benzenoids	7.2%	3	44	45	25	83	43	251	264	275	1,852								Nitrobenzenes	1.7%			5	7	13	42	33	14	231	247	471	221
Aryl ketones	7.1%	11	91	58	11	139	49	294	317	1,227	1,546									Ether thioethers	1.7%		9	102	48	11	201	29	143	156	4,310	199
Monosaccharides	7.0%	0	49	58	9	203	66	145	171	182	4,174										Vinylglyoxal	1.7%	0	4	7	0	2	2	22	22	22	141
Organic phosphoric acids and derivatives	6.9%	0	3	19	6	2	22	92	99	1		5-hydroxylflavonoids										1.7%	0	1	0	2	2	2	14	15	39	515
O-glycosyl compounds	6.8%	0	44	42	8	1	43	102	107	27	1,883		Cinnamic acids and derivatives									1.7%	2	3	6	1	9	4	41	40	63	549
Phosphate esters	6.8%	0	3	16	6	2	17	84	86	8	439			Disaccharides								1.6%	0	9	11	4	0	15	33	36	4	519
Alkyl phosphates	6.6%	0	0	3	13	5	2	14	67	71	422				Benzylamines							1.6%	230	23	26	7	229	20	84	292	2,168	191
Pyrides	6.6%	129	66	186	5	376	85	170	186	9,364	3,312					Dihydrofurans						1.6%	7	2	1	0	2	2	2	2	63	263
Cyclic ketones	6.3%	2	168	121	50	51	125	285	312	589	1,051						Pyridinones					1.6%	10	7	10	1	131	8	20	22	2,162	241
Vinyllogous acids	6.0%	0	51	41	10	57	37	179	203	330	1,465							Dicarboximides				1.6%	0	66	33	52	45	28	90	94	2,032	246
Piperidines	5.9%	48	241	133	7	718	129	991	422	20,664	1,171								Phenethylamines			1.6%	455	34	57	0	189	85	113	125	189	570
Indoles and derivatives	5.6%	79	74	66	60	25	121	173	182	420	1,306									Fatty acyl glycosides		1.6%	2	3	2	0	2	2	2	2	2	2
Glycophospholipids	5.5%	0	0	2	0	0	1	5	6	0	217										Dicarboxyphenylamines	1.6%	5	19	27	6	280	24	78	84	4,645	271
Dialkyl phosphates	5.5%	0	0	3	2	0	5	28	30	0	250	Thiophenes										1.6%	52	28	28	2	246	33	83	91	7,254	171
Diarylamines	5.4%	530	139	138	1	404	147	342	385	4,459	831		Alpha,beta-unsaturated carbonyl compounds									1.6%	1	14	17	16	17	78	84	1,244	354	
Tricarboxylic acids and derivatives	5.3%	2	77	12	1	1	18	73	14	3	368			Benzodioxoles								1.6%	14	9	9	3	40	12	50	52	1,856	262
Alpha,beta-unsaturated carboxylic esters	5.2%	7	7	11	27	1	24	191	170	2,572	1,071				Bromocyclohexanes							1.6%	22	42	28	2	22	42	28	22	42	28
Enoate esters	5.2%	7	12	27	7	11	24	189	196	2,572	1,071					Thiazoles						1.6%	25	42	29	7	276	33	122	134	5,998	195
Pyridones	5.0%	92	113	85	7	193	121	215	247	10,646	867						Terpene lactones					1.5%	0	0	0	1	0	0	5	4	14	253
Tetracarboxylic acids and derivatives	4.9%	1	3	7	0	0	10	24	27	2	162							Phenanthrenes and derivatives				1.5%	0	41	26	26	25	36	72	84	31	242
Steroids and steroid derivatives	4.9%	0	129	119	48	14	154	40	6	130	110								Guanidines			1.5%	255	20	36	78	5	64	78	84	78	230
Tetrahydrofurans	4.5%	0	22	41	12	15	43	138	146	831	831									Phenylmethanamines		1.5%	230	22	23	1	190	18	74	80	1,794	162
Glycerolipids	4.5%	0	2	3	2	1	3	24	20	1	145										Flavones	1.5%	0	1	1	1	5	2	14	15	39	600
Indoles	4.4%	70	53	49	2	228	68	138	148	2,713	1,811	Oligosaccharides										1.5%	0	3	7	3	0	6	20	19	1	311
Naphthalenes	4.4%	3	44	39	46	88	41	207	234	421	740		Terpene glycosides									1.5%	0	0	0	0	0	0	7	9	0	456
Monosaccharamines	4.3%	42	96	117	50	135	117	281	334	1,417	1,441			Pyridinecarboxylic acids and derivatives								1.5%	84	41	241	246	241	246	241	246	241	3,804
Glycophosphoglycerophosphoglycerols	4.2%	0	0	0	0	0	0	0	0	0	0				Oxazones							1.4%	6	23	15	0	285	25	84	92	4,669	163
Cardiolipins	4.1%	0	0	0	0	0	0	0	0	0	0					Sugar acids and derivatives						1.4%	0	1	5	0	0	8	9	23	31	292
Organic salts	4.1%	7	77	81	11	46	133	441	1,623	344	593						Alkyl chlorides					1.4%	0	60	33	72	10	44	274	298	32	159
Furans	4.0%	106	26	43	8	81	16	84	88	4,250	451							Tetrals				1.4%	13	80	45	9	12	53	91	97	224	264
Substituted pyridines	3.9%	106	26	43	8	81	16	84	88	4,250	451								Ketals			1.4%	0	33	3	0	0	0	0	0	0	0
N-acyl-alpha amino acids and derivatives	3.9%	55	121	118	0	106	139	164	201	945	1,358									Catechols		1.4%	0	9	16	2	6	13	37	45	24	381
Chromones	3.9%	0	3	4	4	50	6	33	34	207	1,134										4-hydroxylflavonoids	1.4%	0	0	0	2	4	1	7	8	18	428
Aryl alkyl ketones	3.8%	9	63	30	5	83	23	168	177	848	833	Alkylaryliethers										1.4%	9	37	23	10	143	12	71	82	3,688	132
Triaryl glycol ethers	3.8%	0	0	0	0	1	4	0	3	0	21		Allyl-pyren-10-yl									1.4%	0	1	1	2	98	40	10	10	10	277
Methyl esters	3.8%	4	19	23	18	30	28	167	179	1,737	783			Morpholines								1.4%	6	21	15	0	285	24	82	89	4,628	156
Alpha amino acid amides	3.8%	65	55	89	4	149	121	128	138	6,551	978				Amphetamines and derivatives							1.4%	454	21	47	0	89	76	80	88	99	468
Alkanolamines	3.7%	524	147	134	6	215	135	317	356	743	530					Diarylthioethers						1.4%	12	116	63	13	37	62	123	152	691	287
Triacylglycerols	3.6%	0	0	0	1	0	1	4	3	0	21						Oxepanes					1.4%	0	1	3	1	0	1	15	15	14	203
Amino acids	3.4%	7	8	127	56	246	156	247	269	699	1,220							Furanones				1.4%	0	1	4	2	2	1	25	25	25	25
Coumarins and derivatives	3.4%	7	6	7	21	3	26	31	151	874	1,874								Isoflavonoids			1.3%	0	0	0	1	25	0	9	9	20	314
Aldehydes	3.4%	1	17	6	16	10	10	136	153	94	527									Alpha,beta-unsaturated ketones		1.3%	1	14	17	6	16	17	56	61	1,244	322
Triterpenoids	3.3%	0	0	7	0	0	5	22	25	4	505										Methoxanilines	1.3%	5	13	6	3	150	10	61	70	5,481	115
Organotinorimides	3.3%	38	384	13	49	95	14	203	128	2,943	233	Thiazoles										1.3%	5	18	16	21	9	65	25	81	68	5,773
Sulfenyl compounds	3.2%	11	66	53	49	212	58	250	272	14,691	292		Benzodiazines									1.3%	5	10	22	6	250	21	61	57	3,821	220
Benzenofurans	3.2%	1	11	12	16	6	23	11	53	61	406			322								2-heteroaryl carboxamides	1.3%	56	11	20	1	245	22	49	48	10,889
Diazinones	3.1%	8	167	85	34	406	69	188	206	10,959	436			Vinyl halides	1.3%								7	24	31	31	112	21	152	170	20,463	183
Hydroxysteroids	3.0%	0	122	81	37	9	91	158	169	43	804				Imidazopyridines	1.2%							0	28	28	10	96	32	70	73</		



Supplementary Fig. 10. UMAP embedding using Tanimoto coefficients. We use and display all 19,994 biomolecular structures. Jaccard distances (one minus Tanimoto coefficients) were computed from MACCS fingerprints (a), ECFP4 (Morgan) fingerprints (b), and MAP4 fingerprints (c). The resulting distances are used to create corresponding UMAP embeddings. Compound classes are color-coded as in Fig. 1. Different from the MCES-based plots, for both types of fingerprints we observe numerous outlier clusters, as well as singleton outliers. Certain compound classes are separated in distant clusters. Compare to Fig. 1 where outliers were not removed.



Supplementary Fig. 11. Joint UMAP embedding of ChEMBL and biomolecular structures. We uniformly subsample 20,000 structures from ChEMBL “small molecules”, downloaded on July 31st, 2024. For each structure that could not be standardized, not processed by rdkit, or consisted of single ions, another structure was drawn. UMAP embeddings were computed from the union of 20,000 ChEMBL structures and 19,994 biomolecular structures. We clearly observe different structure distributions for ChEMBL and the biomolecular structures. Compare to Fig. 1 where only biomolecular structures are shown, and to Supplementary Fig. 12 where ChEMBL structures are used as the background distribution.



Supplementary Fig. 12. Maps of nine training datasets using ChEMBL as the background distribution. The 20,000 structures from ChEMBL (see Supplementary Fig. 11) were used to derive the UMAP embedding, and show in light gray. The same training datasets as in Fig. 5 are shown, see there for details. Notably, the background ChEMBL distribution is clearly separated between molecular structures that contain a sulfonyl group (upper-left cluster, 99.4% of molecular structures contain a sulfonyl group) and those that do not (rest, 0.5% of molecular structures contain a sulfonyl group). Similarly, 77.4% of molecular structures in the upper-left cluster contain a sulfonamide group, vs. 0.2% of the other molecular structures. Consistent with our analysis for biomolecular structures, most datasets show areas of missing coverage, with Tox21 and Toxcast providing the best coverage. Notably, the Lipo and SMRT datasets appear to be more uniform than in our previous analyses.

Supplementary Table 1. Structure databases that contribute to the set of biomolecular structures used herein.

For each database the total number of molecular structures (total) and the number of structures that are absent of all preceding databases (additional) is shown. A subset of structures from PubChem (PC) which are annotated with categories of biological relevance is included.

Database	total	additional	Ref.	URL
KEGG	14,362	14,362	[82]	https://www.kegg.jp
ChEBI	56,219	42,621	[83]	https://www.ebi.ac.uk/chebi
HMDB	95,973	85,942	[84]	https://www.hmdb.ca
YMDB	1,864	135	[85]	http://www.ymdb.ca
PlantCyc	5,503	1,072	[86]	https://plantcyc.org/
MetaCyc	13,059	3,175	[87]	https://metacyc.org
KNAPSAcK	42,168	31,967	[88]	http://www.knapsackfamily.com/KNAPSAcK/
UNPD	161,976	112,057	[89]	http://pkuxxj.pku.edu.cn/UNPD [offline]
MaConDa	91	31	[90]	https://www.maconda.bham.ac.uk/
HSDB	4,319	1,080	[91]	ftp://ftp.nlm.nih.gov/nlmda/.hsdblease/
Super Natural II	227,931	95,291	[92]	http://bioinformatics.charite.de/supernatural
COCONUT	380,169	104,404	[93]	https://coconut.naturalproducts.net
NORMAN suspect list	64,121	47,251	[94]	https://www.norman-network.com/nds/SLE/
MeSH-annotated PC	82,337	37,523	[66, 95]	https://pubchem.ncbi.nlm.nih.gov
PC "bio and metabolites"	138,753	30,500	[66]	https://pubchem.ncbi.nlm.nih.gov/classification
PC "drug"	7,649	651	[66]	https://pubchem.ncbi.nlm.nih.gov/classification
PC "safety and toxic"	148,962	109,792	[66]	https://pubchem.ncbi.nlm.nih.gov/classification
PC "food"	4,305	243	[66]	https://pubchem.ncbi.nlm.nih.gov/classification

Supplementary Table 2. Molecular structure dataset sizes. Number of molecular structures in the considered molecular structure training datasets. Some molecular structures failed the standardization via PubChem and were discarded, see column "After standardization". Few molecular structures resulted in failed MCES computations and were also discarded, see column "After MCES computations". For the *SMRT* dataset we uniformly subsampled 10,000 standardized molecular structures. The number "Total in plot" refers to the union of the molecular structures in the dataset and the 18,096 biomolecular structures.

Dataset	Number of compounds	After standardization	After MCES computations	Total in plot	ref.
BACE	1513	1194	1194	19,290	[41]
BBBP	2039	1662	1662	19,758	[42]
ClinTox	1478	1476	1475	19,571	[43]
Delaney	1128	1128	1127	19,223	[44]
Lipo	4200	4200	4200	22,296	[45]
SIDER	1427	1420	1389	19,485	[46, 47]
SMRT	80,037	10,000	10,000	28,096	[48]
ToxCast	8576	8517	8502	26,598	[49]
Tox21	7831	7783	7763	25,859	[50, 51]
MS/MS	18,848	18,848	18,848	36,944	[52, 53]