

Additional file 1

How to Approach Machine Learning-based Prediction of Drug/Compound-Target Interactions

Heval Atas Guvenilir^{1,2}, Tunca Doğan^{1,3,4*}

¹ Biological Data Science Laboratory, Department of Computer Engineering, Hacettepe University, Ankara, Turkey

² Department of Health Informatics, Graduate School of Informatics, METU, Ankara, Turkey

³ Institute of Informatics, Hacettepe University, Ankara, Turkey

⁴ Department of Bioinformatics, Graduate School of Health Sciences, Hacettepe University, Ankara, Turkey

* To whom correspondence should be addressed, e-mail: tuncadogan@gmail.com

Supplementary Tables

Table S1. Model performance scores (in terms of MCC) in the small-scale analysis (on the compound-centric datasets) for; **(a)** random forest, and **(b)** SVM models. The 3 best performances for each dataset are shown in bold font. For the results based on other performance metrics please see the supplementary spreadsheets attached as external files named: “small-scale_rf_overall_test_results.xlsx” and “small-scale_svm_overall_test_results.xlsx”.

(a)

Model	ChEMBL id (only the numeric part) of the center compound of each compound cluster										Mean	Standard error
	44	50	83	91	104	633	808	116438	295698			
aac	0.283	0.112	0.359	0.238	0.413	0.199	0.227	0.099	0.195	0.236	0.035	
aac_pssm	0.226	0.278	0.395	0.432	0.170	0.212	0.399	0.153	0.270	0.282	0.035	
aadp_pssm	0.166	0.211	0.396	0.453	0.177	0.347	0.330	0.240	0.380	0.300	0.035	
aatp_pssm	0.231	0.312	0.432	0.477	0.332	0.280	0.398	0.254	0.387	0.345	0.028	
ab_pssm	0.237	0.231	0.397	0.408	0.157	0.323	0.487	0.206	0.364	0.312	0.037	
apaac	0.263	0.262	0.463	0.346	0.303	0.514	0.422	0.293	0.384	0.361	0.030	
cksaagp	0.243	0.034	0.390	0.333	0.222	0.323	0.261	0.120	0.260	0.243	0.037	
cksaap	0.324	0.268	0.510	0.410	0.297	0.489	0.405	0.143	0.431	0.364	0.039	
ctdc	0.117	-0.035	0.361	0.213	0.332	0.280	0.223	0.185	0.109	0.199	0.041	
ctdd	0.176	0.209	0.446	0.204	0.142	0.512	0.471	0.185	0.390	0.304	0.049	
ctdt	0.212	0.033	0.309	0.201	0.363	0.169	0.379	0.130	0.301	0.233	0.038	
ctriad	0.231	0.179	0.355	0.230	0.243	0.449	0.285	0.280	0.396	0.294	0.029	
d_fpssm	0.364	0.288	0.429	0.371	0.398	0.518	0.492	0.250	0.345	0.384	0.029	
dde	0.302	0.293	0.528	0.492	0.356	0.512	0.536	0.174	0.446	0.404	0.043	
dp_pssm	0.279	0.376	0.369	0.387	0.297	0.316	0.325	0.278	0.396	0.336	0.016	
dpc	0.318	0.170	0.467	0.482	0.197	0.453	0.395	0.121	0.412	0.335	0.046	
dpc_pssm	0.227	0.210	0.396	0.491	0.198	0.369	0.360	0.233	0.433	0.324	0.036	
edp_pssm	0.144	0.172	0.375	0.388	0.164	0.141	0.358	0.152	0.273	0.241	0.036	
eedp_pssm	0.186	0.225	0.419	0.433	0.164	0.310	0.336	0.022	0.262	0.262	0.043	
gaac	0.089	-0.081	0.266	0.020	0.088	0.200	0.135	-0.057	0.129	0.088	0.038	
gdpc	0.314	0.045	0.311	0.292	0.253	0.310	0.262	0.076	0.232	0.233	0.034	
geary	0.337	0.182	0.390	0.394	0.294	0.339	0.360	0.345	0.322	0.329	0.021	
gtpc	0.304	0.134	0.344	0.198	0.122	0.442	0.339	0.184	0.404	0.275	0.039	
k-sep_pssm	0.330	0.312	0.561	0.326	0.452	0.506	0.313	0.209	0.355	0.374	0.037	
ksctriad	0.295	0.190	0.394	0.268	0.316	0.448	0.315	0.172	0.387	0.310	0.031	
medp_pssm	0.167	0.164	0.417	0.431	0.216	0.309	0.375	0.023	0.285	0.265	0.045	
moran	0.325	0.236	0.367	0.391	0.264	0.412	0.415	0.247	0.377	0.337	0.024	
nmbroto	0.254	0.190	0.365	0.444	0.277	0.327	0.344	0.195	0.391	0.310	0.029	
paac	0.230	0.181	0.502	0.417	0.255	0.505	0.500	0.262	0.347	0.356	0.043	
pfam	0.325	0.286	0.486	0.432	0.430	0.203	0.273	0.132	0.399	0.329	0.039	
pse_pssm	0.209	0.267	0.438	0.327	0.386	0.228	0.340	0.314	0.396	0.323	0.026	
pssm_ac	0.371	0.379	0.393	0.426	0.324	0.578	0.451	0.210	0.478	0.401	0.034	
pssm_cc	0.293	0.276	0.401	0.413	0.302	0.483	0.400	0.274	0.398	0.360	0.025	
pssm_composition	0.220	0.221	0.455	0.451	0.333	0.373	0.507	0.142	0.437	0.349	0.043	
qso	0.229	0.247	0.410	0.224	0.378	0.433	0.291	0.141	0.348	0.300	0.033	
random200	0.071	0.080	-0.035	-0.033	-0.046	-0.100	-0.055	-0.111	0.054	-0.019	0.024	
rpm_pssm	0.289	0.132	0.401	0.471	0.440	0.410	0.358	0.189	0.422	0.346	0.039	
rpssm	0.055	0.197	0.355	0.456	0.187	0.340	0.418	0.087	0.338	0.270	0.048	
spmap	0.338	0.259	0.431	0.405	0.194	0.556	0.489	0.153	0.457	0.365	0.046	
taap	0.174	0.179	0.455	0.283	0.431	0.442	0.402	0.102	0.309	0.309	0.044	
tpc	0.245	0.268	0.476	0.499	0.270	0.529	0.528	0.124	0.448	0.376	0.050	
tpc_pssm	0.220	0.263	0.330	0.492	0.352	0.241	0.356	0.241	0.412	0.323	0.030	
tri-gram_pssm	0.350	0.349	0.541	0.362	0.426	0.545	0.386	0.274	0.352	0.398	0.030	

(b)

Model	ChEMBL id (only the numeric part) of the center compound of each compound cluster									Mean	Standard error
	44	50	83	91	104	633	808	116438	295698		
aac	0.122	0.049	0.183	0.102	0.243	0.279	0.091	0.161	0.136	0.152	0.025
aac_pssm	0.178	0.146	0.396	0.488	0.285	0.306	0.378	0.144	0.242	0.285	0.040
aadp_pssm	0.195	0.142	0.360	0.287	0.218	0.254	0.394	0.132	0.359	0.260	0.032
aatp_pssm	0.096	0.180	0.390	0.392	0.311	0.305	0.471	0.132	0.262	0.282	0.042
ab_pssm	0.296	0.311	0.385	0.372	0.201	0.470	0.370	0.195	0.295	0.322	0.030
apaac	0.159	0.292	0.508	0.430	0.358	0.559	0.438	0.200	0.300	0.360	0.045
cksaagp	0.215	0.045	0.238	0.248	0.240	0.216	0.146	0.229	0.283	0.207	0.024
cksaap	0.329	0.260	0.336	0.380	0.433	0.316	0.491	0.164	0.506	0.357	0.036
ctdc	0.197	-0.055	0.280	0.225	0.353	0.158	0.134	0.079	0.121	0.166	0.039
ctdd	0.151	0.226	0.375	0.357	0.225	0.310	0.414	0.237	0.024	0.258	0.041
ctdt	0.229	-0.013	0.233	0.099	0.498	0.300	0.284	0.096	0.053	0.198	0.052
ctriad	0.171	0.174	0.293	0.215	0.408	0.329	0.233	0.164	0.271	0.251	0.027
d_fpssm	0.205	0.123	0.242	0.388	0.025	0.140	0.394	0.033	0.362	0.212	0.048
dde	0.307	0.301	0.360	0.283	0.389	0.386	0.358	0.238	0.409	0.337	0.019
dp_pssm	0.228	0.434	0.370	0.363	0.229	0.329	0.335	0.186	0.313	0.310	0.027
dpc	0.202	0.106	0.308	0.291	0.236	0.333	0.263	0.193	0.337	0.252	0.025
dpc_pssm	0.195	0.142	0.418	0.287	0.218	0.254	0.394	0.132	0.358	0.266	0.035
edp_pssm	0.131	0.253	0.391	0.331	0.215	0.256	0.262	0.004	0.334	0.242	0.039
eedp_pssm	0.220	0.248	0.373	0.415	0.080	0.266	0.491	0.059	0.283	0.271	0.048
gaac	0.151	0.067	0.172	0.120	-0.089	0.230	0.111	0.065	0.206	0.115	0.032
gdpc	0.088	0.136	0.273	0.319	0.055	0.134	0.018	0.139	0.234	0.155	0.034
geary	0.277	0.201	0.357	0.327	0.308	0.288	0.280	0.195	0.358	0.288	0.020
gtpc	0.140	0.056	0.135	0.060	0.105	0.212	0.243	0.259	0.362	0.175	0.034
k-sep_pssm	0.277	0.347	0.513	0.288	0.418	0.474	0.456	0.241	0.312	0.370	0.033
ksctriad	0.124	0.231	0.406	0.277	0.031	0.402	0.338	0.148	0.262	0.247	0.043
medp_pssm	0.198	0.270	0.365	0.415	0.080	0.280	0.491	0.091	0.271	0.274	0.046
moran	0.264	0.215	0.345	0.347	0.181	0.315	0.384	0.184	0.303	0.282	0.025
nmbroto	0.231	0.228	0.354	0.289	0.439	0.282	0.386	0.236	0.258	0.300	0.025
paac	0.257	0.268	0.480	0.314	0.246	0.528	0.443	0.167	0.338	0.338	0.040
pfam	0.288	0.328	0.478	0.452	0.327	0.353	0.432	0.367	0.398	0.380	0.021
pse_pssm	0.248	0.280	0.367	0.334	0.144	0.348	0.342	0.176	0.343	0.287	0.027
pssm_ac	0.331	0.353	0.290	0.392	0.351	0.423	0.357	0.164	0.428	0.343	0.027
pssm_cc	0.294	0.368	0.484	0.311	0.517	0.446	0.373	0.192	0.349	0.370	0.034
pssm_composition	0.154	0.204	0.386	0.411	0.257	0.419	0.376	0.114	0.348	0.297	0.039
qso	0.220	0.078	0.249	0.211	0.346	0.268	0.073	0.213	0.113	0.197	0.031
random200	-0.044	-0.054	-0.110	-0.206	-0.046	-0.083	0.248	-0.140	0.205	-0.026	0.051
rpm_pssm	0.331	0.177	0.376	0.426	0.300	0.430	0.435	0.062	0.377	0.324	0.042
rpssm	0.133	0.312	0.260	0.431	0.249	0.305	0.361	0.076	0.361	0.276	0.038
spmap	0.342	0.260	0.486	0.254	0.249	0.492	0.394	0.026	0.370	0.319	0.048
taap	0.211	0.284	0.456	0.325	0.210	0.408	0.337	0.167	0.289	0.299	0.032
tpc	0.356	0.218	0.417	0.354	0.300	0.419	0.269	0.239	0.418	0.332	0.026
tpc_pssm	0.111	0.260	0.285	0.232	0.237	0.276	0.264	0.209	0.425	0.256	0.027
tri-gram_pssm	0.340	0.309	0.487	0.368	0.444	0.485	0.490	0.284	0.331	0.393	0.028

Table S2. Model performance scores in the medium-scale analysis (on the mDavis dataset). The best performance for each metric is shown in bold font.

Model	RMSE	Spearman	F1-score	MCC
seqvec	0.794	0.571	0.530	0.445
k-sep_pssm	0.817	0.545	0.531	0.434
unirep1900	0.823	0.541	0.510	0.418
apaac	0.831	0.532	0.519	0.418
unirep5700	0.831	0.531	0.506	0.412
transformer-avg	0.839	0.519	0.508	0.410
transformer-pool	0.840	0.515	0.506	0.412
qso	0.843	0.519	0.486	0.384
dde	0.845	0.508	0.480	0.384
geary	0.847	0.519	0.473	0.377
protvec	0.850	0.503	0.506	0.403
ctdd	0.851	0.503	0.484	0.376
ctriad	0.852	0.508	0.476	0.387
pfam	0.854	0.497	0.538	0.410
taap	0.863	0.492	0.467	0.349
spmap	0.871	0.491	0.477	0.362
random200	0.957	0.403	0.368	0.251
random200_random-ecfp4	0.968	0.388	0.346	0.235

Table S3. Model performance scores (in terms of the median corrected MCC) in the large-scale analysis on the protein family specific datasets of; **(a)** the random-split, **(b)** dissimilar-compound-split, and **(c)** the fully-dissimilar-split. The 3 best performances for each protein family are shown in bold font (ran200_ran-ecfp4: random200_random-ecfp4, only-ran-ecfp4: only-random-ecfp4). For the results based on other performance metrics please see the supplementary spreadsheet attached as an external file named: “large-scale_overall_test_results.xlsx”.

(a)

Random-split	epigenetic-regulators	hydrolases	ion-channels	membrane-receptors	other-enzymes	oxidoreduc-tases	proteases	transcripti-on-factors	transferase-s	transporter-s
apaac	0.745	0.755	0.697	0.689	0.754	0.692	0.735	0.714	0.696	0.728
ctdd	0.741	0.747	0.700	0.686	0.757	0.694	0.730	0.711	0.694	0.732
ctriad	0.734	0.749	0.701	0.686	0.752	0.694	0.731	0.706	0.694	0.726
dde	0.741	0.756	0.703	0.689	0.754	0.692	0.735	0.709	0.691	0.722
geary	0.733	0.754	0.701	0.694	0.752	0.681	0.735	0.721	0.696	0.728
k-sep_pssm	0.757	0.749	0.709	0.688	0.754	0.690	0.735	0.706	0.704	0.720
pfam	0.678	0.694	0.679	0.458	0.609	0.561	0.635	0.645	0.628	0.622
qso	0.734	0.757	0.700	0.685	0.754	0.690	0.733	0.704	0.691	0.728
random200	0.728	0.751	0.687	0.680	0.746	0.685	0.734	0.709	0.687	0.726
spmap	0.737	0.748	0.697	0.682	0.757	0.680	0.728	0.709	0.683	0.720
taap	0.760	0.747	0.712	0.687	0.750	0.693	0.736	0.721	0.700	0.730
protvec	0.741	0.742	0.703	0.693	0.758	0.696	0.733	0.714	0.696	0.726
seqvec	0.745	0.749	0.699	0.690	0.757	0.678	0.728	0.709	0.700	0.724
transformer-avg	0.736	0.748	0.707	0.691	0.760	0.681	0.734	0.701	0.702	0.718
transformer-pool	0.734	0.746	0.695	0.689	0.741	0.684	0.733	0.714	0.694	0.730
unirep1900	0.744	0.745	0.703	0.686	0.753	0.696	0.731	0.716	0.703	0.728
unirep5700	0.729	0.749	0.690	0.688	0.755	0.690	0.734	0.706	0.705	0.726
only-ecfp4	0.591	0.665	0.643	0.426	0.600	0.514	0.519	0.534	0.576	0.503
ran200_ran-ecfp4	0.382	0.481	0.400	0.256	0.449	0.401	0.320	0.265	0.319	0.235
only-ran-ecfp4	0.296	0.175	0.082	0.165	0.358	0.137	0.189	0.171	0.224	0.173

(b)

Dissimilar-compound-split	epigenetic-regulators	hydrolases	ion-channels	membrane-receptors	other-enzymes	oxidoreduc-tases	proteases	transcripti-on-factors	transferase-s	transporter-s
apaac	0.137	0.355	0.342	0.249	0.419	0.391	0.381	0.058	0.358	0.362
ctdd	0.021	0.407	0.311	0.241	0.386	0.423	0.391	0.048	0.342	0.405
ctriad	0.045	0.351	0.276	0.243	0.354	0.405	0.353	0.006	0.346	0.425
dde	0.089	0.371	0.327	0.223	0.359	0.398	0.341	0.071	0.340	0.403
geary	0.044	0.375	0.291	0.242	0.397	0.400	0.394	0.036	0.347	0.417
k-sep_pssm	0.239	0.382	0.419	0.298	0.381	0.449	0.354	0.071	0.318	0.368
pfam	0.455	0.329	0.448	0.146	0.452	0.339	0.319	0.257	0.308	0.366
qso	0.247	0.373	0.338	0.278	0.345	0.356	0.369	0.071	0.324	0.419
random200	0.152	0.386	0.273	0.266	0.348	0.409	0.341	0.103	0.341	0.388
spmap	0.158	0.351	0.289	0.274	0.361	0.395	0.383	0.036	0.335	0.369
taap	0.289	0.371	0.434	0.243	0.443	0.360	0.398	0.187	0.322	0.438
protvec	-0.024	0.348	0.437	0.222	0.363	0.381	0.366	0.118	0.344	0.390
seqvec	0.192	0.349	0.310	0.228	0.374	0.435	0.373	0.033	0.359	0.378
transformer-avg	0.075	0.349	0.288	0.250	0.447	0.428	0.391	0.043	0.328	0.390
transformer-pool	0.104	0.348	0.364	0.258	0.411	0.402	0.357	0.061	0.316	0.376
unirep1900	0.161	0.334	0.277	0.256	0.402	0.400	0.376	0.076	0.329	0.370
unirep5700	0.061	0.350	0.289	0.255	0.387	0.374	0.346	0.090	0.320	0.376
only-ecfp4	0.428	0.244	0.243	0.168	0.419	0.281	0.307	0.058	0.306	0.309
ran200_ran-ecfp4	-0.076	0.254	0.293	0.133	0.210	0.270	0.138	0.083	0.178	0.284
only-ran-ecfp4	0.004	-0.020	0.018	0.001	-0.028	-0.015	-0.008	0.016	-0.020	-0.035

(c)

Fully-dissimilar-split	epigenetic-regulators	hydrolases	ion-channels	membrane-receptors	other-enzymes	oxidoreduc-tases	proteases	transcripti-on-factors	transferase-s	transporter-s
apaac	0.403	0.156	0.146	0.243	0.129	-0.044	0.192	0.063	0.300	0.240
ctdd	0.396	0.132	-0.074	0.253	0.162	0.074	0.207	0.101	0.247	-0.017
ctriad	0.420	0.203	0.086	0.220	0.125	0.030	0.212	0.238	0.273	0.276
dde	0.375	0.170	0.124	0.206	0.232	-0.050	0.195	0.150	0.295	0.029
geary	0.319	0.160	0.155	0.195	0.172	0.044	0.268	0.066	0.275	-0.027
k-sep_pssm	0.252	0.181	0.157	0.137	0.043	0.052	-0.134	0.086	0.300	0.297
pfam	0.446	0.208	0.174	0.270	0.221	0.088	0.142	0.156	0.301	0.198
qso	0.397	0.111	0.044	0.166	0.187	0.141	0.215	0.129	0.300	0.202
random200	0.289	0.040	0.146	0.226	0.282	0.070	0.149	0.051	0.284	-0.194
spmap	0.361	0.103	0.182	0.213	0.114	0.015	0.209	0.091	0.287	0.118
taap	0.289	0.155	0.181	0.286	0.208	-0.028	0.205	0.129	0.310	0.194
protvec	0.275	0.146	0.174	0.235	0.131	0.077	0.184	0.160	0.301	0.204
seqvec	0.372	0.154	0.032	0.155	0.199	0.018	0.046	0.150	0.276	-0.023
transformer-avg	0.367	0.129	0.058	0.265	0.176	0.092	0.227	0.144	0.311	0.187
transformer-pool	0.403	0.148	-0.052	0.244	0.079	0.133	0.170	0.117	0.313	-0.040
unirep1900	0.325	0.186	0.143	0.217	0.205	0.132	0.159	-0.008	0.332	0.004
unirep5700	0.339	0.154	0.143	0.204	0.229	0.226	0.189	0.127	0.324	0.114
only-ecfp4	0.426	0.205	0.133	0.254	0.240	0.016	0.287	0.169	0.294	0.172
ran200_ran-ecfp4	0.087	0.067	-0.012	0.022	-0.001	0.123	0.061	-0.003	0.057	0.032
only-ran-ecfp4	0.041	-0.003	-0.013	0.001	0.010	-0.031	-0.019	0.001	0.018	0.025

Table S4: Prediction error percentages of transformer-avg models with different thresholds on random, dissimilar-compound, and fully-dissimilar splits of transferases family dataset.

	PE > 0.5 (%)	PE > 1 (%)	PE > 1.5 (%)	PE > 2 (%)
Random-split	36.6	11.8	3.6	0.9
Dissimilar-compound-split	69.6	40.8	17.8	6.9
Fully-dissimilar-split	64.9	39.0	22.5	10.7

PE: Prediction error

Table S5. Statistics of the protein family specific datasets used in the large-scale analysis.

Protein family	Datapoints		Proteins		Compounds	
	initial	discarded	initial	discarded	initial	discarded
epigenetic regulators	19,219	1,055	125	8	10,637	276
hydrolases	71,996	9,051	623	12	56,313	7,282
ion channels	38,097	3,865	241	26	32,376	2,073
membrane receptors	220,810	14,635	640	3	133,440	8,357
other enzymes	36,857	4,007	456	9	21,428	3,372
oxidoreductases	59,023	4,545	404	2	42,338	2,895
proteases	99,100	11,839	409	20	64,688	8,369
transcription factors	25,292	216	119	0	17,244	80
transferases	191,864	31,900	1,020	39	124,266	18,963
transporters	29,165	804	190	2	18,493	523

Table S6. Split-based statistics of the protein family specific datasets for; **(a)** the random-split, **(b)** dissimilar-compound-split, and **(c)** the fully-dissimilar-split sets (ratio represents the ratio of train to test).

(a)

Random-split	Bioactivity datapoints			Proteins			Compounds		
	Protein family	train	test	ratio	train	test	ratio	train	test
epigenetic-regulators	16,675	1,489	11.2	117	83	1.4	9,797	1,388	7.1
hydrolases	59,021	3,924	15.0	610	304	2.0	46,575	3,851	12.1
ion-channels	31,287	2,945	10.6	213	142	1.5	27,961	2,898	9.6
membrane-receptors	197,919	8,256	24.0	636	428	1.5	121,954	8,030	15.2
other-enzymes	29,465	3,385	8.7	442	256	1.7	16,757	3,040	5.5
oxidoreductases	52,277	2,201	23.8	402	240	1.7	38,277	2,169	17.6
proteases	83,561	3,700	22.6	389	219	1.8	54,628	3,598	15.2
transcription-factors	23,498	1,578	14.9	118	80	1.5	16,465	1,529	10.8
transferases	150,247	9,717	15.5	980	634	1.5	100,462	9,070	11.1
transporters	26,313	2,048	12.8	188	114	1.6	17,141	1,958	8.8

(b)

Dissimilar-compound-split	Bioactivity datapoints			Proteins			Compounds		
Protein family	train	test	ratio	train	test	ratio	train	test	ratio
epigenetic-regulators	16,763	1,401	12.0	117	52	2.3	9,536	825	11.6
hydrolases	58,953	3,992	14.8	597	178	3.4	45,913	3,118	14.7
ion-channels	31,329	2,903	10.8	208	69	3.0	27,656	2,647	10.4
membrane-receptors	197,652	8,523	23.2	627	252	2.5	118,914	6,169	19.3
other-enzymes	29,325	3,525	8.3	433	157	2.8	15,428	2,628	5.9
oxidoreductases	52,190	2,288	22.8	396	107	3.7	37,763	1,680	22.5
proteases	83,405	3,856	21.6	383	130	2.9	53,493	2,826	18.9
transcription-factors	23,444	1,632	14.4	116	38	3.1	16,015	1,149	13.9
transferases	150,216	9,748	15.4	956	393	2.4	98,103	7,200	13.6
transporters	26,273	2,088	12.6	182	63	2.9	16,545	1,425	11.6

(c)

Fully-dissimilar-split	Bioactivity datapoints			Proteins			Compounds		
Protein family	train	test	ratio	train	test	ratio	train	test	ratio
epigenetic-regulators	16,675	1,489	11.2	117	52	2.3	9,536	825	11.6
hydrolases	59,021	3,924	15.0	597	178	3.4	45,913	3,118	14.7
ion-channels	31,287	2,945	10.6	208	69	3.0	27,656	2,647	10.4
membrane-receptors	197,919	8,256	24.0	627	252	2.5	118,914	6,169	19.3
other-enzymes	29,465	3,385	8.7	433	157	2.8	15,428	2,628	5.9
oxidoreductases	52,277	2,201	23.8	396	107	3.7	37,763	1,680	22.5
proteases	83,561	3,700	22.6	383	130	2.9	53,493	2,826	18.9
transcription-factors	23,498	1,578	14.9	116	38	3.1	16,015	1,149	13.9
transferases	150,247	9,717	15.5	956	393	2.4	98,103	7,200	13.6
transporters	26,313	2,048	12.8	182	63	2.9	16,545	1,425	11.6

Table S7. Statistics of the compound-centric datasets used in the small-scale analysis.

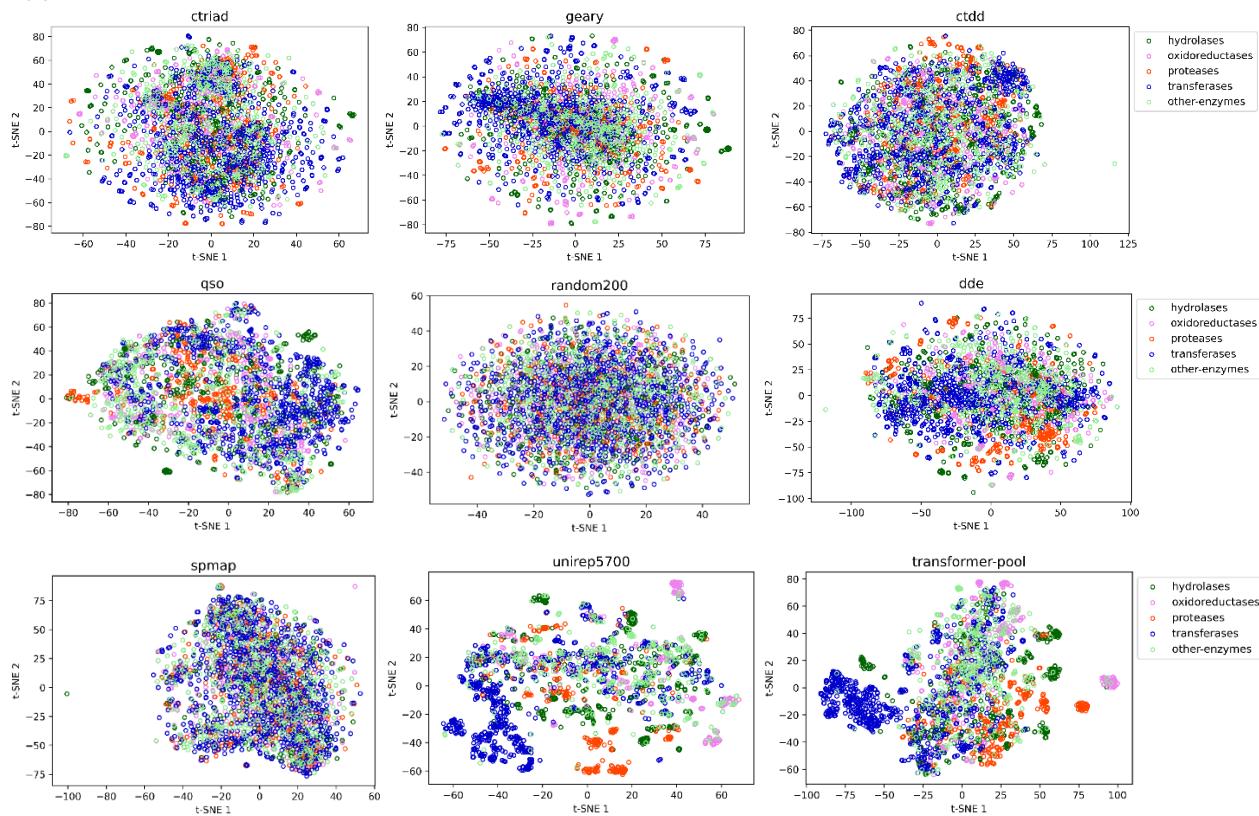
Name of the center compound	ChEMBL id of the center compounds	cluster size (total # of compounds)	# of active targets	# of inactive targets
Curcumin	CHEMBL116438	1648	94	66
Tamoxifen	CHEMBL83	754	147	35
Quercetin	CHEMBL50	543	92	64
Genistein	CHEMBL44	479	102	57
Econazole	CHEMBL808	246	54	23
Levoketoconazole	CHEMBL295698	243	79	27
Amiodarone	CHEMBL633	194	73	28
Miconazole	CHEMBL91	170	49	22
Clotrimazole	CHEMBL104	52	38	37

Table S8. Enzyme sub-classes of the protein family specific datasets used in the large-scale analysis.

Main class EC number(s) of the merged enzyme classes	Names of merged enzyme classes
1	Oxidoreductases, Cytochrome P450s
2	Transferases, Kinases
3 (except 3.4)	Hydrolases, Phosphatases, Phosphodiesterases
3.4	Proteases (a.k.a. Peptidases)
All remaining	Other, Lyases

Supplementary Figures

(a)



(b)

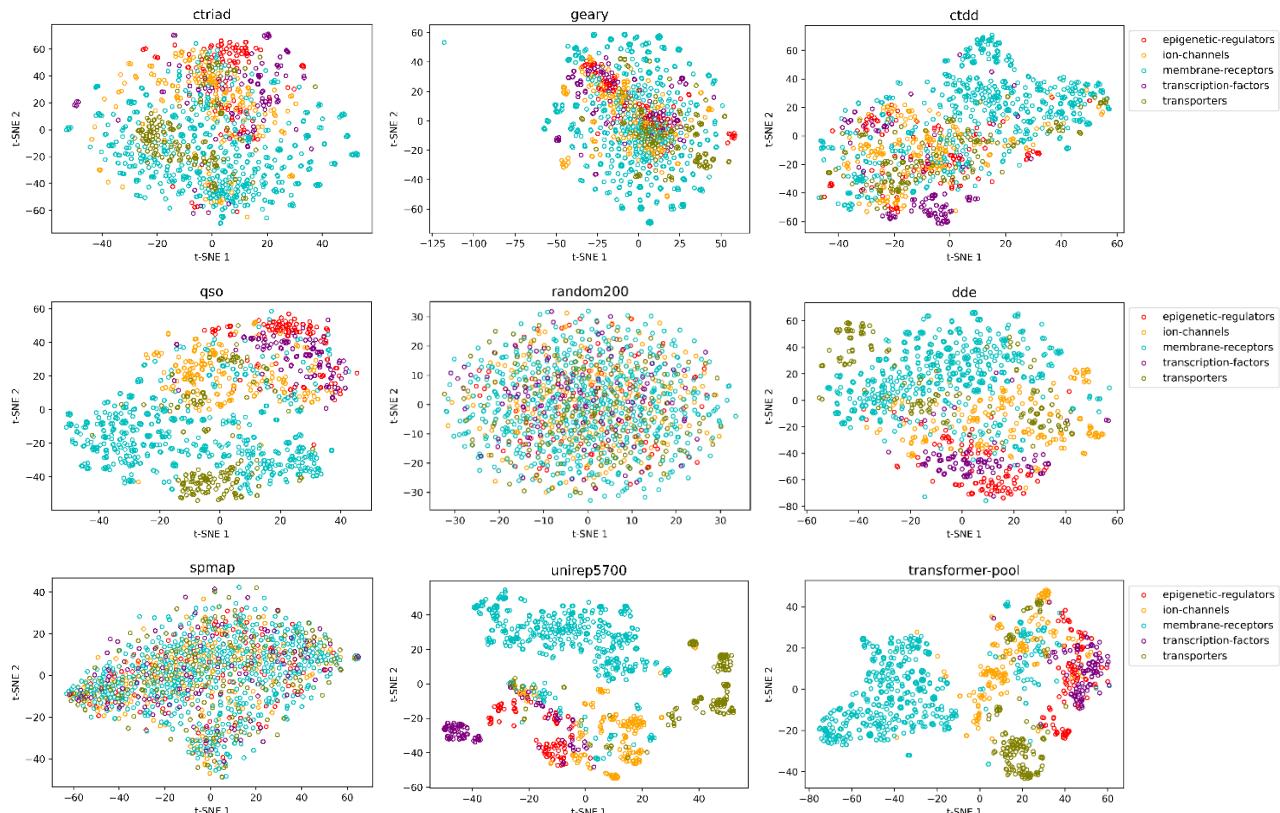
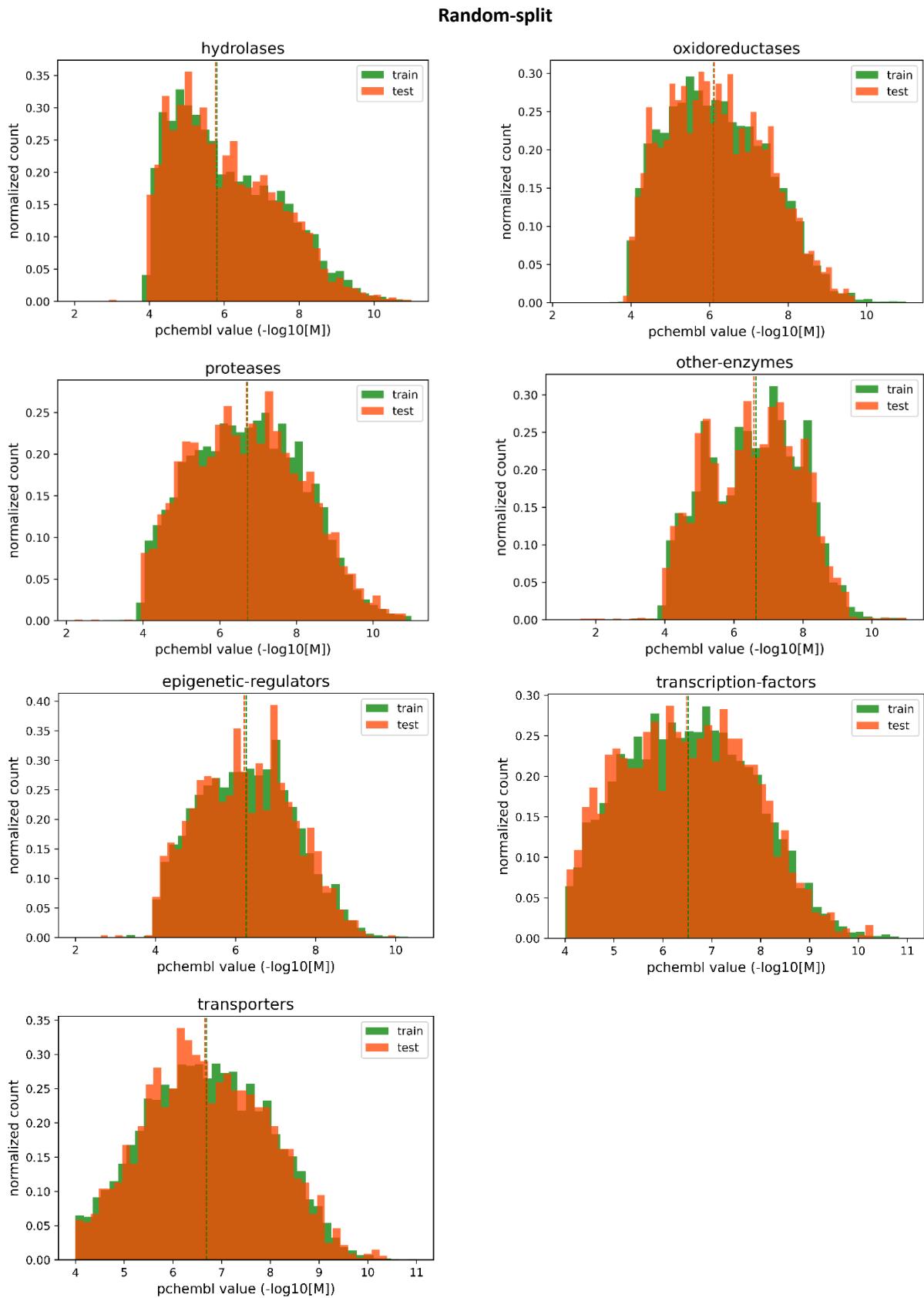


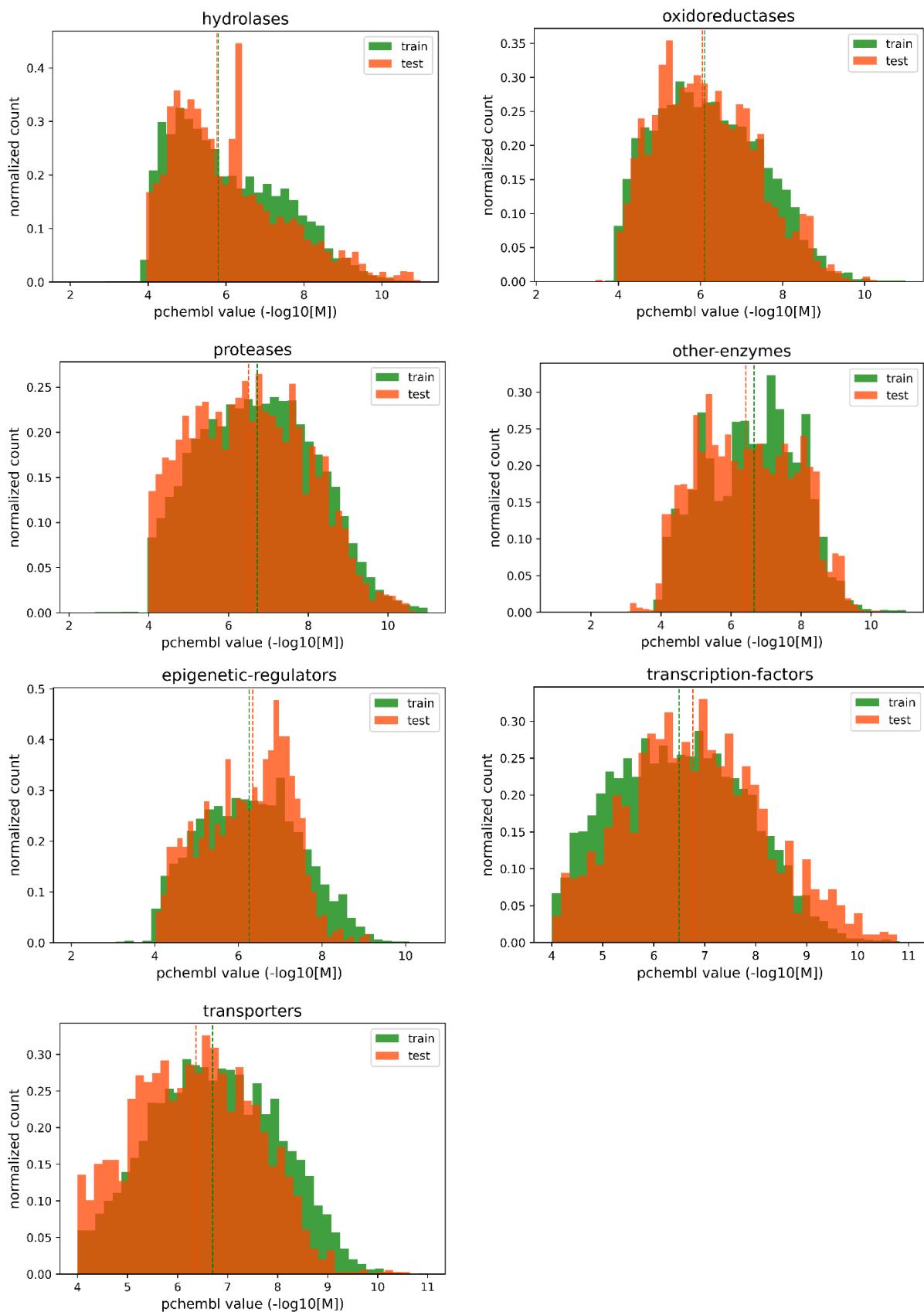
Figure S1. t-SNE based visualization of conventional and learned protein representations on; (a) enzymes including hydrolases, oxidoreductases, proteases, transferases, and other-enzymes, and (b) other protein families (non-enzymes) including epigenetic regulators, ion channels, membrane receptors, transcription factors and transporters.

(a)



(b)

Dissimilar-compound-split



(c)

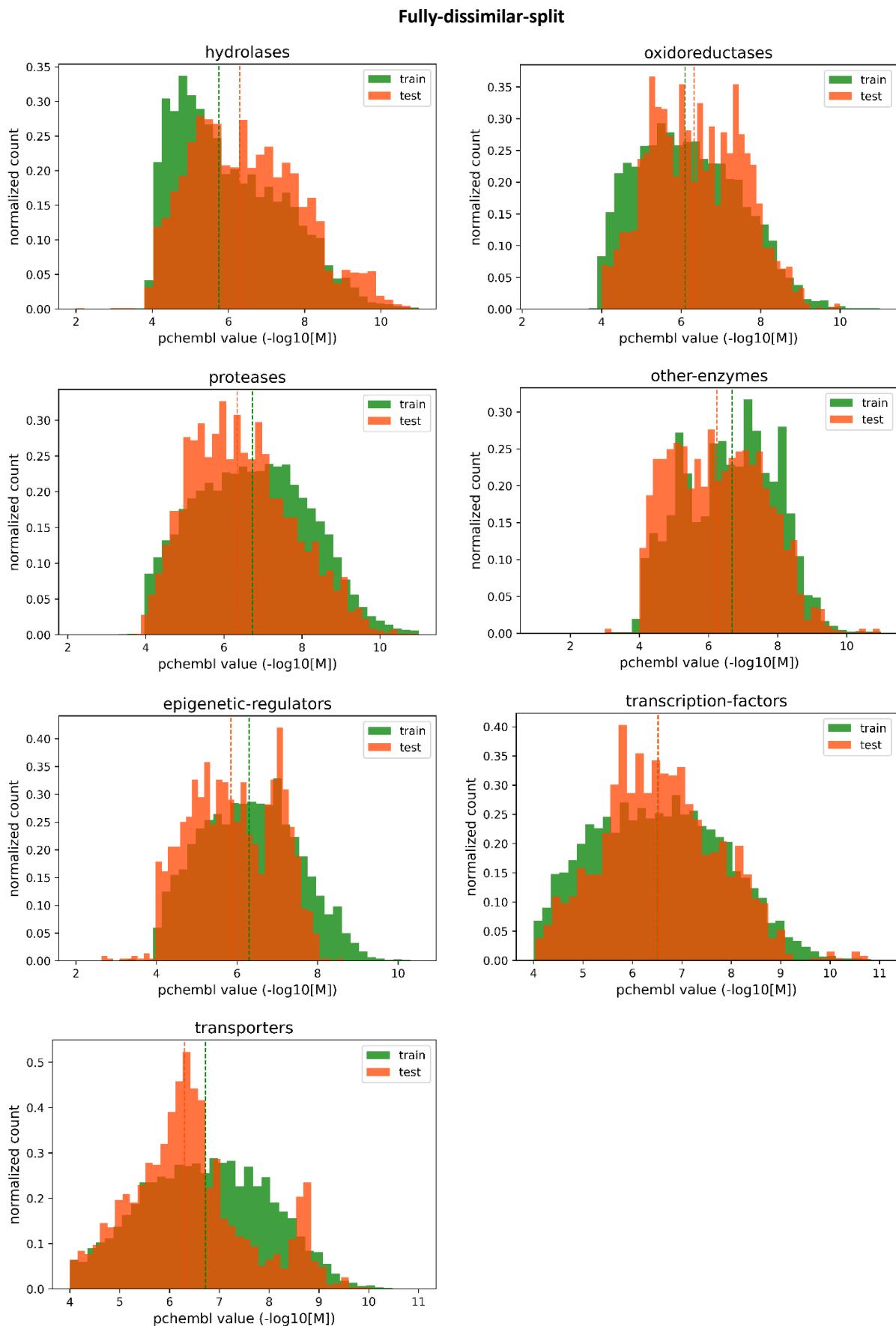


Figure S2. Bioactivity distributions of protein family-specific datasets in terms of; **(a)** random split, **(b)** dissimilar-compound split, and **(c)** fully-dissimilar split sets, together with the median values shown as vertical dashed lines.

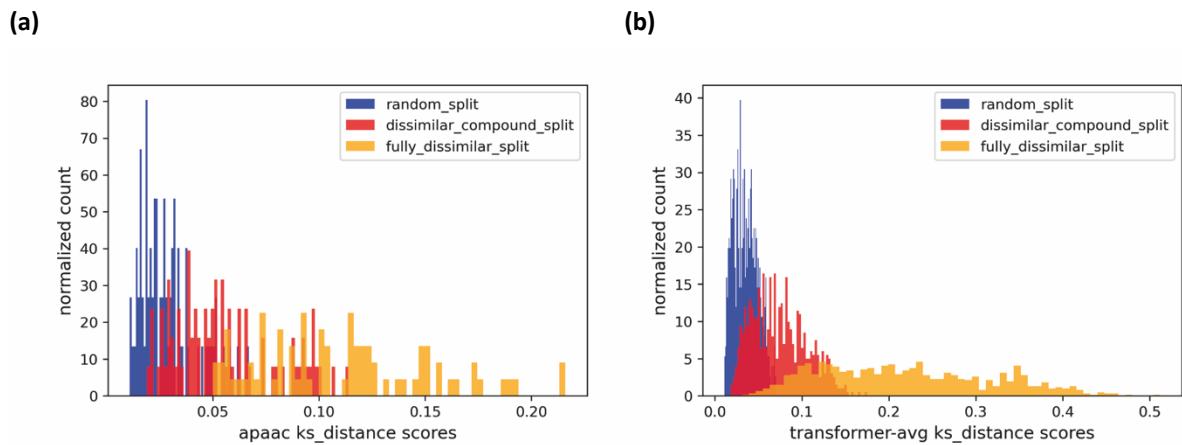
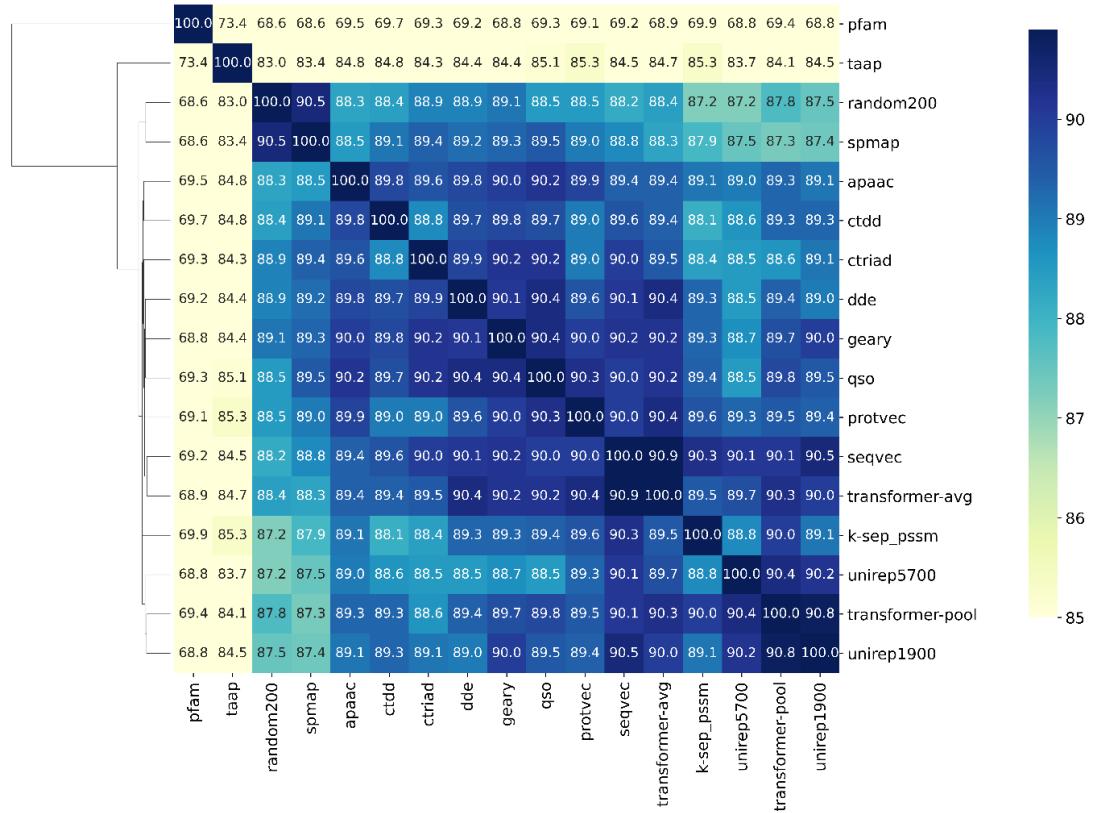
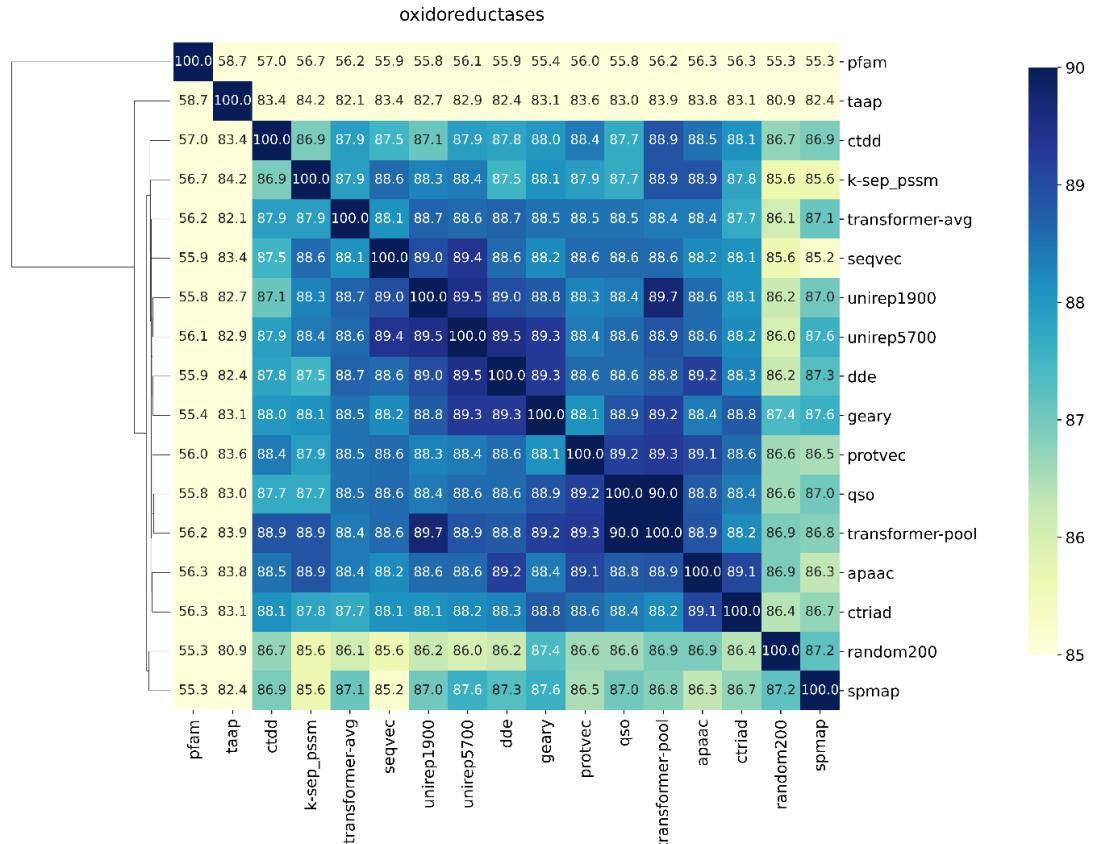
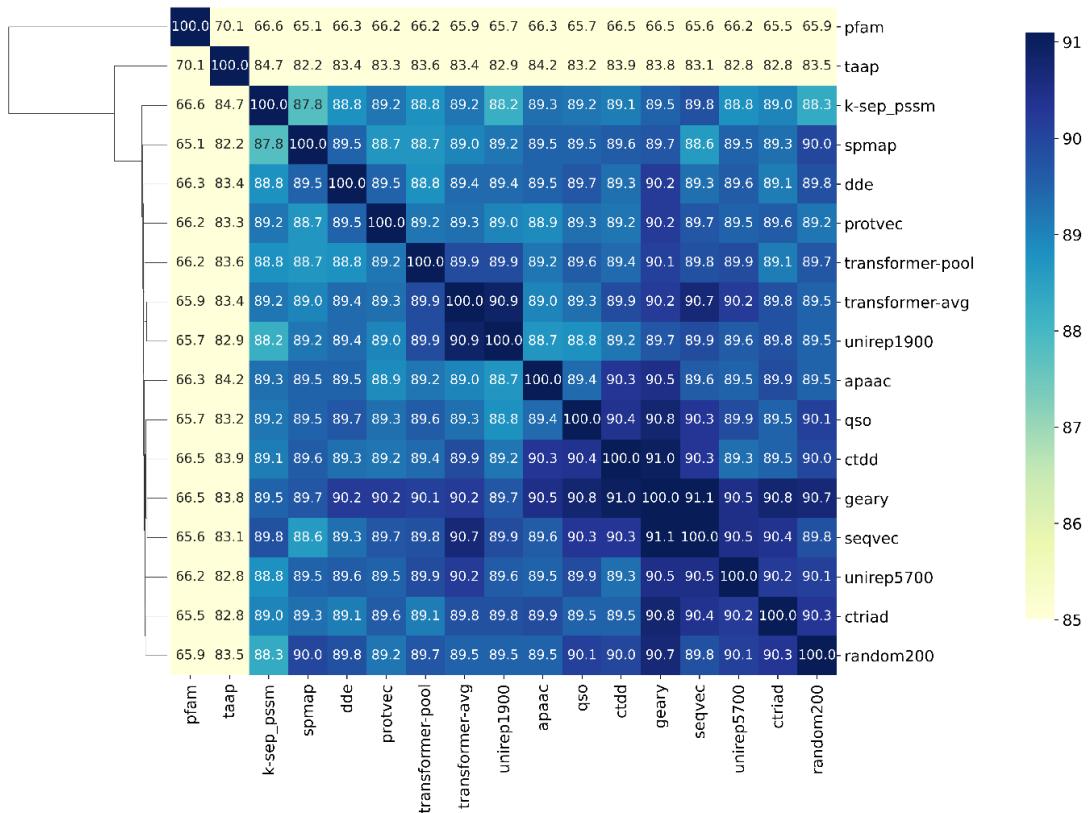


Figure S3: KS distance (between train and test samples) score distributions of **(a)** apaac, and **(b)** transformer-avg representations among random, dissimilar-compound, and fully-dissimilar splits in the transferases family proteins.

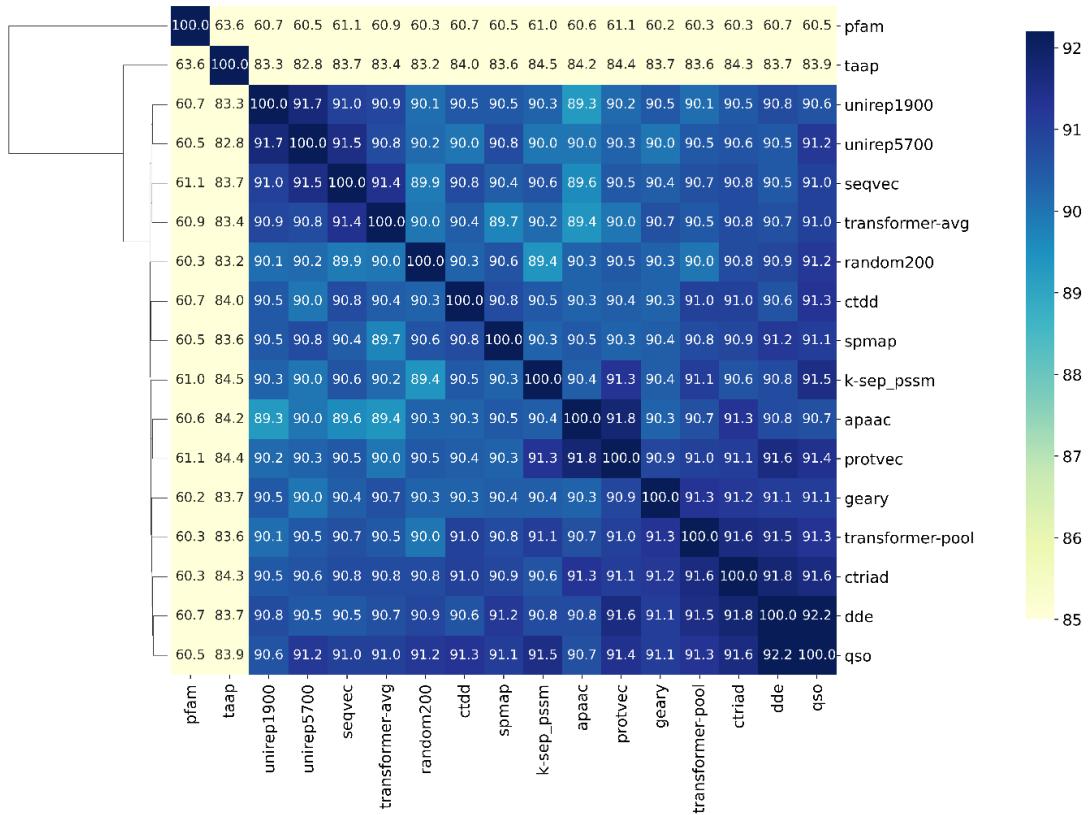
(a)

Random-split**hydrolases****oxidoreductases**

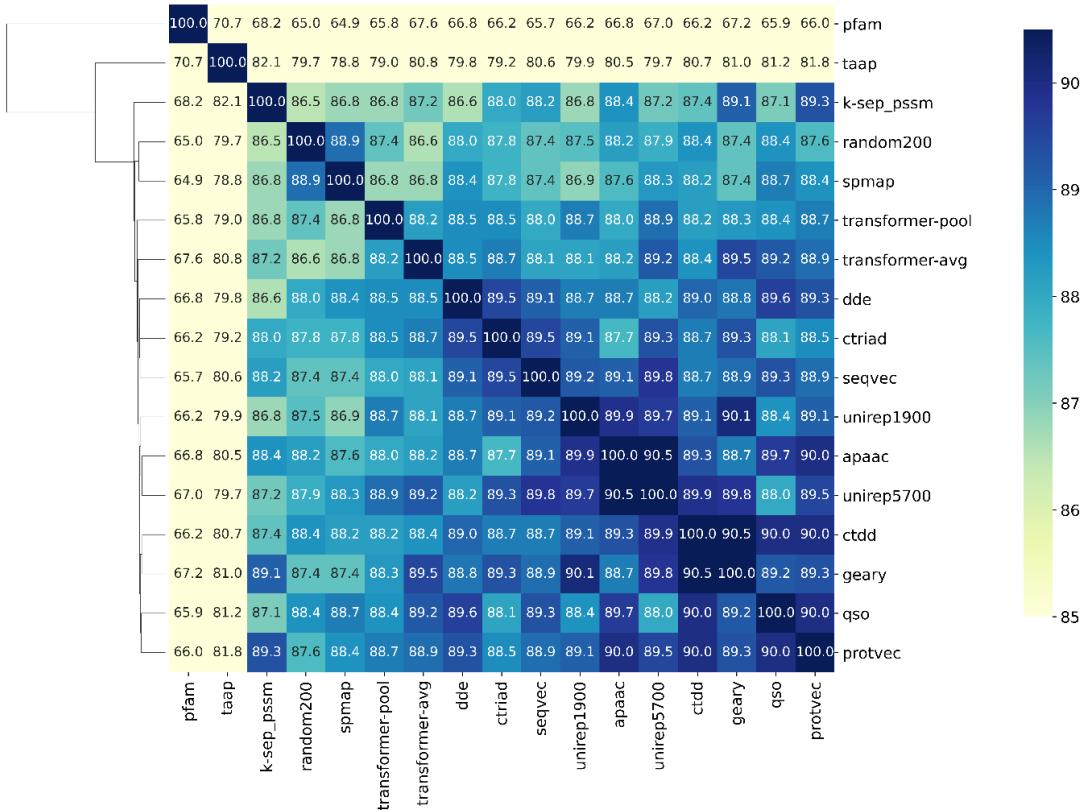
proteases



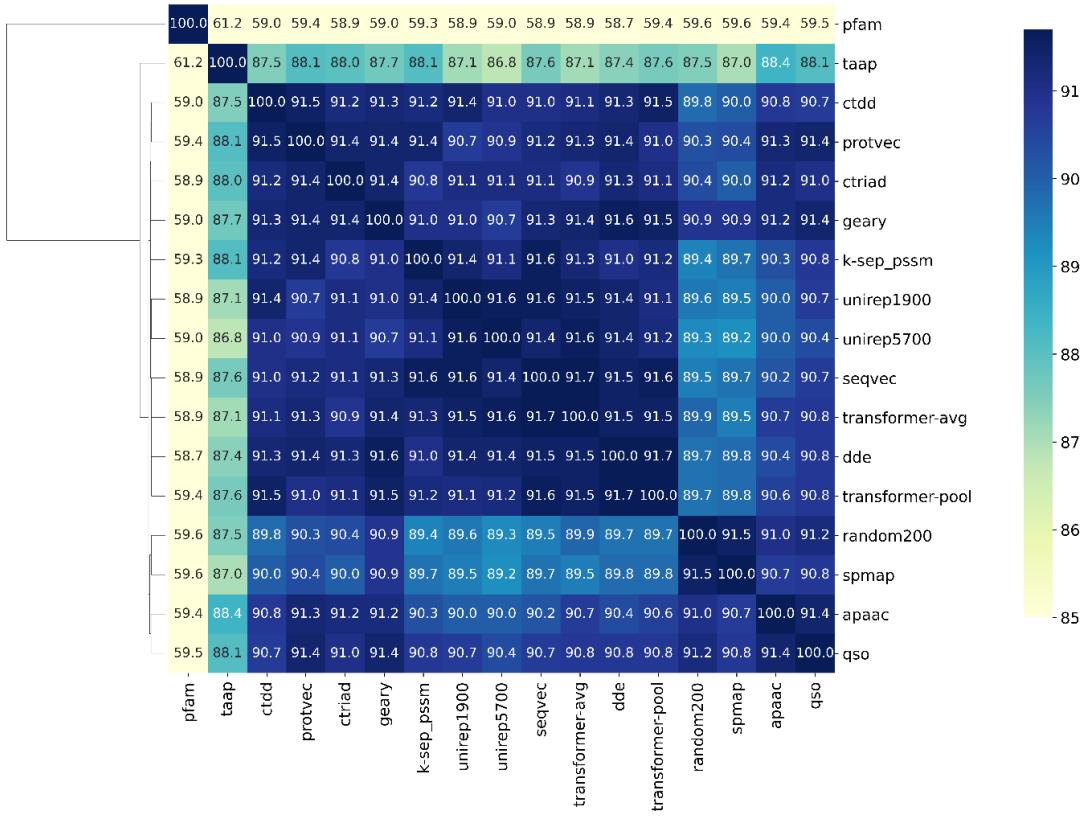
other-enzymes



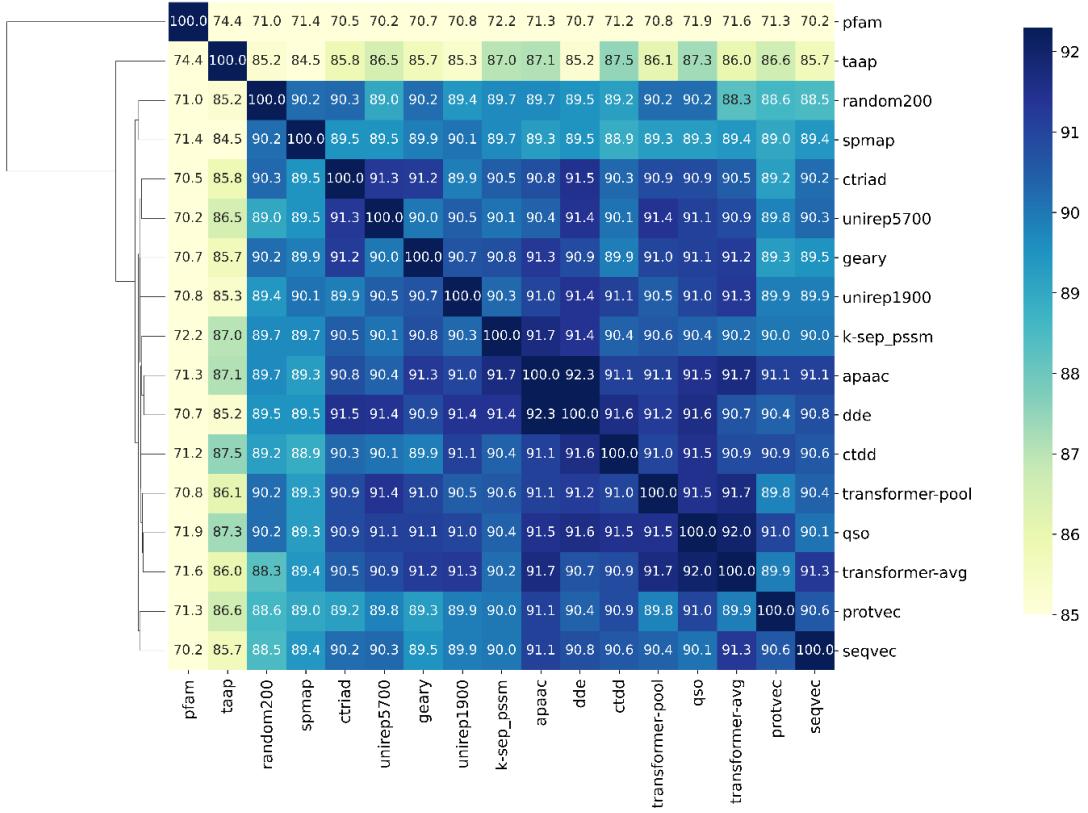
epigenetic-regulators



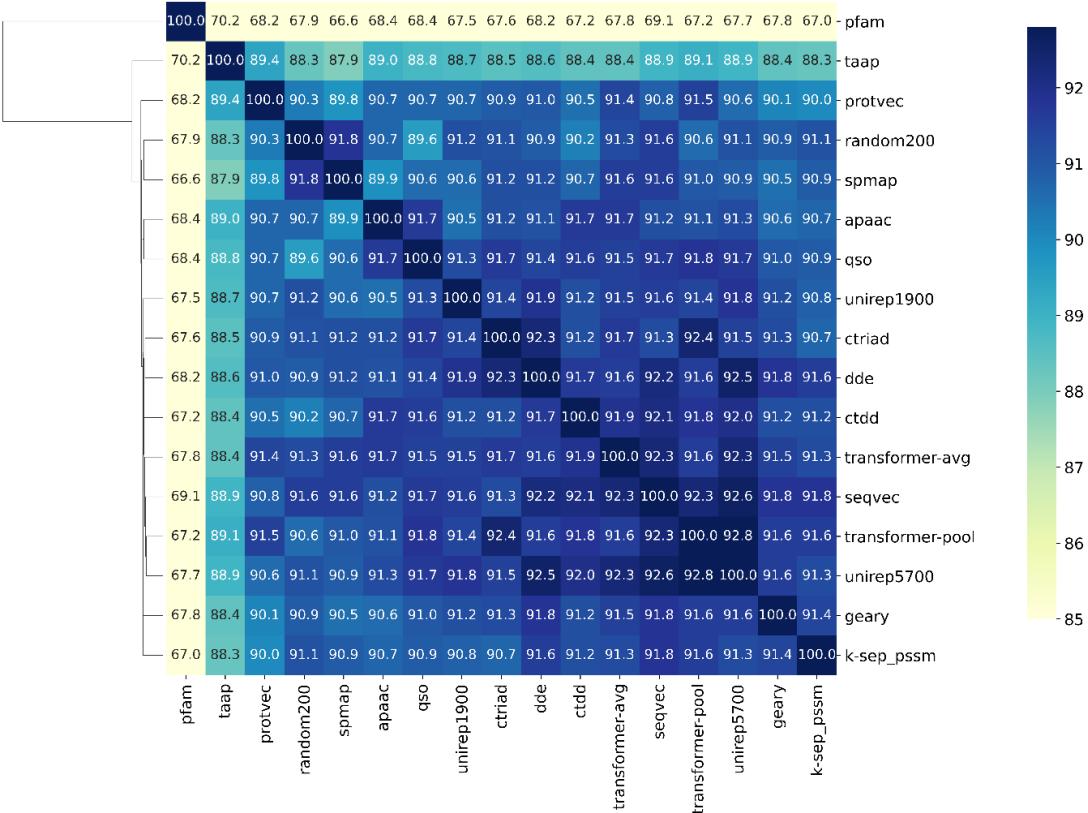
membrane-receptors



transcription-factors

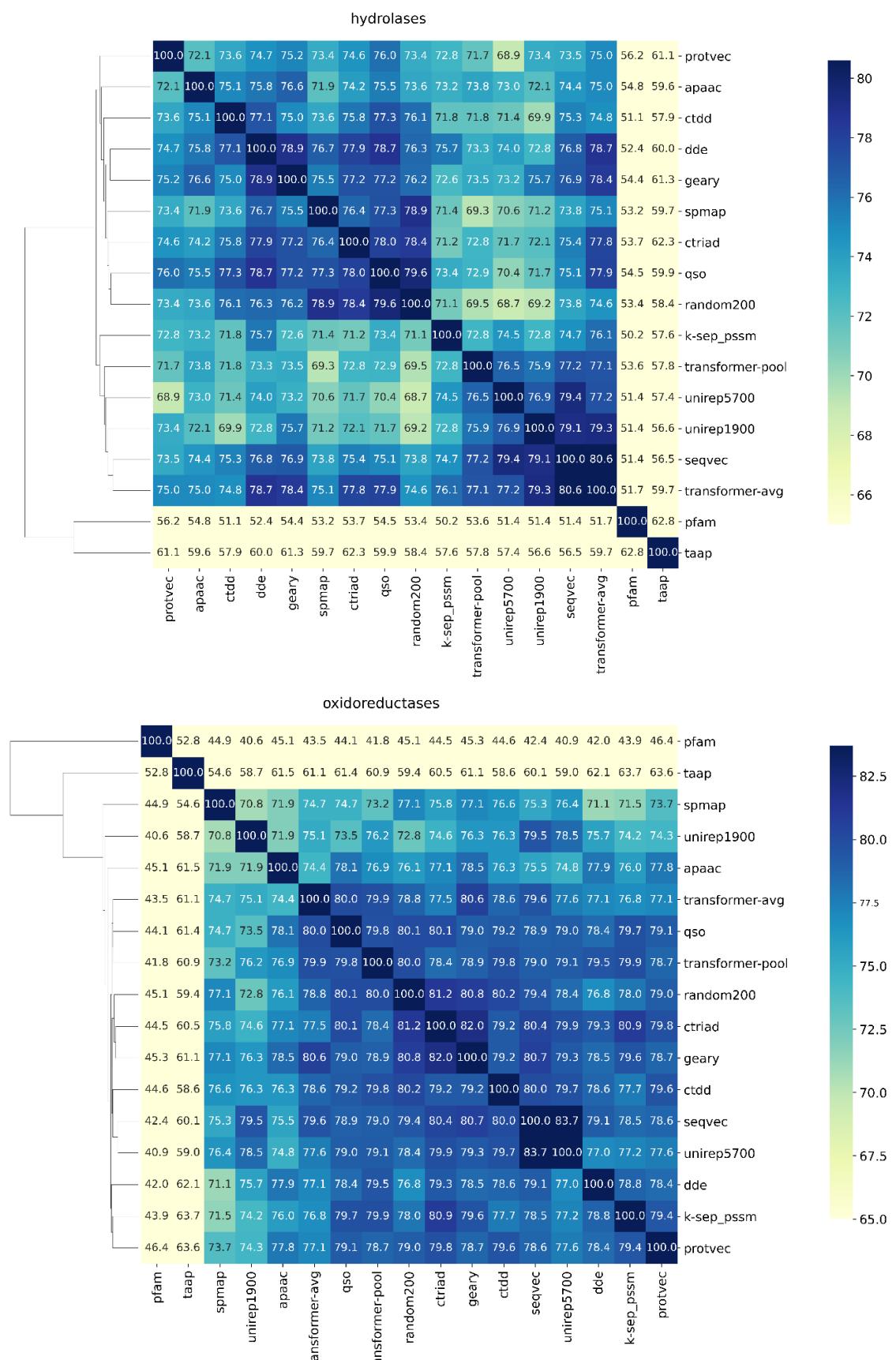


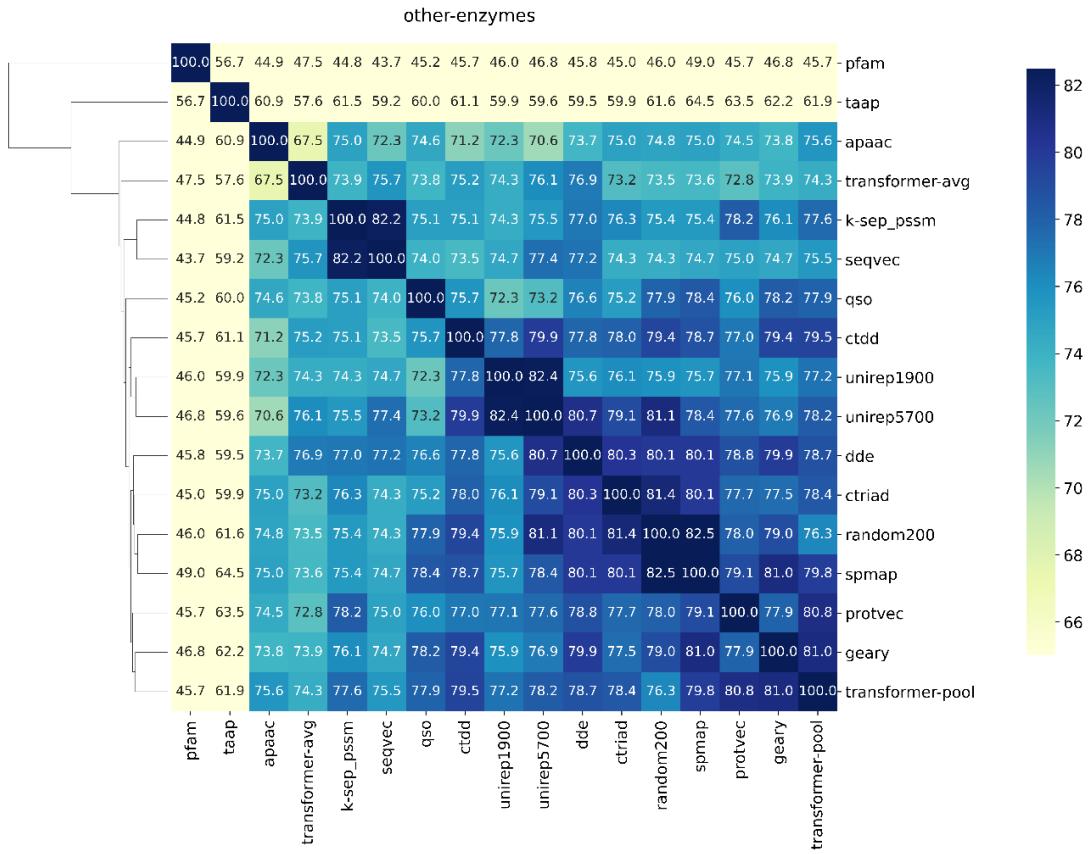
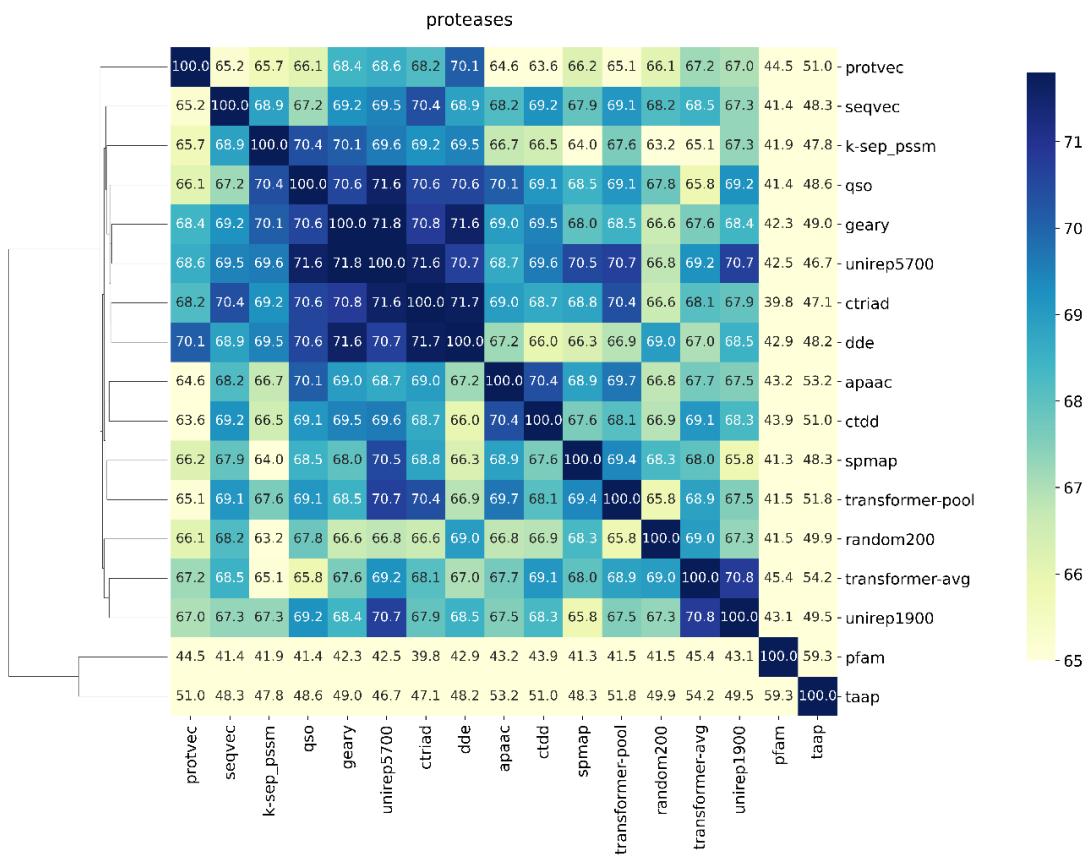
transporters



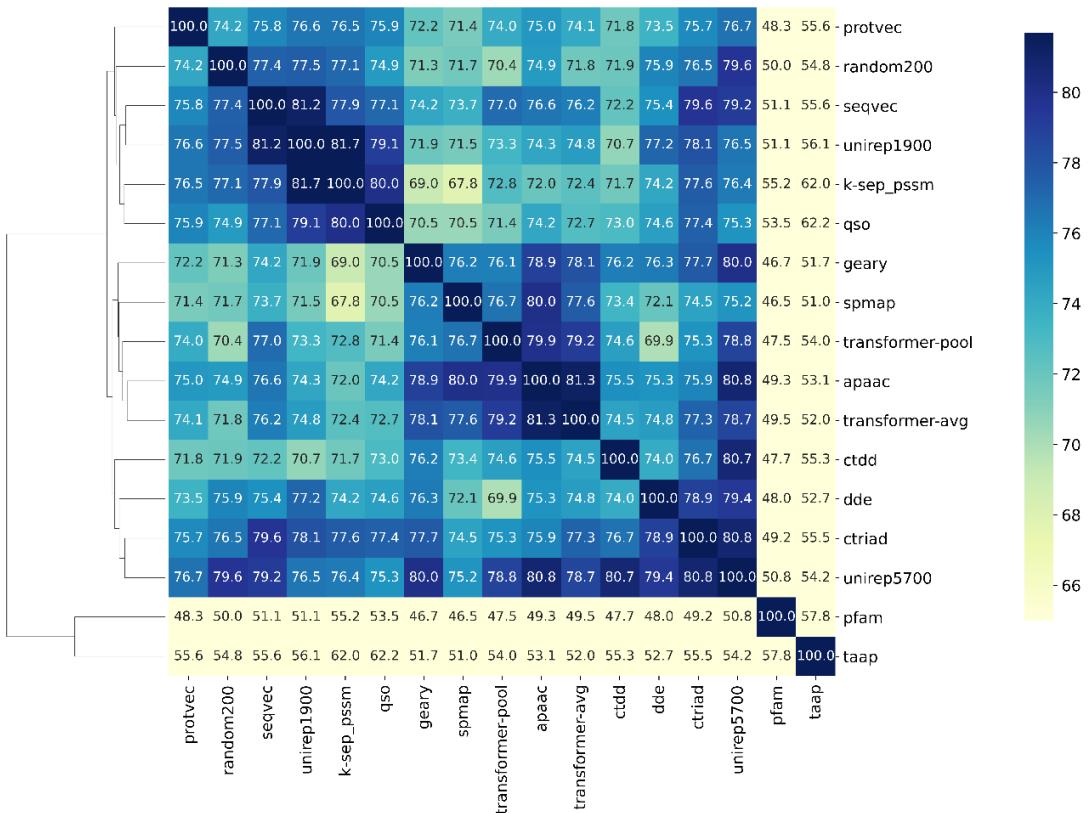
(b)

Dissimilar-compound-split

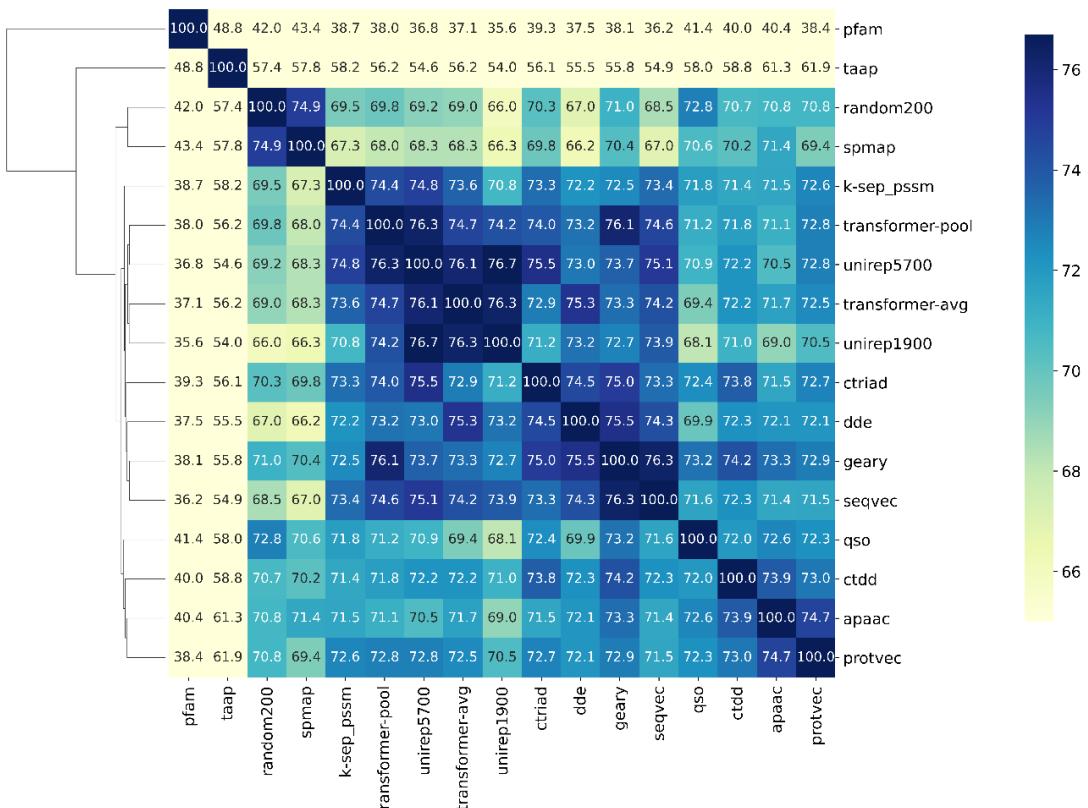


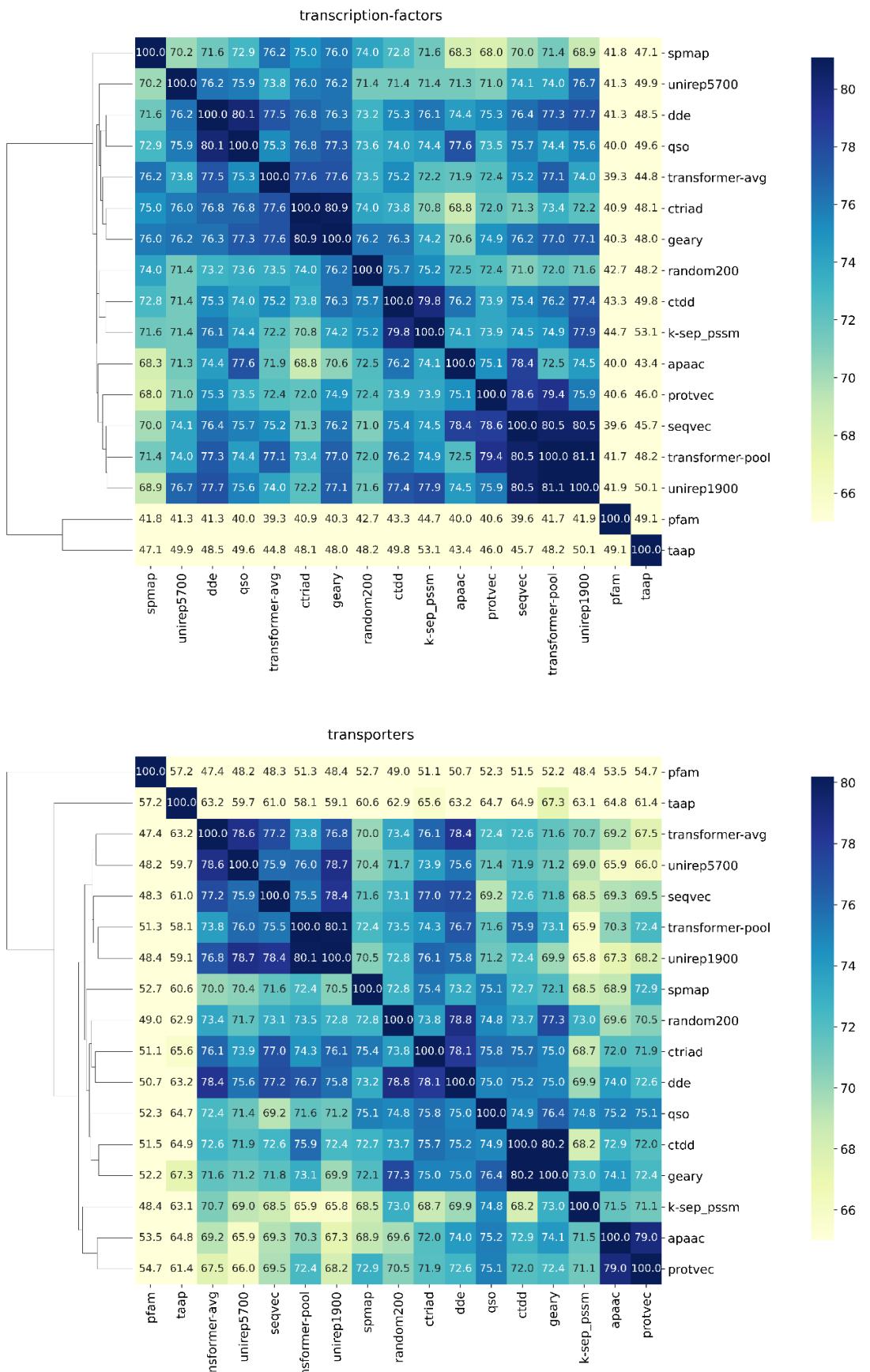


epigenetic-regulators



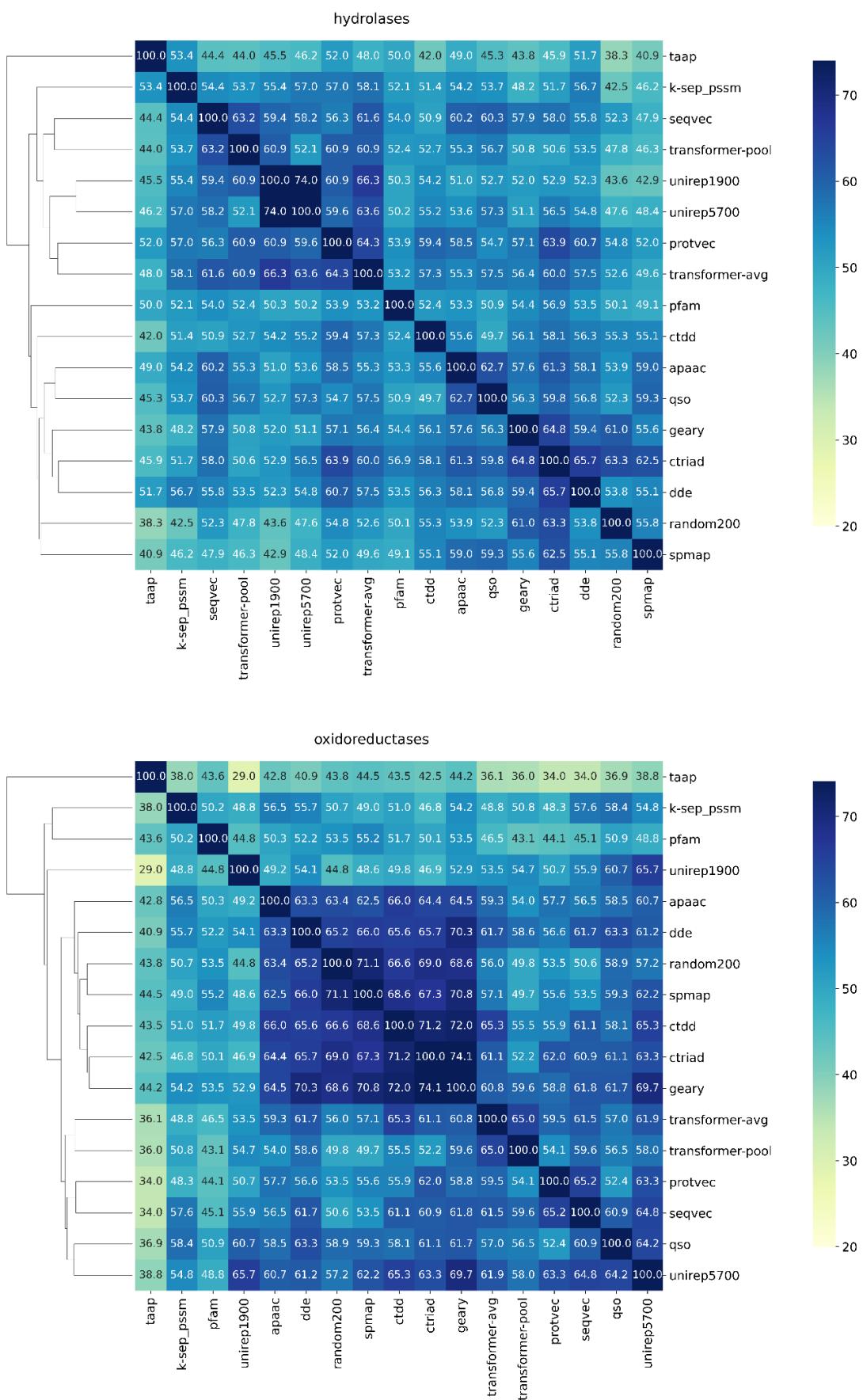
membrane-receptors

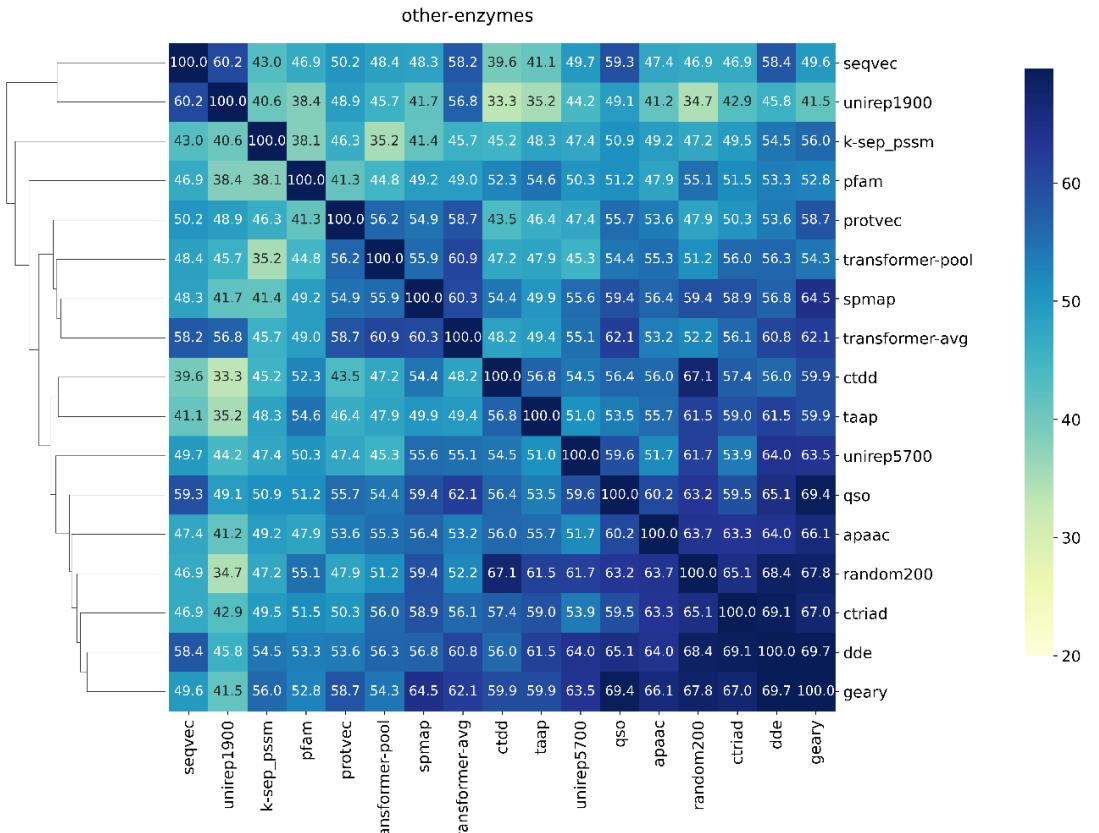
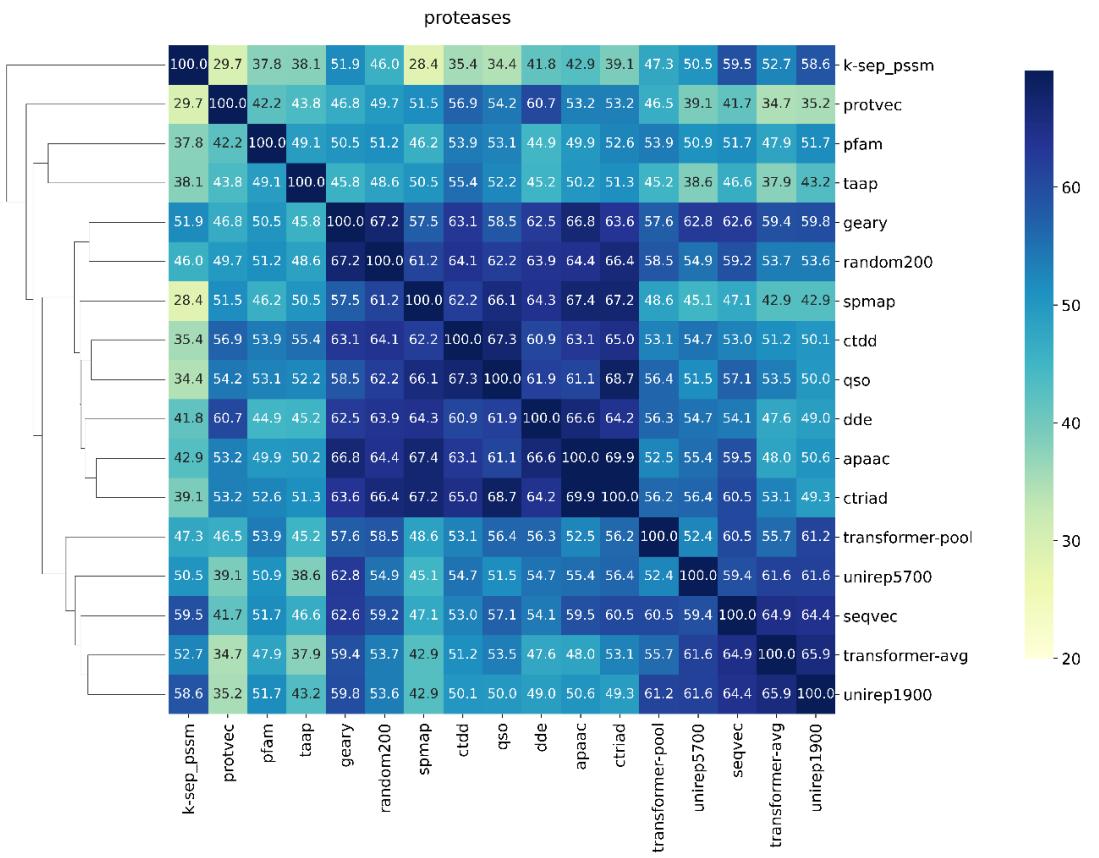




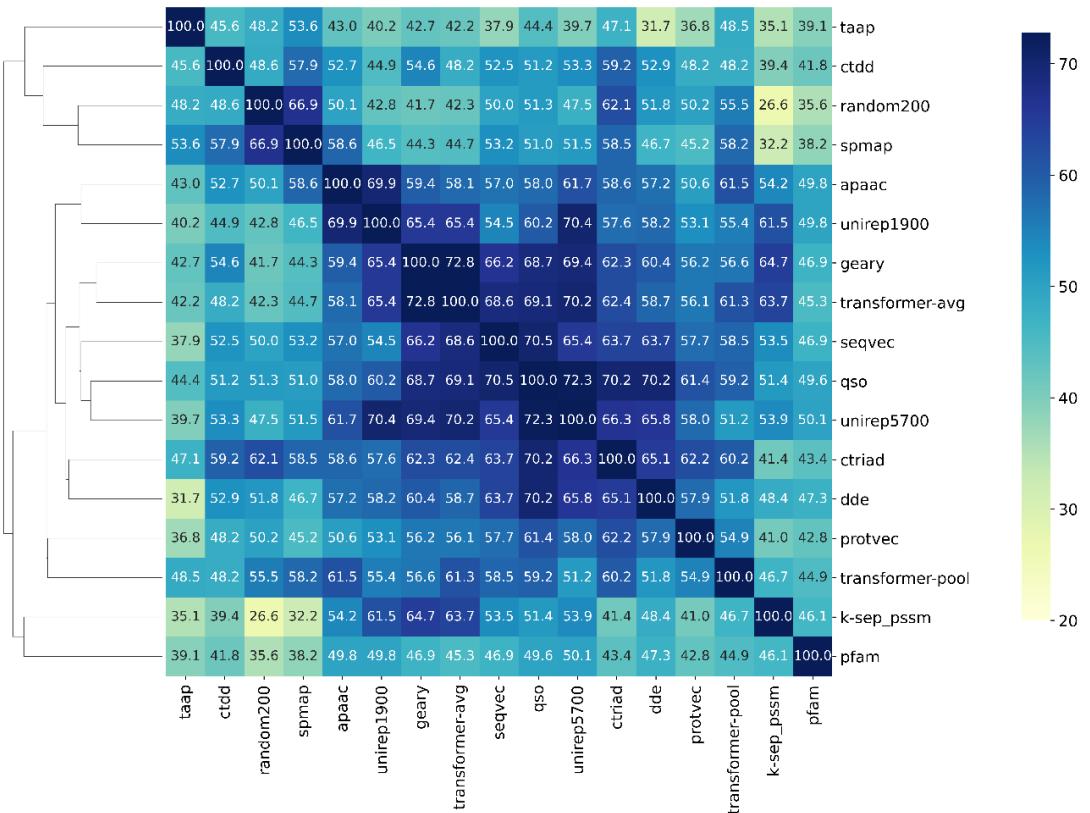
(c)

Fully-dissimilar-split

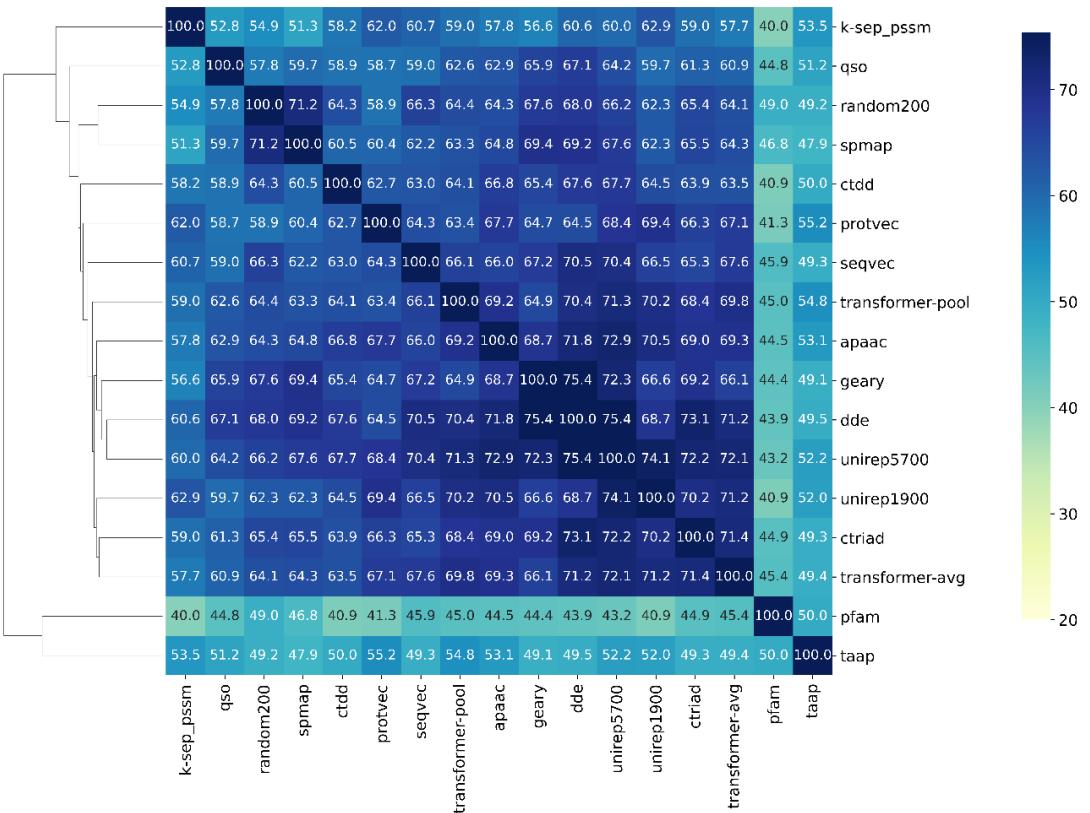




epigenetic-regulators



membrane-receptors



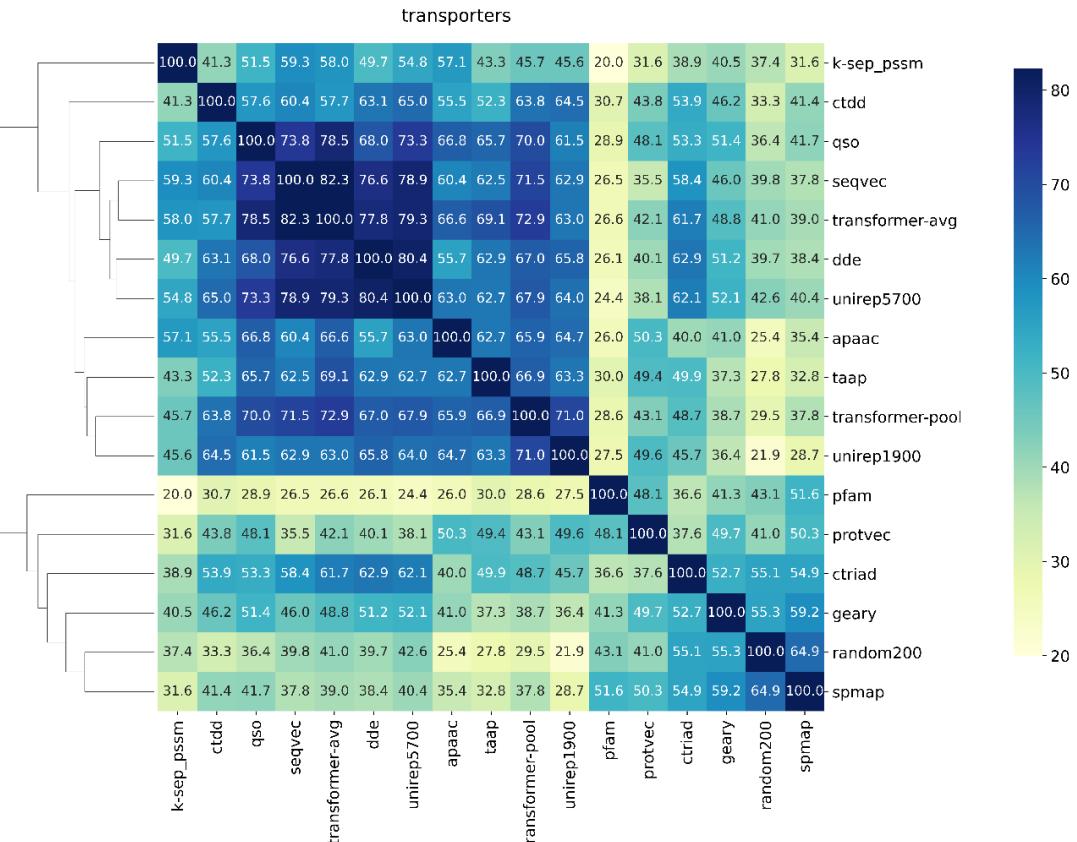
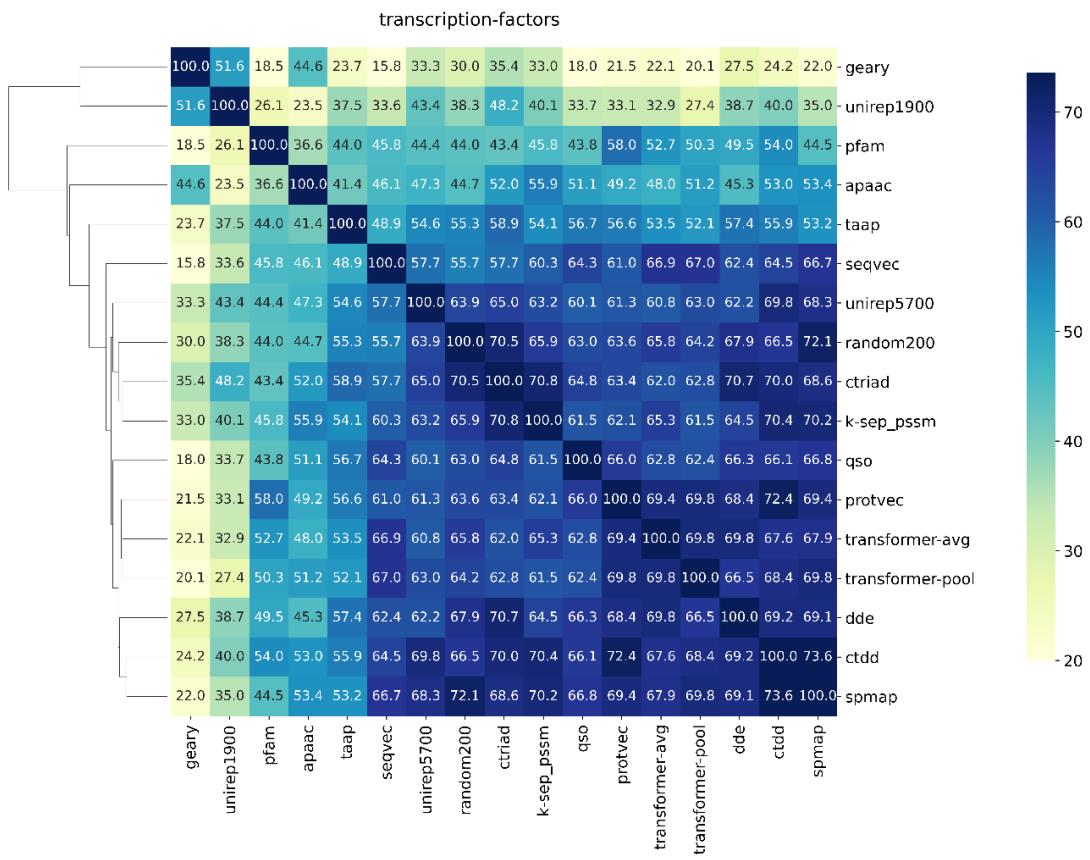


Figure S4. Clustered heatmaps of different protein representation approaches for protein families on (a) the random-split, (b) dissimilar-compound-split, and (c) fully-dissimilar-split datasets.