

Supplementary Information for the independent and comprehensive evaluation of CAZyme classifiers

Emma E. M. Hobbs, Tracey M. Gloster, Leighton Pritchard

June 2023

This document contains all supplementary information for the paper: Independent and comprehensive evaluation of CAZyme classifiers (Hobbs *et al.*, 2023). The tables and figures are presented in the same order as they are referenced in the main manuscript.

Contents

1	Test set composition	2
2	Evaluation of CAZyme/non-CAZyme classification	6
2.1	Summary of CAZyme/non-CAZyme classification	6
2.2	Output of testing for statistically significant difference in performance between the tools	6
3	Taxonomic performance of CAZyme/non-CAZyme classification	9
3.1	Summary of taxonomic kingdom performance of CAZyme/non-CAZyme classification	9
4	Evaluation of CAZy class classification across all CAZy class	12
5	Evaluation of CAZy class classification per CAZy class	14
6	Taxonomic performance of CAZy class classification across all CAZy classes	16
7	Overall performance of CAZy family classification (across all CAZy families and classes)	18
8	Multi-label classification of CAZyme families	19

1 Test set composition

SI Table 1: Genomes used to compile test sets (overleaf)

SI Table 1 lists the taxonomic classification of each genomic assembly used to generate test sets to evaluate the CAZyme classifiers. The number of CAZymes listed per species (* encompassing all strains) and per genus is taken from the July 2023 CAZy database release. In the July 2023 CAZy database release: *Streptomyces antimycoticus* (GCA_009936315.1) is classified *Brettanomyces nanus*; *Brettanomyces nanus* (GCA_011074865.2) is classified as *Brettanomyces bruxellensis* UCD 2041; and *Nibricoccus aquaticus* (GCA_002310495.1) is classified as *Verrucomicrobiota bacterium* HZ-65 .

NCBI:txid	Assembly name	Version accession	Organism	Kingdom	Group	Number of CAZymes for species in CAZy*	Number of CAZymes for genus in CAZy*
515635	ASM2164v1	GCA_000021645.1	Dictyoglomus turgidum DSM 6724	Bacteria	Gram negative	102	208
509190	ASM9228v1	GCA_000092285.1	Caulobacter segnis ATCC 21756	Bacteria	Gram negative; a-proteobacteria	586	1975
192	ASM131501v1	GCA_001315015.1	Azospirillum brasilense	Bacteria	Gram negative; a-proteobacteria	1550	3057
661488	ASM797018v1	GCA_007970185.1	Pseudobacter ginsenosidimutans	Bacteria	Gram negative; CFB group bacteria	237	237
562	ASM522158v1	GCA_005221585.1	Escherichia coli	Bacteria	Gram negative; E. coli	110387	122982
562	ASM522190v1	GCA_005221905.1	Escherichia coli	Bacteria	Gram negative; E. coli	110387	122982
561229	ASM2356v1	GCA_000023565.1	Dickeya chrysanthemi Ech1591	Bacteria	Gram negative; Enterobacteriaceae	226	6498
1334564	ASM82877v1	GCA_000828775.1	Serratia marcescens SM39	Bacteria	Gram negative; Enterobacteriaceae	819	17744
571	ASM290639v1	GCA_002906395.1	Klebsiella oxytoca	Bacteria	Gram negative; Enterobacteriaceae	5943	118339
1940567	ASM340313v1	GCA_003403135.1	Dickeya dianthicola	Bacteria	Gram negative; Enterobacteriaceae	1811	6498
61645	ASM394076v1	GCA_003940765.1	Enterobacter asburiae	Bacteria	Gram negative; Enterobacteriaceae	3772	38478
1134687	ASM1009300v1	GCA_010093005.1	Klebsiella michiganensis	Bacteria	Gram negative; Enterobacteriaceae	9145	118339
548	ASM1160472v1	GCA_011604725.1	Klebsiella aerogenes	Bacteria	Gram negative; Enterobacteriaceae	6884	118339
615	ASM1342615v1	GCA_013426155.1	Serratia marcescens	Bacteria	Gram negative; Enterobacteriaceae	8192	17744
59203	31885_G02	GCA_900635675.1	Salmonella enterica subsp. arizonae	Bacteria	Gram negative; Enterobacteriaceae	48446	51679
498211	ASM1922v1	GCA_000019225.1	Cellvibrio japonicus Ueda107	Bacteria	Gram negative; g-proteobacteria	912	2091
1137651	ASM34922v1	GCA_000349225.1	Xanthomonas citri subsp. citri Aw12879	Bacteria	Gram negative; g-proteobacteria	18260	70327
1308541	ASM81688v1	GCA_000816885.1	Xanthomonas citri subsp. citri A306	Bacteria	Gram negative; g-proteobacteria	18260	70327
1583341	PCPL58T	GCA_900074915.1	Pseudomonas cerasi	Bacteria	Gram negative; g-proteobacteria	304	97188
718	55685_B01	GCA_900638075.1	Actinobacillus equuli	Bacteria	Gram negative; g-proteobacteria	1688	4578
1637999	ASM97276v1	GCA_000972765.1	Verrucomicrobia bacterium IMCC26134	Bacteria	Gram negative; verrucomicrobia	375	375
2026799	ASM231049v1	GCA_002310495.1	Nibricoccus aquaticus	Bacteria	Gram negative; verrucomicrobia	284	284
203119	ASM1586v1	GCA_000015865.1	Acetivibrio thermocellus ATCC 27405	Bacteria	Gram positive; firmicutes	828	1111
394503	ASM2206v1	GCA_000022065.1	Ruminiclostridium cellulolyticum H10	Bacteria	Gram positive; firmicutes	168	524
720554	ASM23708v1	GCA_000237085.1	Acetivibrio clariflavus DSM 19732	Bacteria	Gram positive; firmicutes	145	1111
1520	ASM83310v2	GCA_000833105.2	Clostridium beijerinckii	Bacteria	Gram positive; firmicutes	1427	28279
1292358	ASM83514v1	GCA_000835145.1	Bacillus amyloliquefaciens KHG19	Bacteria	Gram positive; firmicutes	7463	111030
36745	ASM200330v1	GCA_002003305.1	Clostridium saccharoperbutylacetonicum	Bacteria	Gram positive; firmicutes	453	28279
1352	ASM202504v1	GCA_002025045.1	Enterococcus faecium	Bacteria	Gram positive; firmicutes	11568	27760
304207	ASM869410v1	GCA_008694105.1	Schleiferilactobacillus harbinensis	Bacteria	Gram positive; firmicutes	435	435
37734	ASM970734v1	GCA_009707345.1	Enterococcus casseliflavus	Bacteria	Gram positive; firmicutes	1368	22760
1429244	ASM50720v2	GCA_000507205.2	Paenibacillus polymyxa CR1	Bacteria	Gram positive; firmicutes	8515	36966
1039	ASM1479206v1	GCA_014792065.1	Bacillus amyloliquefaciens	Bacteria	Gram positive; firmicutes	7463	111030
2665646	ASM1640612v1	GCA_016406125.1	Alicyclobacillus sp. SO9	Bacteria	Gram positive; firmicutes	442	1232
479432	ASM2486v1	GCA_000024865.1	Streptosporangium roseum DSM 43021	Bacteria	Gram positive; high G+C	257	605
749414	ASM9238v1	GCA_000092385.1	Streptomyces bingchenggensis BCW-1	Bacteria	Gram positive; high G+C	387	100828
212767	ASM32856v1	GCA_000328565.1	Mycobacterium sp. JS623	Bacteria	Gram positive; high G+C	9195	30430
284038	ASM993631v1	GCA_009936315.1	Streptomyces antimycoticus	Bacteria	Gram positive; high G+C	140	100828
228602	ASM1180114v1	GCA_011801145.1	Nocardia arthritidis	Bacteria	Gram positive; high G+C	180	3932
2704468	ASM1408410v1	GCA_014084105.1	Streptacidiphilus sp. P02-A3a	Bacteria	Gram positive; high G+C	449	499
332648	ASM14353v4	GCA_000143535.4	Botrytis cinerea B05.10	Eukaryote	Ascomycete fungi	1309	1371
403677	ASM159280v2	GCA_001592805.2	Peltaster fructicola	Eukaryote	Ascomycete fungi	266	266
318829	ASM434696v1	GCA_004346965.1	Pyricularia oryzae	Eukaryote	Ascomycete fungi	1757	1825
73501	ASM808049v1	GCA_008080495.1	Cordyceps militaris	Eukaryote	Ascomycete fungi	323	356
227321	ASM901741v1	GCA_009017415.1	Aspergillus flavus	Eukaryote	Ascomycete fungi	2516	7327
660027	ASM1308505v1	GCA_013085055.1	Fusarium oxysporum Fo47	Eukaryote	Ascomycete fungi	1827	8833
500148	ASM1342620v1	GCA_013426205.1	Metarhizium brunneum	Eukaryote	Ascomycete fungi	393	490
36651	ASM1676781v1	GCA_016767815.1	Penicillium digitatum	Eukaryote	Ascomycete fungi	345	1037
182096	AchevalieriM1_assembly01	GCA_016861735.1	Aspergillus chevalieri	Eukaryote	Ascomycete fungi	311	7327
101028	ASM1695230v1	GCA_016952305.1	Fusarium pseudograminearum	Eukaryote	Ascomycete fungi	1019	8833
5516	ASM1695235v1	GCA_016952355.1	Fusarium culmorum	Eukaryote	Ascomycete fungi	485	8833
2747967	ASM2310122v1	GCA_023101225.1	Fusarium solani-melongenae CRI 24-3	Eukaryote	Ascomycete fungi	681	8833
5059	ASM1478422v2	GCA_014784225.2	Aspergillus flavus CA14	Eukaryote	Ascomycete fungi	2516	7327
5499	Cfulv_R5_v5	GCA_020509005.2	Fulvia fulva Race5_Kim	Eukaryote	Ascomycete fungi	576	576
63577	ASM2064779v1	GCA_020647795.1	Trichoderma atroviride P1	Eukaryote	Ascomycete fungi	586	3361
101201	ASM2064786v1	GCA_020647865.1	Trichoderma asperellum FT101	Eukaryote	Ascomycete fungi	550	3361
1491479	ASM1956561v1	GCA_019565615.1	Trichoderma simmonsii GH-Sj1	Eukaryote	Ascomycete fungi	483	3361
170446	ASM1690657v1	GCA_016906575.1	Ceratobasidium sp. AG-Ba	Eukaryote	Basidiomycota fungi	1287	1918
284590	ASM251v1	GCA_000002515.1	Kluyveromyces lactis	Eukaryote	Budding yeasts	296	730
573826	ASM2694v1	GCA_000026945.1	Candida dubliniensis CD36	Eukaryote	Budding yeasts	146	3700
284811	ASM9102v4	GCA_000091025.4	Eremothecium gossypii ATCC 10895	Eukaryote	Budding yeasts	270	477
796027	ASM164002v2	GCA_001640025.2	Sugiyamaella lignohabitans	Eukaryote	Budding yeasts	150	150
4911	ASM185444v2	GCA_001854445.2	Kluyveromyces marxianus	Eukaryote	Budding yeasts	434	730
1365886	ASM198439v2	GCA_001984395.2	Zygosaccharomyces parabaillii	Eukaryote	Budding yeasts	220	450
498019	ASM301371v2	GCA_003013715.2	[Candida] auris	Eukaryote	Budding yeasts	657	3700
4909	ASM305444v1	GCA_003054445.1	Pichia kudriavzevii	Eukaryote	Budding yeasts	188	229
2163413	ASM421770v1	GCA_004217705.1	Metschnikowia aff. pulcherrima	Eukaryote	Budding yeasts	135	144
28985	ASM799369v1	GCA_007993695.1	Kluyveromyces lactis	Eukaryote	Budding yeasts	296	730
498019	ASM827514v1	GCA_008275145.1	[Candida] auris	Eukaryote	Budding yeasts	657	3700
36911	ASM949811v1	GCA_009498115.1	Clavispora lusitaniae	Eukaryote	Budding yeasts	686	686
13502	ASM1107486v2	GCA_011074865.2	Brettanomyces nanus	Eukaryote	Budding yeasts	138	279
5007	ASM1107488v2	GCA_011074885.2	Brettanomyces bruxellensis	Eukaryote	Budding yeasts	138	279
5478	ASM1421772v1	GCA_014217725.1	Nakaseomyces glabratus	Eukaryote	Budding yeasts	2152	2252
4652	ASM1449061v1	GCA_014490615.1	Yarrowia lipolytic	Eukaryote	Budding yeasts	407	407
230603	ASM2755758v1	GCA_027557585.1	Saccharomyces uvarum CBS7001	Eukaryote	Budding yeasts	541	17183
296587	ASM9098v2	GCA_000090985.2	Micromonas commoda	Eukaryote	Green algae	150	152
436017	ASM9206v1	GCA_000092065.1	Ostreococcus lucimarinus CCE9901	Eukaryote	Green algae	116	246
1764295	ASM785969v1	GCA_007859695.1	Chloropicon primus	Eukaryote	Green algae	443	443
6239	WBcel235	GCA_000002985.3	Caenorhabditis elegans BRISTOL N2	Eukaryote	Nematodes	1340	1880
573729	ASM22609v1	GCA_000226095.1	Thermothelomyces thermophilus ATCC 42464	Eukaryote	Thermophile	400	401

SI Table 2: Coverage of CAZy families (overleaf)

SI Table 2 lists the number of unique NCBI protein version accessions associated with each CAZy family in the CAZy database July 2023 release, as well as the total number and percentage of proteins from each CAZy family included across the test sets. CAZy families are grouped by their respective parent CAZy class.

2 Evaluation of CAZyme/non-CAZyme classification

2.1 Summary of CAZyme/non-CAZyme classification

2.2 Output of testing for statistically significant difference in performance between the tools

SI Figure 1: Evaluation of binary CAZyme/non-CAZyme classification (overleaf)

SI figure ?? plots the value of each statistical parameter for each test, per CAZyme classifier.

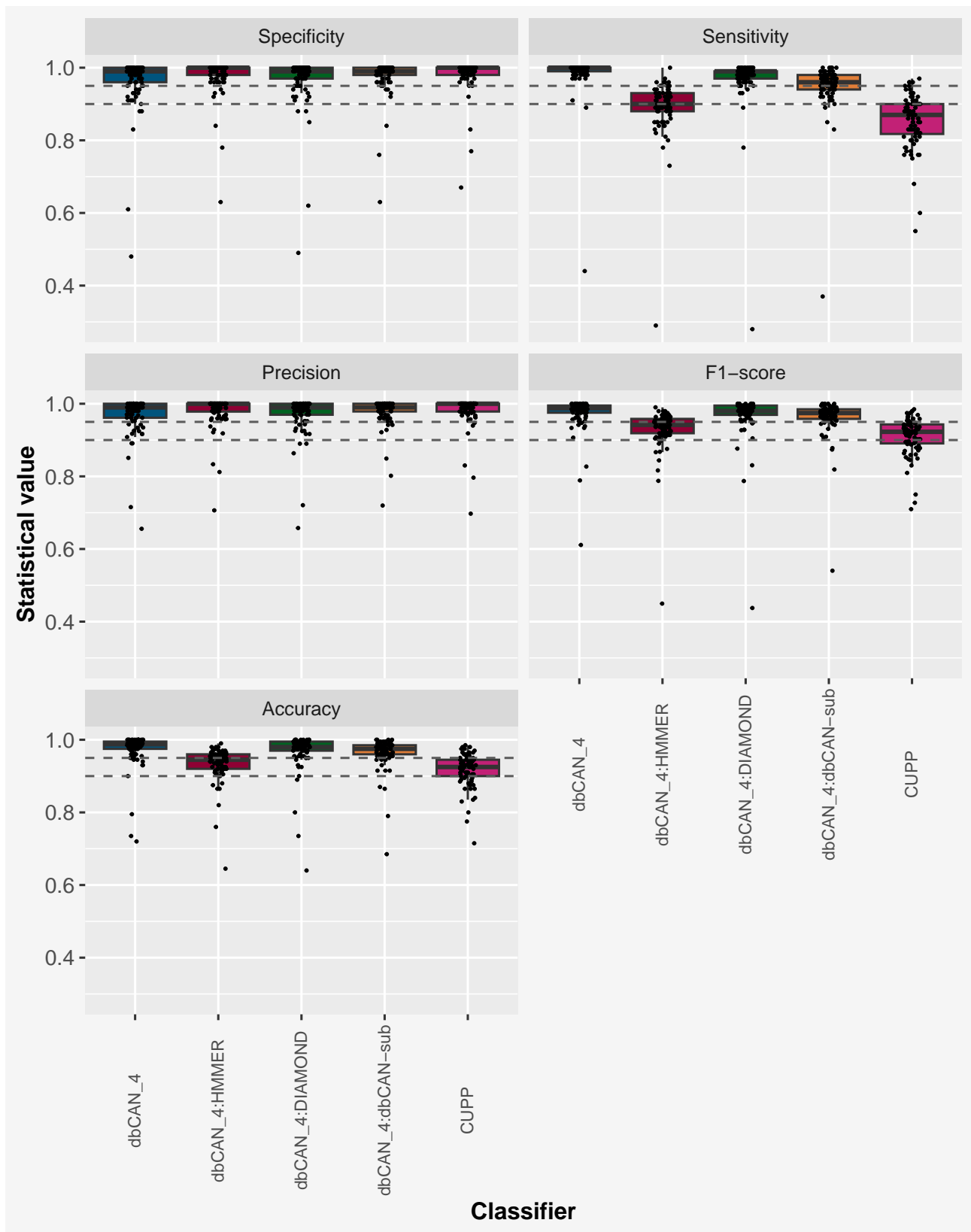


Figure 1: One-dimensional scatter plot overlaying a box and whisker plot, where each scatter plot point represents a test set and the corresponding statistical parameter value. Dashed lines indicate statistical values of 0.9 and 0.95.

??

SI Table 3: Tukey HSD test to measure the statistically significant difference between the mean F1-score for CAZyme/non-CAZyme classification (overleaf)

Tukey HSD test of the mean F1-score between the classifiers, reporting the lower and upper 95% confidence interval, and the adjusted p-value (P-Adj).

	Mean Difference	Lower	Upper	P-Adj
dbCAN_4:HMMER-dbCAN_4	-0.04391	-0.06969	-0.01813	4.09E-05
dbCAN_4:DIAMOND-dbCAN_4	-0.00668	-0.03246	0.019095	0.954057
dbCAN_4:dbCAN-sub-dbCAN_4	-0.01143	-0.03721	0.014346	0.742466
CUPP-dbCAN_4	-0.06232	-0.08809	-0.03654	1.12E-09
dbCAN_4:DIAMOND-dbCAN_4:HMMER	0.037227	0.011449	0.063005	0.00085
dbCAN_4:dbCAN-sub-dbCAN_4:HMMER	0.032478	0.0067	0.058256	0.005531
CUPP-dbCAN_4:HMMER	-0.01841	-0.04418	0.007373	0.28939
dbCAN_4:dbCAN-sub-dbCAN_4:DIAMOND	-0.00475	-0.03053	0.021029	0.986863
CUPP-dbCAN_4:DIAMOND	-0.05563	-0.08141	-0.02985	7.21E-08
CUPP-dbCAN_4:dbCAN-sub	-0.05088	-0.07666	-0.02511	1.09E-06

SI Table 4: Tukey HSD test to measure the statistically significant difference between the mean sensitivity for CAZyme/non-CAZyme classification (overleaf)

Tukey HSD test of the mean sensitivity between the classifiers, reporting the lower and upper 95% confidence interval, and the adjusted p-value (P-Adj).

	Mean Difference	Lower	Upper	P-Adj
dbCAN_4:HMMER-dbCAN_4	-0.09412	-0.12691	-0.06134	0
dbCAN_4:DIAMOND-dbCAN_4	-0.01475	-0.04753	0.018034	0.732105
dbCAN_4:dbCAN-sub-dbCAN_4	-0.03438	-0.06716	-0.00159	0.034555
CUPP-dbCAN_4	-0.13025	-0.16303	-0.09747	0
dbCAN_4:DIAMOND-dbCAN_4:HMMER	0.079375	0.046591	0.112159	1.05E-09
dbCAN_4:dbCAN-sub-dbCAN_4:HMMER	0.05975	0.026966	0.092534	8.77E-06
CUPP-dbCAN_4:HMMER	-0.03613	-0.06891	-0.00334	0.022499
dbCAN_4:dbCAN-sub-dbCAN_4:DIAMOND	-0.01963	-0.05241	0.013159	0.472431
CUPP-dbCAN_4:DIAMOND	-0.1155	-0.14828	-0.08272	0
CUPP-dbCAN_4:dbCAN-sub	-0.09587	-0.12866	-0.06309	0

SI Table 5: Tukey HSD test to measure the statistically significant difference between the mean accuracy for CAZyme/non-CAZyme classification (overleaf)

Tukey HSD test of the mean accuracy between the classifiers, reporting the lower and upper 95% confidence interval, and the adjusted p-value (P-Adj).

	Mean Difference	Lower	Upper	P-Adj
dbCAN_4:HMMER-dbCAN_4	-0.03881	-0.05992	-0.01771	7.03E-06
dbCAN_4:DIAMOND-dbCAN_4	-0.00512	-0.02623	0.015979	0.963624
dbCAN_4:dbCAN-sub-dbCAN_4	-0.00956	-0.03067	0.011542	0.726883
CUPP-dbCAN_4	-0.05538	-0.07648	-0.03427	1.44E-11
dbCAN_4:DIAMOND-dbCAN_4:HMMER	0.033688	0.012583	0.054792	0.000151
dbCAN_4:dbCAN-sub-dbCAN_4:HMMER	0.02925	0.008146	0.050354	0.001577
CUPP-dbCAN_4:HMMER	-0.01656	-0.03767	0.004542	0.200982
dbCAN_4:dbCAN-sub-dbCAN_4:DIAMOND	-0.00444	-0.02554	0.016667	0.978491
CUPP-dbCAN_4:DIAMOND	-0.05025	-0.07135	-0.02915	2.07E-09
CUPP-dbCAN_4:dbCAN-sub	-0.04581	-0.06692	-0.02471	5.94E-08

3 Taxonomic performance of CAZyme/non-CAZyme classification

3.1 Summary of taxonomic kingdom performance of CAZyme/non-CAZyme classification

SI Figure 2: The specificity of CAZyme/non-CAZyme classification per taxonomic kingdom for each test set

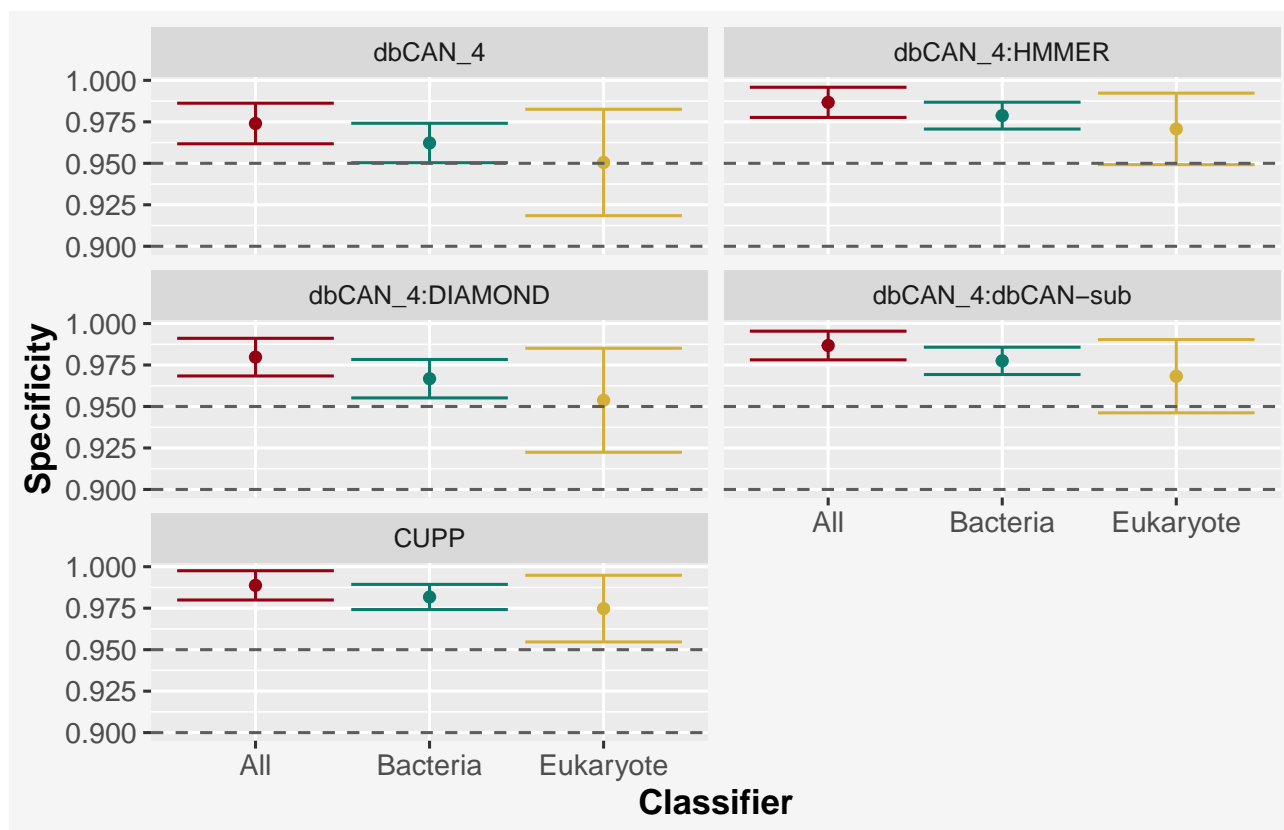


Figure 2: Mean and 95% confidence interval of the specificity across all test sets per taxonomic kingdom (Bacteria shaded green; Eukaryote shaded Yellow; Both/All shaded red) for the CAZyme/non-CAZyme classification of protein sequences.

SI Figure 3: The precision of CAZyme/non-CAZyme classification per taxonomic kingdom for each test set

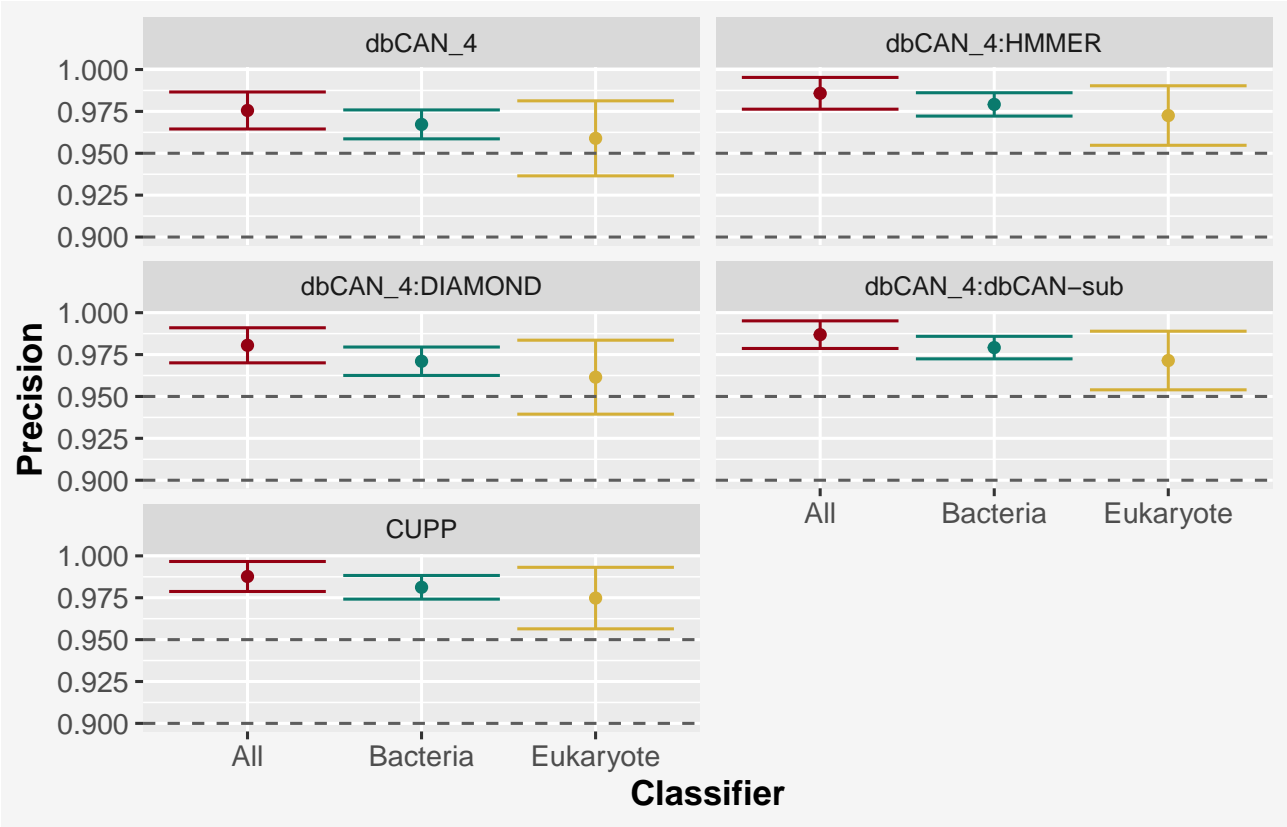


Figure 3: Mean and 95% confidence interval of the precision across all test sets per taxonomic kingdom (Bacteria shaded green; Eukaryote shaded Yellow; Both/All shaded red) for the CAZyme/non-CAZyme classification of protein sequences.

SI Figure 4: The accuracy of CAZyme/non-CAZyme classification per taxonomic kingdom for each test set

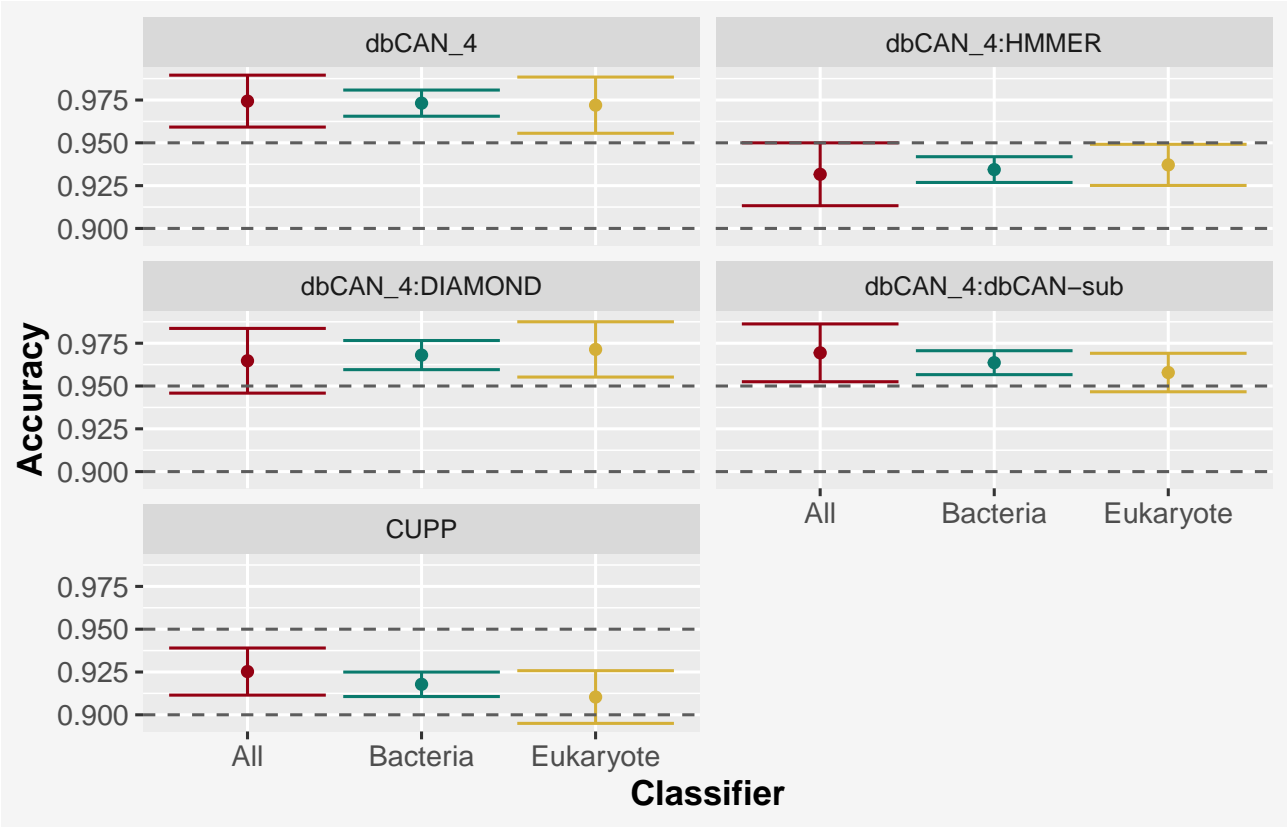


Figure 4: Mean and 95% confidence interval of the accuracy across all test sets per taxonomic kingdom (Bacteria shaded green; Eukaryote shaded Yellow; Both/All shaded red) for the CAZyme/non-CAZyme classification of protein sequences.

4 Evaluation of CAZy class classification across all CAZy class

SI Table 6: Tukey HSD test to measure the statistically significant difference between the mean F1-score for CAZy class classification

Tukey HSD test of the mean F1-score between the classifiers, evaluating CAZyme class classification after aggregating all CAZyme classes, reporting the lower and upper 95% confidence interval, and the adjusted p-value (P-Adj).

	Mean Difference	Lower	Upper	P-Adj
dbCAN_4:HMMER-dbCAN_4	-0.06649	-0.10568	-0.0273	3.78E-05
dbCAN_4:DIAMOND-dbCAN_4	0.010131	-0.02906	0.04932	0.955216
dbCAN_4:dbCAN-sub-dbCAN_4	-0.00071	-0.0399	0.038481	0.999999
CUPP-dbCAN_4	-0.20421	-0.2434	-0.16502	5.55E-11
dbCAN_4:DIAMOND-dbCAN_4:HMMER	0.076621	0.037432	0.115809	1.04E-06
dbCAN_4:dbCAN-sub-dbCAN_4:HMMER	0.065781	0.026592	0.10497	4.77E-05
CUPP-dbCAN_4:HMMER	-0.13772	-0.17691	-0.09853	5.56E-11
dbCAN_4:dbCAN-sub-dbCAN_4:DIAMOND	-0.01084	-0.05003	0.028349	0.943195
CUPP-dbCAN_4:DIAMOND	-0.21434	-0.25353	-0.17515	5.55E-11
CUPP-dbCAN_4:dbCAN-sub	-0.2035	-0.24269	-0.16431	5.55E-11

SI Table 7: Tukey HSD test to measure the statistically significant difference between the mean sensitivity for CAZy class classification

Tukey HSD test of the mean sensitivity between the classifiers, evaluating CAZyme class classification after aggregating all CAZyme classes, reporting the lower and upper 95% confidence interval, and the adjusted p-value (P-Adj).

	Mean Difference	Lower	Upper	P-Adj
dbCAN_4:HMMER-dbCAN_4	-0.09245	-0.13391	-0.05098	1.37E-08
dbCAN_4:DIAMOND-dbCAN_4	0.022676	-0.01879	0.064138	0.567019
dbCAN_4:dbCAN-sub-dbCAN_4	0.001643	-0.03982	0.043104	0.999969
CUPP-dbCAN_4	-0.22089	-0.26235	-0.17942	5.55E-11
dbCAN_4:DIAMOND-dbCAN_4:HMMER	0.115122	0.07366	0.156583	5.61E-11
dbCAN_4:dbCAN-sub-dbCAN_4:HMMER	0.094089	0.052627	0.13555	7.05E-09
CUPP-dbCAN_4:HMMER	-0.12844	-0.1699	-0.08698	5.56E-11
dbCAN_4:dbCAN-sub-dbCAN_4:DIAMOND	-0.02103	-0.06249	0.020428	0.637387
CUPP-dbCAN_4:DIAMOND	-0.24356	-0.28502	-0.2021	5.55E-11
CUPP-dbCAN_4:dbCAN-sub	-0.22253	-0.26399	-0.18107	5.55E-11

SI Table 8: Tukey HSD test to measure the statistically significant difference between the mean accuracy for CAZy class classification

Tukey HSD test of the mean accuracy between the classifiers, evaluating CAZyme class classification after aggregating all CAZyme classes, reporting the lower and upper 95% confidence interval, and the adjusted p-value (P-Adj).

	Mean Difference	Lower	Upper	P-Adj
dbCAN_4:HMMER-dbCAN_4	-0.01489	-0.02277	-0.00702	2.66E-06
dbCAN_4:DIAMOND-dbCAN_4	0.003211	-0.00467	0.011088	0.799843
dbCAN_4:dbCAN-sub-dbCAN_4	-0.00056	-0.00844	0.007314	0.999678
CUPP-dbCAN_4	-0.03001	-0.03788	-0.02213	5.56E-11
dbCAN_4:DIAMOND-dbCAN_4:HMMER	0.018104	0.010227	0.025981	4.31E-09
dbCAN_4:dbCAN-sub-dbCAN_4:HMMER	0.01433	0.006453	0.022207	7.29E-06
CUPP-dbCAN_4:HMMER	-0.01511	-0.02299	-0.00724	1.78E-06
dbCAN_4:dbCAN-sub-dbCAN_4:DIAMOND	-0.00377	-0.01165	0.004104	0.686372
CUPP-dbCAN_4:DIAMOND	-0.03322	-0.04109	-0.02534	5.56E-11
CUPP-dbCAN_4:dbCAN-sub	-0.02944	-0.03732	-0.02157	5.56E-11

SI Table 9: Tukey HSD test to measure the statistically significant difference between the mean precision for CAZy class classification

Tukey HSD test of the mean precision between the classifiers, evaluating CAZyme class classification after aggregating all CAZyme classes, reporting the lower and upper 95% confidence interval, and the adjusted p-value (P-Adj).

	Mean Difference	Lower	Upper	P-Adj
dbCAN_4:HMMER-dbCAN_4	-0.01489	-0.02277	-0.00702	2.66E-06
dbCAN_4:DIAMOND-dbCAN_4	0.003211	-0.00467	0.011088	0.799843
dbCAN_4:dbCAN-sub-dbCAN_4	-0.00056	-0.00844	0.007314	0.999678
CUPP-dbCAN_4	-0.03001	-0.03788	-0.02213	5.56E-11
dbCAN_4:DIAMOND-dbCAN_4:HMMER	0.018104	0.010227	0.025981	4.31E-09
dbCAN_4:dbCAN-sub-dbCAN_4:HMMER	0.01433	0.006453	0.022207	7.29E-06
CUPP-dbCAN_4:HMMER	-0.01511	-0.02299	-0.00724	1.78E-06
dbCAN_4:dbCAN-sub-dbCAN_4:DIAMOND	-0.00377	-0.01165	0.004104	0.686372
CUPP-dbCAN_4:DIAMOND	-0.03322	-0.04109	-0.02534	5.56E-11
CUPP-dbCAN_4:dbCAN-sub	-0.02944	-0.03732	-0.02157	5.56E-11

5 Evaluation of CAZy class classification per CAZy class

SI Table 10: Tukey HSD test to measure the statistically significant difference between the mean F1-score between CAZy classes and prediction tools

Tukey HSD test of the mean F1-score between the classifiers and CAZyme classes, reporting the lower and upper 95% confidence interval, and the adjusted p-value (P-Adj).

Showing only hits with a P-value ≤ 0.05 where the classifier is the same and the CAZyme class is different.

Class 1	Class 2	Classifier 1	Classifier 2	Mean Difference	Lower 95% Confidence Interval	Upper 95% Confidence Interval	Adjusted P-value
CBM	AA	CUPP	CUPP	-0.90105	-0.97347	-0.82863	5.07E-11
CE	CBM	CUPP	CUPP	0.925011	0.861476	0.988547	5.07E-11
GH	CBM	CUPP	CUPP	0.946062	0.883346	1.008778	5.07E-11
GT	CBM	CUPP	CUPP	0.910716	0.848	0.973432	5.07E-11
PL	CBM	CUPP	CUPP	0.885018	0.81212	0.957915	5.07E-11
GH	CBM	dbCAN_4:DIAMOND	dbCAN_4:DIAMOND	0.069222	0.006506	0.131937	0.011938
GT	CBM	dbCAN_4:DIAMOND	dbCAN_4:DIAMOND	0.067879	0.005163	0.130594	0.016339
PL	CBM	dbCAN_4:DIAMOND	dbCAN_4:DIAMOND	0.075289	0.002391	0.148187	0.032579
CBM	AA	dbCAN_4:HMMER	dbCAN_4:HMMER	-0.34806	-0.42047	-0.27564	5.07E-11
CE	CBM	dbCAN_4:HMMER	dbCAN_4:HMMER	0.353329	0.289794	0.416865	5.07E-11
GH	CBM	dbCAN_4:HMMER	dbCAN_4:HMMER	0.356254	0.293538	0.418969	5.07E-11
GT	CBM	dbCAN_4:HMMER	dbCAN_4:HMMER	0.32167	0.258954	0.384386	5.07E-11
PL	CBM	dbCAN_4:HMMER	dbCAN_4:HMMER	0.378704	0.305806	0.451602	5.07E-11
CBM	AA	dbCAN_4:sub	dbCAN_4:sub	-0.08535	-0.15777	-0.01293	0.003723
CE	CBM	dbCAN_4:sub	dbCAN_4:sub	0.104081	0.040545	0.167616	4.02E-07
GH	CBM	dbCAN_4:sub	dbCAN_4:sub	0.107924	0.045208	0.170639	5.66E-08
GT	CBM	dbCAN_4:sub	dbCAN_4:sub	0.108752	0.046037	0.171468	4.10E-08
PL	CBM	dbCAN_4:sub	dbCAN_4:sub	0.1171	0.044202	0.189998	8.34E-07
CBM	AA	dbCAN_4	dbCAN_4	-0.09171	-0.16413	-0.0193	0.000834
CE	CBM	dbCAN_4	dbCAN_4	0.109065	0.04553	0.172601	6.27E-08
GH	CBM	dbCAN_4	dbCAN_4	0.112331	0.049615	0.175047	9.99E-09
GT	CBM	dbCAN_4	dbCAN_4	0.113027	0.050312	0.175743	7.56E-09
PL	CBM	dbCAN_4	dbCAN_4	0.119362	0.046464	0.192259	4.09E-07

SI Table 11: Tukey HSD test to measure the statistically significant difference between the mean F1-score between CAZy classes and prediction tools

Tukey HSD test of the mean F1-score between the classifiers and CAZyme classes, reporting the lower and upper 95% confidence interval, and the adjusted p-value (P-Adj).

Showing only hits with a P-value ≤ 0.05 where the CAZyme class is the same and the classifier is different.

Class 1	Class 2	Classifier 1	Classifier 2	Mean Difference	Lower 95% Confidence Interval	Upper 95% Confidence Interval	Adjusted P-value
GT	GT	CUPP	dbCAN_4:DIAMOND	-0.06433	-0.12704	-0.00161	0.035824
PL	PL	CUPP	dbCAN_4	-0.08906	-0.17089	-0.00724	0.014939
PL	PL	CUPP	dbCAN_4:HMMER	-0.08906	-0.17089	-0.00724	0.014939
PL	PL	CUPP	dbCAN_4:DIAMOND	-0.09744	-0.17926	-0.01561	0.003047
PL	PL	CUPP	dbCAN_4:sub	-0.08971	-0.17153	-0.00789	0.013311
CBM	CBM	dbCAN_4:HMMER	dbCAN_4	-0.25934	-0.32206	-0.19663	5.07E-11
CBM	CBM	CUPP	dbCAN_4	-0.85472	-0.91744	-0.792	5.07E-11
CBM	CBM	dbCAN_4:DIAMOND	dbCAN_4:HMMER	0.311787	0.249071	0.374503	5.07E-11
CBM	CBM	dbCAN_4:sub	dbCAN_4:HMMER	0.262249	0.199533	0.324964	5.07E-11
CBM	CBM	CUPP	dbCAN_4:HMMER	-0.59538	-0.65809	-0.53266	5.07E-11
CBM	CBM	CUPP	dbCAN_4:DIAMOND	-0.90716	-0.96988	-0.84445	5.07E-11
CBM	CBM	CUPP	dbCAN_4:sub	-0.85763	-0.92034	-0.79491	5.07E-11

SI Table 12: Tukey HSD test to measure the statistically significant difference between the mean sensitivity between CAZy classes and prediction tools

Tukey HSD test of the mean sensitivity between the classifiers and CAZyme classes, reporting the lower and upper 95% confidence interval, and the adjusted p-value (P-Adj).

Showing only hits with a P-value ≤ 0.05 where the CAZyme class is the same and the classifier is different.

Class 1	Class 2	Classifier 1	Classifier 2	Mean Difference	Lower 95% Confidence Interval	Upper 95% Confidence Interval	Adjusted P-value
GT	GT	dbCAN_4:HMMER	dbCAN_4	-0.09213	-0.16198	-0.02228	0.000321
GT	GT	CUPP	dbCAN_4	-0.10425	-0.1741	-0.0344	1.01E-05
GT	GT	dbCAN_4:DIAMOND	dbCAN_4:HMMER	0.109403	0.039555	0.179252	2.05E-06
GT	GT	dbCAN_4:sub	dbCAN_4:HMMER	0.08812	0.018271	0.157968	0.00091
GT	GT	CUPP	dbCAN_4:DIAMOND	-0.12153	-0.19138	-0.05168	3.56E-08
GT	GT	CUPP	dbCAN_4:sub	-0.10024	-0.17009	-0.0304	3.34E-05
PL	PL	CUPP	dbCAN_4	-0.12285	-0.21398	-0.03172	0.000186
PL	PL	CUPP	dbCAN_4:HMMER	-0.12285	-0.21398	-0.03172	0.000186
PL	PL	CUPP	dbCAN_4:DIAMOND	-0.14767	-0.2388	-0.05655	6.04E-07
PL	PL	CUPP	dbCAN_4:sub	-0.1241	-0.21523	-0.03297	0.000143
CBM	CBM	dbCAN_4:HMMER	dbCAN_4	-0.32544	-0.39529	-0.2556	5.07E-11
CBM	CBM	dbCAN_4:DIAMOND	dbCAN_4	0.077849	0.008001	0.147698	0.010154
CBM	CBM	CUPP	dbCAN_4	-0.79948	-0.86933	-0.72964	5.07E-11
CBM	CBM	dbCAN_4:DIAMOND	dbCAN_4:HMMER	0.403294	0.333445	0.473142	5.07E-11
CBM	CBM	dbCAN_4:sub	dbCAN_4:HMMER	0.345283	0.275434	0.415131	5.07E-11
CBM	CBM	CUPP	dbCAN_4:HMMER	-0.47404	-0.54389	-0.40419	5.07E-11
CBM	CBM	CUPP	dbCAN_4:DIAMOND	-0.87733	-0.94718	-0.80749	5.07E-11
CBM	CBM	CUPP	dbCAN_4:sub	-0.81932	-0.88917	-0.74947	5.07E-11

SI Table 13: Tukey HSD test to measure the statistically significant difference between the mean precision between CAZy classes and prediction tools

Tukey HSD test of the mean precision between the classifiers and CAZyme classes, reporting the lower and upper 95% confidence interval, and the adjusted p-value (P-Adj).

Showing only hits with a P-value ; 0.05 where the CAZyme class is the same and the classifier is different.

Class 1	Class 2	Classifier 1	Classifier 2	Mean Difference	Lower 95% Confidence Interval	Upper 95% Confidence Interval	Adjusted P-value
CBM	CBM	CUPP	dbCAN_4	-0.95625	-1.01961	-0.89289	5.07E-11
CBM	CBM	CUPP	dbCAN_4:HMMER	-0.9184	-0.98176	-0.85504	5.07E-11
CBM	CBM	CUPP	dbCAN_4:DIAMOND	-0.95711	-1.02047	-0.89375	5.07E-11
CBM	CBM	CUPP	dbCAN_4:sub	-0.93759	-1.00095	-0.87423	5.07E-11

SI Table 14: Tukey HSD test to measure the statistically significant difference between the mean accuracy between CAZy classes and prediction tools

Tukey HSD test of the mean accuracy between the classifiers and CAZyme classes, reporting the lower and upper 95% confidence interval, and the adjusted p-value (P-Adj).

Showing only hits with a P-value ; 0.05 where the CAZyme class is the same and the classifier is different.

Class 1	Class 2	Classifier 1	Classifier 2	Mean Difference	Lower 95% Confidence Interval	Upper 95% Confidence Interval	Adjusted P-value
GT	GT	dbCAN_4:HMMER	dbCAN_4	-0.02787	-0.04925	-0.00649	0.000427
GT	GT	CUPP	dbCAN_4	-0.03193	-0.05331	-0.01055	9.95E-06
GT	GT	dbCAN_4:DIAMOND	dbCAN_4:HMMER	0.033533	0.012153	0.054913	1.96E-06
GT	GT	dbCAN_4:sub	dbCAN_4:HMMER	0.027036	0.005656	0.048416	0.000863
GT	GT	CUPP	dbCAN_4:DIAMOND	-0.03759	-0.05897	-0.01621	2.26E-08
GT	GT	CUPP	dbCAN_4:sub	-0.0311	-0.05248	-0.00972	2.25E-05
GH	GH	CUPP	dbCAN_4:DIAMOND	-0.02725	-0.04863	-0.00587	0.000724
CBM	CBM	dbCAN_4:HMMER	dbCAN_4	-0.03267	-0.05405	-0.01129	4.74E-06
CBM	CBM	CUPP	dbCAN_4	-0.0912	-0.11258	-0.06982	5.07E-11
CBM	CBM	dbCAN_4:DIAMOND	dbCAN_4:HMMER	0.039355	0.017976	0.060735	2.85E-09
CBM	CBM	dbCAN_4:sub	dbCAN_4:HMMER	0.03251	0.01113	0.05389	5.58E-06
CBM	CBM	CUPP	dbCAN_4:HMMER	-0.05853	-0.07991	-0.03715	5.10E-11
CBM	CBM	CUPP	dbCAN_4:DIAMOND	-0.09788	-0.11926	-0.07651	5.07E-11
CBM	CBM	CUPP	dbCAN_4:sub	-0.09104	-0.11242	-0.06966	5.07E-11

6 Taxonomic performance of CAZy class classification across all CAZy classes

This section of the SI presents the data and figures for evaluation the performance of CAZy class classification across all CAZy classes.

SI table 15: Tukey HSD test to measure the statistically significant difference between the mean F1-score for CE CAZyme class classification (overleaf)

Tukey HSD test of the mean F1-score between the classifiers, reporting the lower and upper 95% confidence interval, and the adjusted p-value (P-Adj).

	Mean Difference	Lower	Upper	P-Adj
Eukaryote: dbCAN_4 - Bacteria: dbCAN_4	0.050244	-0.02676	0.127252	0.650193
Eukaryote: db- CAN_4:HMMER - Bacte- ria: dbCAN_4:HMMER	0.072597	-0.00441	0.149605	0.089366
Eukaryote: db- CAN_4:DIAMOND - Bacteria: db- CAN_4:DIAMOND	0.043919	-0.03309	0.120927	0.831691
Eukaryote :db- CAN_4:dbCAN-sub - Bac- teria: dbCAN_4:dbCAN- sub	0.049899	-0.02711	0.126907	0.661286
Eukaryote: CUPP - Bac- teria: CUPP	0.052876	-0.02413	0.129885	0.5634086

SI table 16: Tukey HSD test to measure the statistically significant difference between the mean F1-score for AA CAZyme class classification (overleaf)

Tukey HSD test of the mean F1-score between the classifiers, reporting the lower and upper 95% confidence interval, and the adjusted p-value (P-Adj).

	Mean Difference	Lower	Upper	P-Adj
Eukaryote: dbCAN_4 - Bacteria: dbCAN_4	-0.06949	-0.20561	0.066629	0.919319
Eukaryote: db- CAN_4:HMMER - Bacte- ria: dbCAN_4:HMMER	-0.07338	-0.2095	0.062736	0.8813
Eukaryote: db- CAN_4:DIAMOND - Bacteria: db- CAN_4:DIAMOND	-0.05068	-0.1868	0.085437	0.995013
Eukaryote: db- CAN_4:dbCAN-sub - Bac- teria: dbCAN_4:dbCAN- sub	-0.07398	-0.2101	0.062141	0.874648
Eukaryote: CUPP - Bac- teria: CUPP	-0.12837	-0.26449	0.007754	0.088815

SI table 17: Tukey HSD test to measure the statistically significant difference between the mean F1-score for CBM CAZyme class classification (overleaf)

Tukey HSD test of the mean F1-score between the classifiers, reporting the lower and upper 95% confidence interval, and the adjusted p-value (P-Adj).

	Mean Difference	Lower	Upper	P-Adj
Eukaryote: dbCAN_4 - Bacteria: dbCAN_4	-0.14981	-0.25398	-0.04564	0.000121
Eukaryote: db- CAN_4:HMMER - Bacte- ria: dbCAN_4:HMMER	-0.24181	-0.34598	-0.13764	0
Eukaryote: db- CAN_4:DIAMOND - Bacteria: db- CAN_4:DIAMOND	0.093003	-0.01117	0.197174	0.141534
Eukaryote: db- CAN_4:dbCAN-sub - Bac- teria: dbCAN_4:dbCAN- sub	-0.18155	-0.28572	-0.07738	4.67E-07
Eukaryote: CUPP - Bac- teria: CUPP	8.53E-16	-0.10417	0.104171	1

7 Overall performance of CAZy family classification (across all CAZy families and classes)

SI Table 18: Tukey HSD test to measure the statistically significant difference between the mean F1-score for CAZy family classification

Tukey HSD test of the mean F1-score between the classifiers, evaluating CAZyme family classification after aggregating all CAZyme classes, reporting the lower and upper 95% confidence interval, and the adjusted p-value (P-Adj).

	Mean Difference	Lower	Upper	P-Adj
dbCAN_4:HMMER-dbCAN_4	-0.03553	-0.10033	0.029273	0.564396
dbCAN_4:DIAMOND-dbCAN_4	0.037334	-0.02782	0.102486	0.520279
dbCAN_4:dbCAN-sub-dbCAN_4	-0.01065	-0.07555	0.054252	0.991671
CUPP-dbCAN_4	-0.23454	-0.29975	-0.16934	0
dbCAN_4:DIAMOND-dbCAN_4:HMMER	0.072864	0.00786	0.137867	0.019015
dbCAN_4:dbCAN-sub-dbCAN_4:HMMER	0.024881	-0.03987	0.089633	0.832207
CUPP-dbCAN_4:HMMER	-0.19901	-0.26407	-0.13396	0
dbCAN_4:dbCAN-sub-dbCAN_4:DIAMOND	-0.04798	-0.11308	0.017119	0.260215
CUPP-dbCAN_4:DIAMOND	-0.27188	-0.33728	-0.20648	0
CUPP-dbCAN_4:dbCAN-sub	-0.2239	-0.28905	-0.15874	0

SI Table 19: Tukey HSD test to measure the statistically significant difference between the mean sensitivity for CAZy family classification

Tukey HSD test of the mean sensitivity between the classifiers, evaluating CAZyme family classification after aggregating all CAZyme classes, reporting the lower and upper 95% confidence interval, and the adjusted p-value (P-Adj).

	Mean Difference	Lower	Upper	P-Adj
dbCAN_4:HMMER-dbCAN_4	-0.04254	-0.1073	0.022217	0.377403
dbCAN_4:DIAMOND-dbCAN_4	0.037767	-0.02734	0.102874	0.507826
dbCAN_4:dbCAN-sub-dbCAN_4	-0.00958	-0.07443	0.05528	0.994441
CUPP-dbCAN_4	-0.25794	-0.32309	-0.19278	0
dbCAN_4:DIAMOND-dbCAN_4:HMMER	0.080307	0.015348	0.145266	0.006742
dbCAN_4:dbCAN-sub-dbCAN_4:HMMER	0.032964	-0.03174	0.097672	0.633438
CUPP-dbCAN_4:HMMER	-0.2154	-0.28041	-0.15039	0
dbCAN_4:dbCAN-sub-dbCAN_4:DIAMOND	-0.04734	-0.1124	0.017715	0.272619
CUPP-dbCAN_4:DIAMOND	-0.2957	-0.36106	-0.23034	0
CUPP-dbCAN_4:dbCAN-sub	-0.24836	-0.31347	-0.18325	0

8 Multi-label classification of CAZyme families

SI Table 20: Tukey HSD test to measure the statistically significant difference between the mean Adjusted Rand Index across prediction tools and taxonomic groups

Showing only hits with a P-value ≤ 0.05 where the CAZyme class is the same and the classifier is different.

Class 1	Class 2	Classifier 1	Classifier 2	Mean Difference	Lower 95% Confidence Interval	Upper 95% Confidence Interval	Adjusted P-value
Eukaryote	Eukaryote	dbCAN_4:HMMER	dbCAN_4	-0.02826	-0.04061	-0.01591	9.65E-13
Eukaryote	Eukaryote	dbCAN_4:DIAMOND	dbCAN_4	0.018578	0.006231	0.030924	3.42E-05
Eukaryote	Eukaryote	CUPP	dbCAN_4	-0.0528	-0.06514	-0.04045	0
Eukaryote	Eukaryote	dbCAN_4:DIAMOND	dbCAN_4:HMMER	0.046837	0.03449	0.059183	0
Eukaryote	Eukaryote	dbCAN_4:sub	dbCAN_4:HMMER	0.026545	0.014199	0.038892	3.23E-11
Eukaryote	Eukaryote	CUPP	dbCAN_4:HMMER	-0.02454	-0.03688	-0.01219	1.67E-09
Eukaryote	Eukaryote	dbCAN_4:sub	dbCAN_4:DIAMOND	-0.02029	-0.03264	-0.00794	2.59E-06
Eukaryote	Eukaryote	CUPP	dbCAN_4:DIAMOND	-0.07137	-0.08372	-0.05903	0
Eukaryote	Eukaryote	CUPP	dbCAN_4:sub	-0.05108	-0.06343	-0.03874	0
All	All	dbCAN_4:HMMER	dbCAN_4	-0.03178	-0.04051	-0.02305	0
All	All	dbCAN_4:DIAMOND	dbCAN_4	0.009785	0.001055	0.018516	0.012146
All	All	CUPP	dbCAN_4	-0.05268	-0.06141	-0.04395	0
All	All	dbCAN_4:DIAMOND	dbCAN_4:HMMER	0.041563	0.032833	0.050293	0
All	All	dbCAN_4:sub	dbCAN_4:HMMER	0.03232	0.023589	0.04105	0
All	All	CUPP	dbCAN_4:HMMER	-0.0209	-0.02963	-0.01217	2.32E-13
All	All	dbCAN_4:sub	dbCAN_4:DIAMOND	-0.00924	-0.01797	-0.00051	0.025872
All	All	CUPP	dbCAN_4:DIAMOND	-0.06246	-0.0712	-0.05373	0
All	All	CUPP	dbCAN_4:sub	-0.05322	-0.06195	-0.04449	0
Bacteria	Bacteria	dbCAN_4:HMMER	dbCAN_4	-0.0353	-0.04764	-0.02295	1.82E-13
Bacteria	Bacteria	CUPP	dbCAN_4	-0.05256	-0.06491	-0.04022	0
Bacteria	Bacteria	dbCAN_4:DIAMOND	dbCAN_4:HMMER	0.03629	0.023943	0.048636	1.25E-13
Bacteria	Bacteria	dbCAN_4:sub	dbCAN_4:HMMER	0.038094	0.025747	0.050441	1.43E-13
Bacteria	Bacteria	CUPP	dbCAN_4:HMMER	-0.01727	-0.02961	-0.00492	0.00021
Bacteria	Bacteria	CUPP	dbCAN_4:DIAMOND	-0.05356	-0.0659	-0.04121	0
Bacteria	Bacteria	CUPP	dbCAN_4:sub	-0.05536	-0.06771	-0.04301	0