4CAC: 4-class classifier of metagenome contigs using machine learning and assembly graphs

Lianrong Pu and Ron Shamir

The Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel.
lianrongpu@mail.tau.ac.il, rshamir@tau.ac.il

Supplementary Material

1 Supplementary tables

Table S1. Performance of 4CAC on simulated datasets created using the validation genomes with initial classifications generated by different score thresholds. To generate an initial four-way classification by XGBoost classifier, we tested score thresholds 0.8, 0.85, 0.9, and 0.95 to classify contigs as viruses and plasmids. Note that this table presents the results of the full 4CAC algorithm.

Datasets	Score threshold	Phage			Plasmid			Prokaryote			Eukaryote			All		
		precision	recall	F1 score	precision	recall	F1 score	precision	recall	F1 score	precision	recall	F1 score	precision	recall	F1 score
Sim_SG	0.95	0.84	0.69	0.76	0.56	0.55	0.55	0.92	0.86	0.89	0.82	0.91	0.86	0.86	0.81	0.83
	0.9	0.76	0.75	0.75	0.50	0.62	0.55	0.92	0.83	0.88	0.83	0.89	0.86	0.83	0.80	0.81
	0.85	0.71	0.78	0.74	0.47	0.66	0.55	0.93	0.81	0.86	0.83	0.88	0.85	0.81	0.79	0.80
	0.8	0.69	0.80	0.74	0.45	0.69	0.54	0.93	0.78	0.85	0.84	0.87	0.85	0.80	0.78	0.79
Sim_SF	0.95	0.56	0.69	0.62	0.74	0.60	0.66	0.95	0.86	0.90	0.92	0.78	0.84	0.92	0.83	0.87
	0.9	0.40	0.75	0.53	0.55	0.67	0.60	0.95	0.85	0.90	0.92	0.78	0.84	0.90	0.83	0.86
	0.85	0.32	0.78	0.46	0.41	0.70	0.52	0.95	0.84	0.89	0.92	0.77	0.84	0.88	0.82	0.85
	0.8	0.26	0.80	0.40	0.32	0.73	0.45	0.95	0.82	0.88	0.92	0.77	0.84	0.85	0.80	0.83
Sim_LG	0.95	0.90	0.85	0.87	0.87	0.82	0.84	0.98	0.96	0.97	0.98	0.88	0.93	0.96	0.93	0.94
	0.9	0.88	0.85	0.87	0.76	0.88	0.82	0.97	0.93	0.95	0.98	0.88	0.93	0.93	0.92	0.92
	0.85	0.83	0.89	0.86	0.73	0.89	0.80	0.97	0.92	0.94	0.98	0.88	0.93	0.92	0.91	0.91
	0.8	0.80	0.89	0.85	0.69	0.92	0.79	0.98	0.90	0.93	0.98	0.88	0.93	0.91	0.90	0.90
Sim_LF	0.95	0.98	0.86	0.91	0.80	0.84	0.82	0.97	0.92	0.94	0.99	0.97	0.98	0.94	0.91	0.92
	0.9	0.98	0.87	0.92	0.74	0.88	0.80	0.97	0.92	0.94	0.99	0.97	0.98	0.93	0.91	0.92
	0.85	0.97	0.89	0.93	0.69	0.90	0.78	0.97	0.89	0.93	0.99	0.97	0.98	0.91	0.90	0.91
	0.8	0.95	0.90	0.92	0.67	0.92	0.77	0.97	0.88	0.92	0.99	0.97	0.98	0.91	0.90	0.90

Table S2. Memory usage of the tested classifiers in GB. ViralV, PPR-M, and DeepVF represent classifiers viralVerify, PPR-Meta, and DeepVirFinder respectively. The binary classifier PLASMe was only run on the four simulated datasets, and it consumed an average of 0.3G of memory.

	4CAC	**DeepMC**	**viralV**	**PPR-M**	**geNomad**	**Tiara**	**PlasClass**	**Platon**	**DeepVF**	**VIBRANT**
Sim_SG	3.9	1.2	1.3	6.5	78.1	2.2	5.7	14.7	21.3	0.2
Sim_SF	15.5	1.1	0.7	6.4	66.4	1.6	16.9	14.8	10.2	0.2
Sim_LG	0.7	1.6	0.2	6.3	66.7	1.4	0.5	12.2	13.8	0.2
Sim_LF	0.6	1	0.3	6.3	68.2	1.4	0.4	12.1	7.9	0.2
Sharon	9.9	0.9	0.1	6.3	65.7	1.4	7.3	14	9.3	0.1
Tara	26.2	0.9	9.8	12.3	67.9	2.8	25.2	15.9	2.8	0.1
Oral_Nano	2.2	1.1	1.5	6.2	55.4	2.8	2.9	15	10	0.2
Gut_HiFi	2.7	1.1	2.4	9.3	68.9	3.7	3.3	13.2	9.4	0.3

2 Supplementary figures

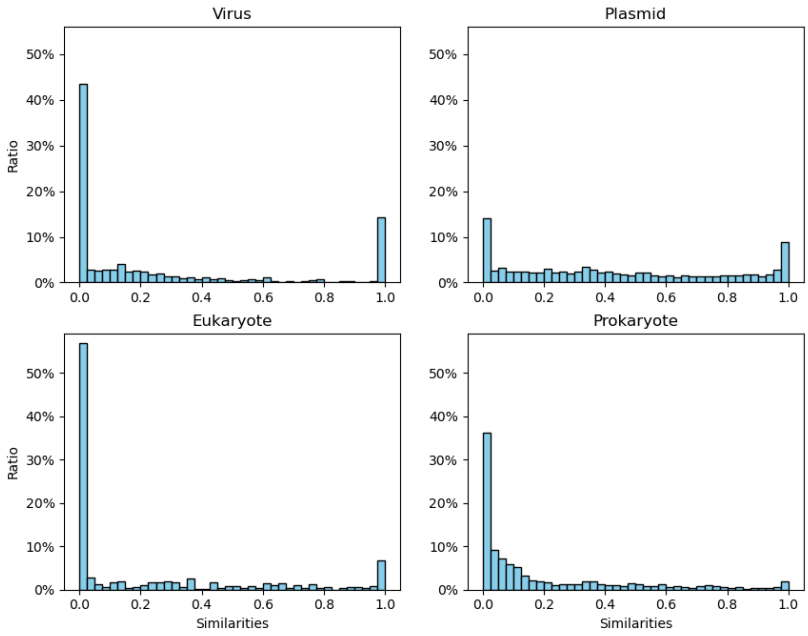

Fig. S1. The distribution of the similarity between the testing and training genomes. Genomes in the test set were matched to the training set of genomes using Minimap2, and the maximum similarity match of each genome is used as its similarity to the training set.

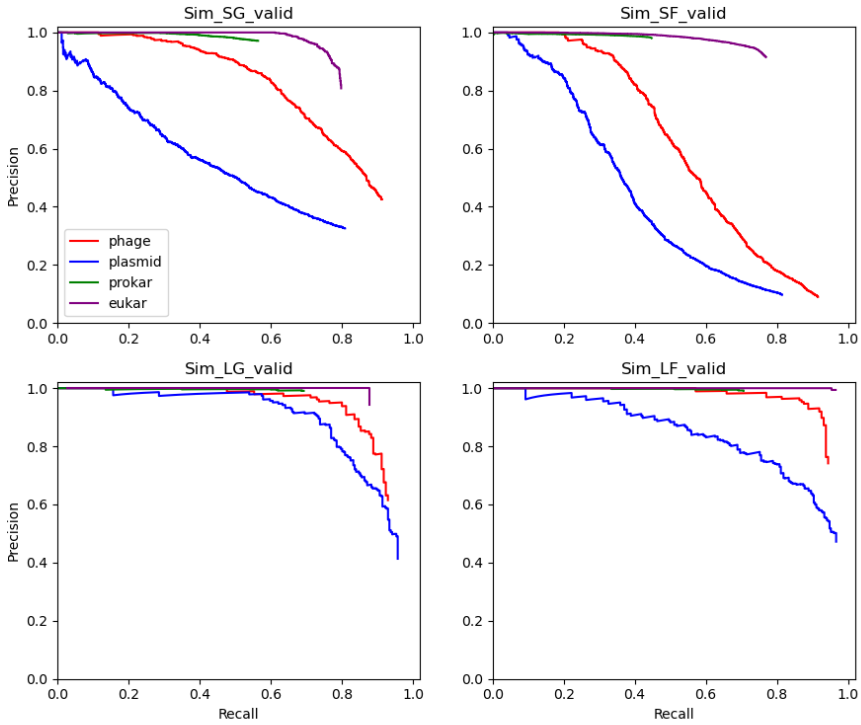

Fig. S2. The precision-recall curve of our XGBoost classifier on simulated metagenomes from the validation dataset. Lines colored red, blue, green, and purple represent precision-recall curves for phage, plasmid, prokaryote, and eukaryote classification, respectively. Validate_Sim_SG and Validate_Sim_SF are assembled from short reads while Validate_Sim_LG and Validate_Sim_LF are assembled from long reads.

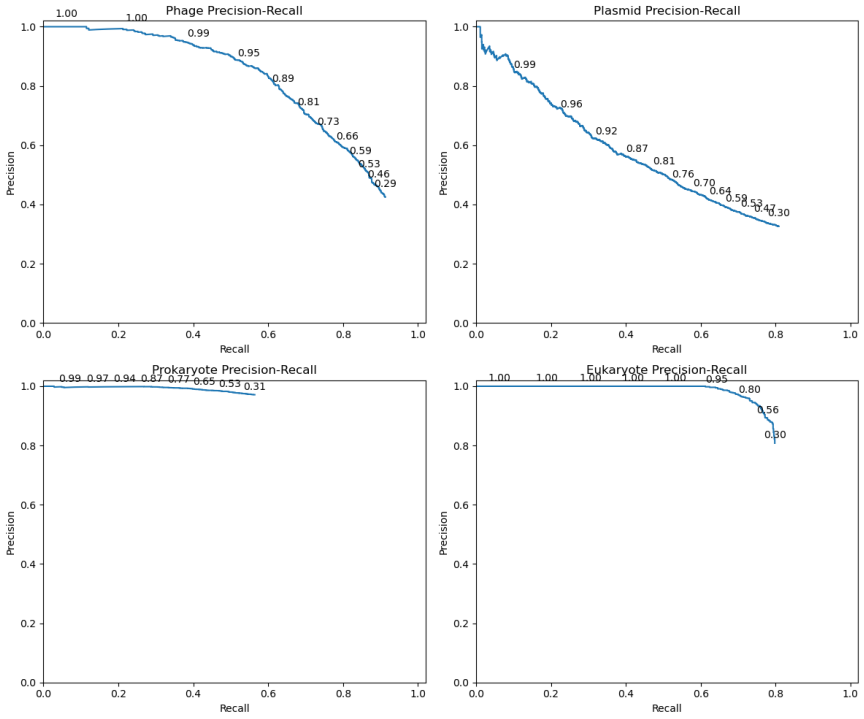

Fig. S3. The precision-recall curve of our XGBoost classifier on Validate_Sim_SG. The numbers on each curve indicate the corresponding score thresholds.

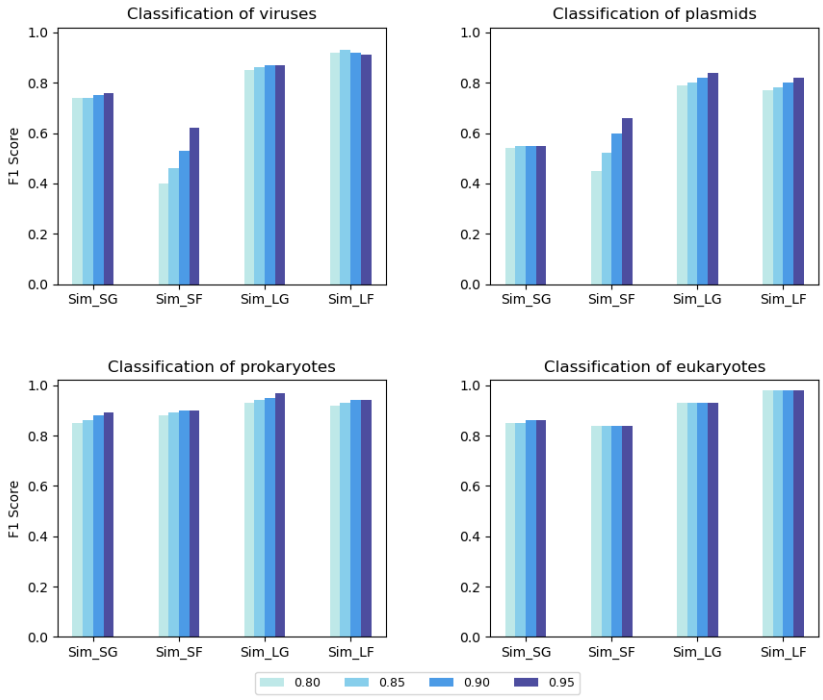

Fig. S4. Performance of 4CAC on validation datasets with initial classifications generated by different score thresholds. To generate an initial four-way classification by XGBoost classifier, we tested score thresholds 0.8, 0.85, 0.9, and 0.95 to classify contigs as viruses and plasmids. Note that this figure presents the results of the full 4CAC algorithm.

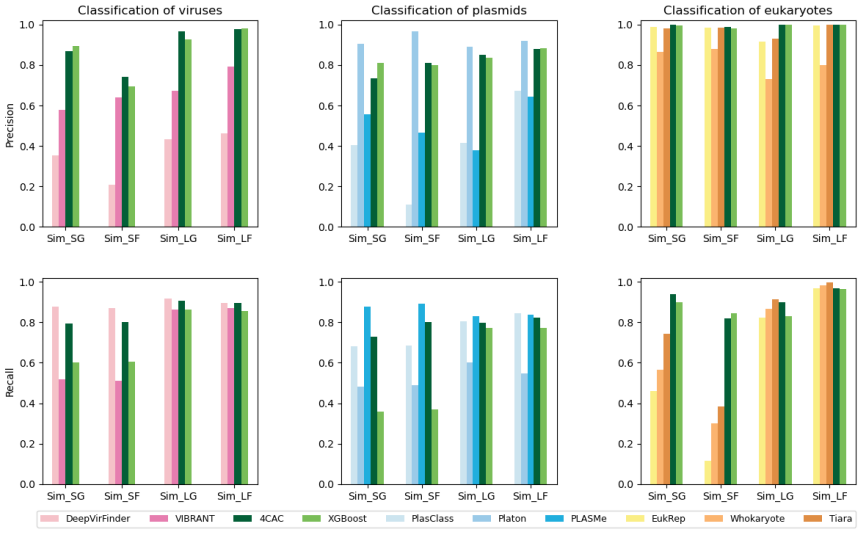

Fig. S5. Performance of binary classifiers and 4CAC on simulated metagenomes. Only two classes of contigs, those from the class and prokaryotes, were considered when benchmarking of each binary classifier. XGBoost represents the XGBoost classifier designed in this study without using graph information.

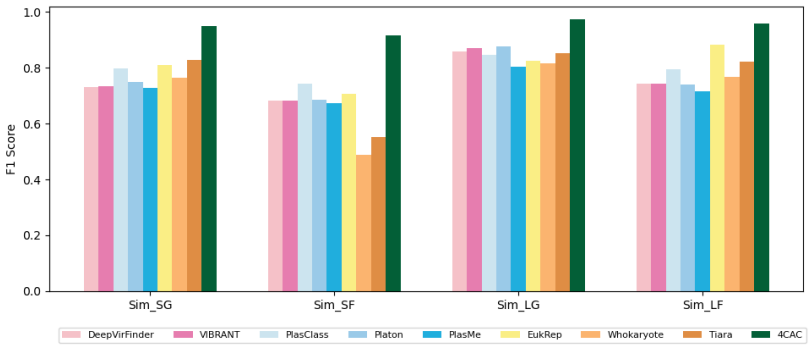

Fig. S6. Performance of binary classifiers and 4CAC on classifying prokaryotes from simulated metagenomes. XGBoost represents the XGBoost classifier designed in this study without using graph information.

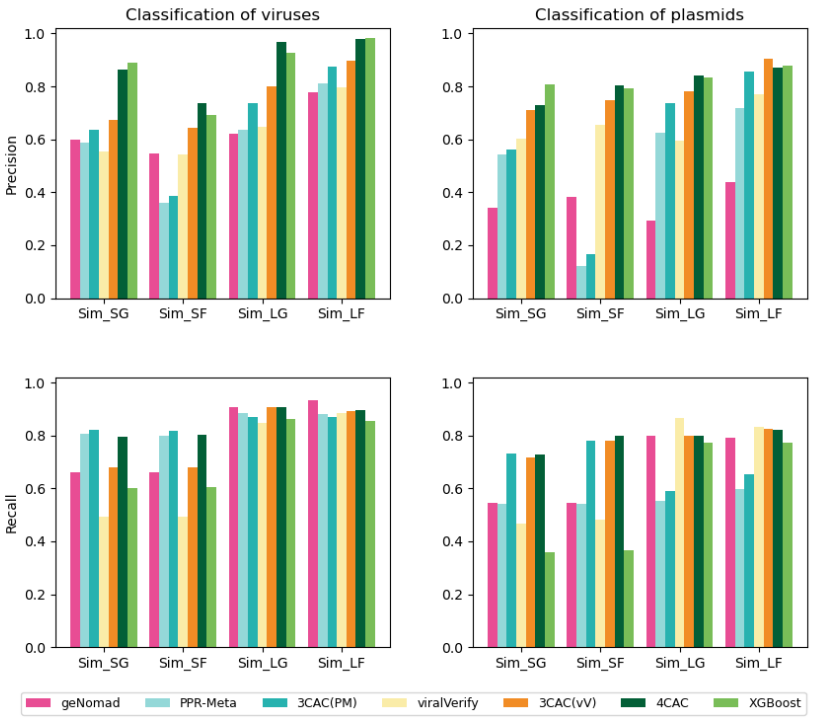

Fig. S7. Performance of three-way classifiers and 4CAC on simulated metagenomes. Eukaryotic contigs were excluded in the benchmarking of each three-way classifier. XGBoost represents the XGBoost classifier designed in this study without using graph information. 3CAC(vV) and 3CAC(PM) represent the execution of 3CAC using viralVerify and PPR-Meta, respectively.

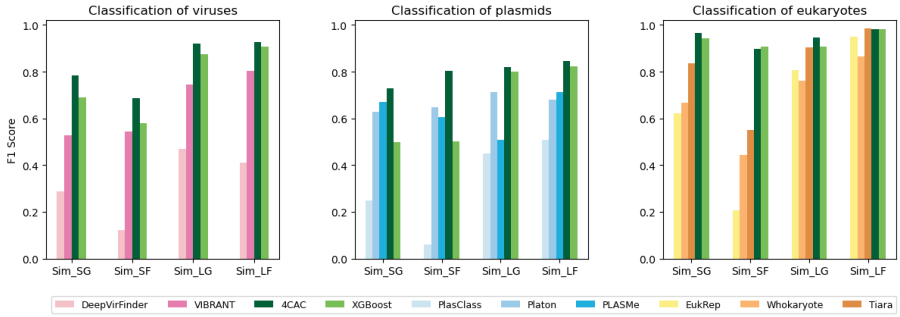

Fig. S8. Performance of binary classifiers and 4CAC on simulated metagenomes. All four classes of contigs were included in the benchmark.

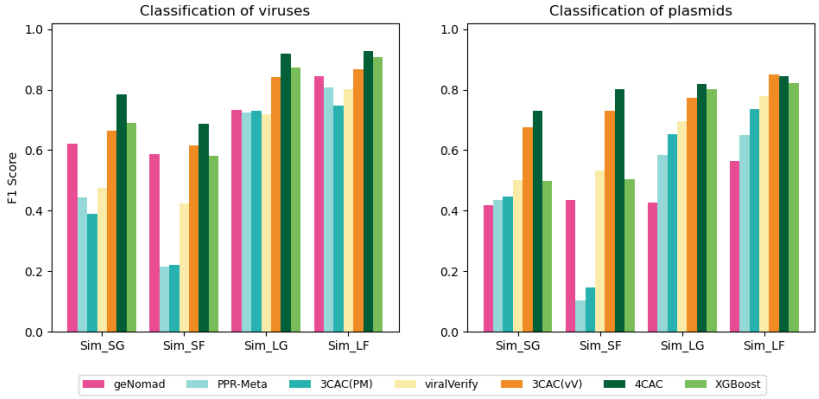

Fig. S9. Performance of three-way classifiers and 4CAC on simulated metagenomes. All four classes of contigs were included in the benchmark.

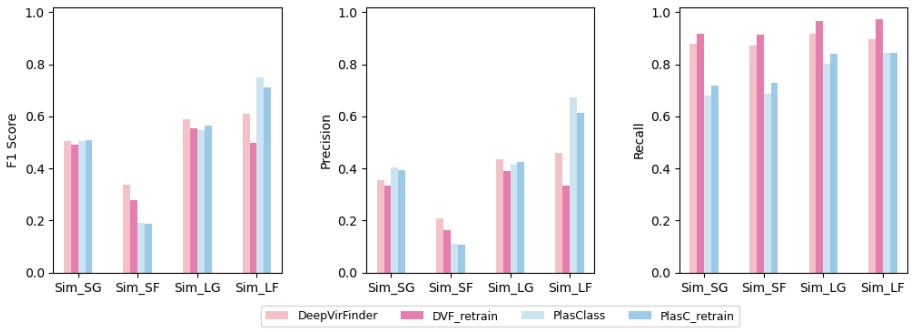

Fig. S10. Performance of DeepVirFinder and PlasClass retrained on the same training dataset as 4CAC. DVF_retrain and PlasC_retrain represent DeepVirFinder and PlasClass, respectively, retrained on the same training dataset as 4CAC. DVF and PlasClass represent the algorithms with their default training.

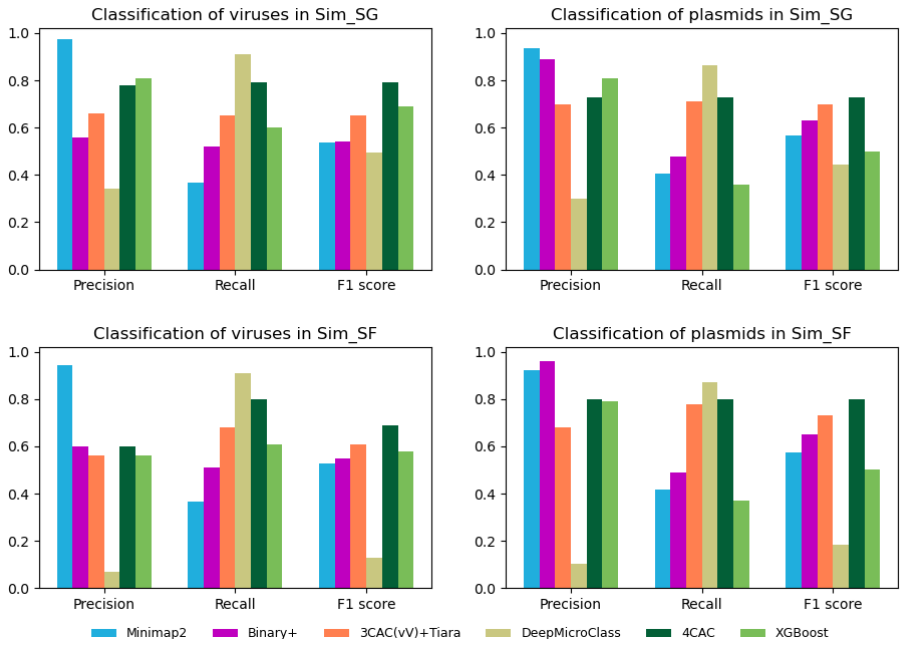

Fig. S11. Performance of four-class classifiers on classifying viruses and plasmids from assemblies of simulated short-read datasets.

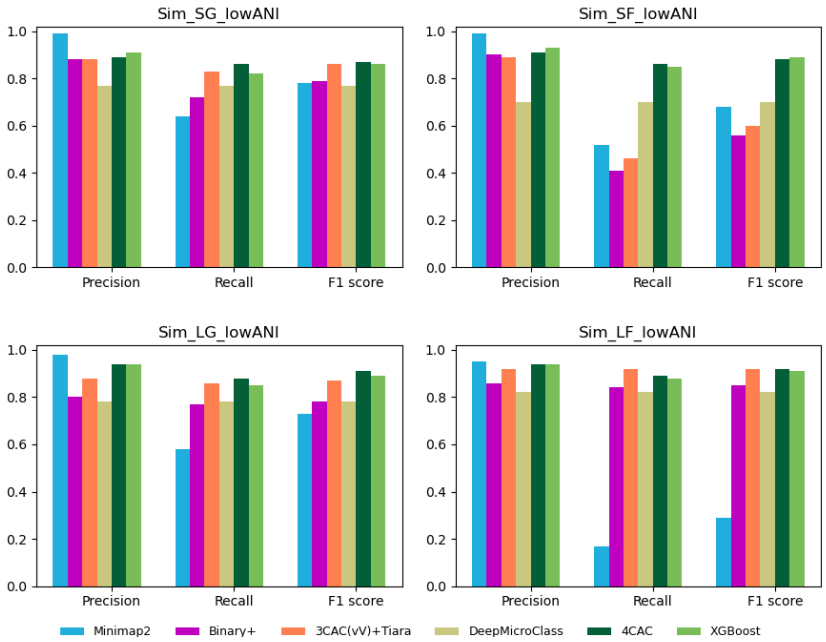

Fig. S12. Performance of four-class classifiers on simulated metagenomes with low similarities to the training dataset. Four simulated datasets Sim_*_lowANI were created using only genomes in the test dataset with maximum similarity <85% to the training dataset. See Methods for details on generating the simulated datasets.