

Supporting Information

Ligand prediction for orphan targets using support vector machines and various target-ligand kernels is dominated by nearest neighbor effects

Anne Mai Wassermann, Hanna Geppert, and Jürgen Bajorath*

Supplementary figure legends

Figure S1. Ligand prediction for target set 1.

For the MACCS and TGD fingerprints, recovery rates for SVM TLK and LC strategies are shown for 11 alternative target kernels and selection sets of 100 database compounds. Recovery rates are averaged over all 12 targets and 10 independent search trials per target.

Figure S2. Ligand prediction for target set 2.

Search results for target set 2 are shown corresponding to Figure S1.

Figure S3. Target-dependent search performance (set 1).

For all targets of set 1, recovery rates are shown for simple SVM ranking and for the SVM TLK search strategy in combination with the MEROPS kernel. Recovery rates are compared for selection sets of 100 compounds averaged over 10 independent trials per target.

Figure S4. Target-dependent search performance (set 2).

For all targets of set 2, recovery rates are shown for simple SVM ranking and for the SVM TLK search strategy in combination with the MEROPS kernel. For simple SVM ranking, ligands of all members of the orphanized target's subfamily were pooled and used as the positive training class. Recovery rates are shown for a selection set of 100 compounds averaged over 10 independent trials per target.

Figure S5. Dependence of search performance on ligands of nearest neighbor targets.

For all set 1 targets, search results are reported for the SVM TLK (MEROPS) search strategy. Blue bars show recovery rates obtained by learning with all target set ligands, whereas red bars show recovery rates obtained when the ligands of the nearest neighbor are excluded from the training set. Recovery rates are shown for a selection set of 100 compounds averaged over 10 independent trials per target.

Table S1. Search results for ligand prediction using SVM LC (set 1).

	uniform	needle	water	PROFEAT	spectrum	SSEA	GO	cleavage	SCOP	Topmatch	MEROPS
MACCS											
ace2	3.5	3.9	3.0	3.5	4.8	3.5	2.2	1.7	0.4	4.4	1.3
cal2	7.3	5.2	7.1	7.5	8.0	7.1	14.1	14.6	22.7	9.8	22.7
cas3	7.1	6.6	7.5	7.0	6.0	7.5	6.3	6.9	7.1	5.6	5.6
catD	28.8	28.2	34.3	29.7	46.6	29.9	36.5	35.7	27.9	37.9	27.5
catL	9.0	8.9	8.9	9.0	8.6	9.2	11.9	14.5	11.8	10.0	12.9
mgcp	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
metap2	1.2	1.2	1.3	1.3	1.2	1.3	1.7	1.4	1.0	1.4	0.6
mmp2	2.6	13.1	5.6	3.0	9.0	3.6	7.1	7.2	27.7	29.6	39.4
mmp8	4.6	25.5	8.2	4.6	12.7	4.6	13.6	11.8	35.5	41.8	44.6
ren	4.1	18.3	7.9	4.5	14.3	7.9	12.3	16.6	29.4	17.6	29.9
thr	11.2	28.2	16.3	12.1	11.4	16.4	28.4	21.4	29.8	28.7	29.8
try	15.6	37.1	21.5	15.4	14.0	22.7	20.2	29.6	43.9	41.0	43.7
average	7.9	14.7	10.1	8.1	11.4	9.5	12.9	13.5	19.8	19.0	21.5
TGD											
ace2	15.2	16.5	15.2	15.2	16.5	15.7	15.2	10.9	10.9	13.5	13.5
cal2	13.6	8.4	12.7	11.8	10.0	12.3	17.7	17.7	10.9	11.8	16.1
cas3	20.1	16.1	20.5	20.0	21.8	19.5	22.2	23.3	23.1	22.2	21.0
catD	34.6	38.0	39.2	34.5	42.6	38.2	42.8	45.5	44.3	41.4	44.0
catL	6.6	6.6	6.6	6.6	6.7	6.7	6.4	7.3	7.8	6.6	7.4
mgcp	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
metap2	7.5	5.3	7.5	7.8	8.0	7.6	7.5	2.5	9.8	4.1	10.7
mmp2	4.4	23.5	7.1	4.9	11.2	5.5	8.6	6.3	39.1	37.2	53.0
mmp8	11.8	28.2	18.2	11.8	22.7	12.7	16.4	20.0	52.7	54.6	53.6
ren	21.8	45.0	31.8	23.8	39.8	29.6	36.7	40.5	46.4	45.7	45.9
thr	19.5	24.9	21.4	19.8	19.9	21.7	27.0	25.9	27.7	26.6	28.0
try	19.4	35.2	25.6	20.8	22.3	25.2	24.8	35.4	44.8	41.5	45.2
average	14.5	20.6	17.1	14.7	18.5	16.2	18.8	19.6	26.5	25.4	28.2

Recovery rates (in %) for all targets in data set 1 are reported for selection sets of 100 compounds averaged over 10 independent trials per target.

Table S2. Search Results for ligand prediction using SVM TLK (set 1).

	uniform	needle	water	PROFEAT	spectrum	SSEA	GO	cleavage	SCOP	Topmatch	MEROPS
MACCS											
ace2	5.7	5.2	4.4	3.9	5.2	3.5	0.0	0.9	0.4	4.4	1.3
cal2	2.3	5.7	6.6	2.3	5.2	2.5	19.1	27.5	26.1	13.6	25.2
cas3	6.7	6.7	8.9	8.1	6.6	8.7	7.4	8.3	8.3	6.3	6.2
catD	21.2	22.2	34.6	27.5	46.0	31.7	37.2	39.5	24.5	24.3	24.8
catL	8.5	9.3	11.0	8.6	8.6	8.8	11.6	17.7	10.8	10.4	13.6
mgcp	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
metap2	1.7	1.2	1.7	1.6	1.4	1.6	1.6	1.5	1.1	1.7	0.7
mmp2	3.3	17.4	16.2	4.0	15.4	8.7	14.9	20.3	34.2	39.6	41.9
mmp8	9.1	30.9	28.2	10.0	20.9	17.3	23.6	27.3	42.7	48.2	46.4
ren	3.0	18.7	9.5	5.0	14.1	9.7	30.5	12.2	26.0	17.9	25.6
thr	17.5	29.2	27.5	18.9	13.4	27.2	13.7	27.5	29.1	29.4	28.9
try	19.6	37.9	30.4	20.4	14.4	34.8	28.3	35.2	41.7	38.5	41.7
average	8.2	15.4	14.9	9.2	12.6	12.9	15.7	18.2	20.4	19.5	21.4
TGD											
ace2	10.4	16.5	11.3	10.9	14.8	13.0	0.4	7.8	6.1	10.9	10.4
cal2	9.3	6.6	9.3	6.4	7.7	6.1	11.6	12.7	11.1	10.7	15.5
cas3	14.9	14.4	17.4	16.0	20.4	15.0	19.0	21.2	21.5	21.6	20.8
catD	36.0	38.2	41.2	36.3	42.5	39.9	43.7	42.9	43.4	40.5	42.3
catL	6.3	6.2	6.3	6.3	6.3	6.2	7.0	7.5	7.4	6.2	7.3
mgcp	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
metap2	11.5	7.6	11.4	11.2	9.7	11.7	5.1	2.7	10.4	5.5	11.7
mmp2	4.9	33.0	15.4	5.5	16.8	8.7	10.9	18.6	49.6	57.3	60.5
mmp8	10.0	37.3	26.4	12.7	28.2	15.5	22.7	34.6	54.6	57.3	57.3
ren	17.5	43.5	33.8	22.9	39.6	30.6	43.0	40.4	44.7	43.5	44.2
thr	19.5	25.1	22.7	20.3	19.3	23.5	25.9	27.5	28.2	27.0	28.2
try	27.3	38.1	32.9	27.5	24.6	35.2	29.6	43.9	45.6	46.0	46.4
average	14.0	22.2	19.0	14.7	19.2	17.1	18.2	21.7	26.9	27.2	28.7

Recovery rates (in %) for all targets in data set 1 are reported for selection sets of 100 compounds averaged over 10 independent trials per target.

Table S3. Search results for ligand prediction with SVM LC (set 2).

	uniform	needle	water	PROFEAT	spectrum	SSEA	GO	cleavage	SCOP	Topmatch	MEROPS
MACCS											
cal1	45.7	48.5	51.2	48.6	49.3	50.0	50.0	50.7	50.0	52.0	50.2
cal2	42.5	67.5	55.7	48.0	63.2	52.5	65.0	65.7	78.6	69.8	74.3
cas1	8.8	6.3	10.0	9.4	12.5	11.3	19.4	20.0	21.3	20.6	21.3
cas3	7.5	16.3	10.2	7.8	14.3	10.3	10.3	17.8	22.7	17.9	22.7
catB	40.0	36.7	40.8	40.8	40.8	39.2	33.3	41.7	35.0	38.3	29.2
catK	17.8	19.2	18.3	18.0	19.3	18.4	19.8	18.7	18.0	18.9	18.1
catL	23.8	25.6	23.8	23.6	26.3	24.5	21.8	21.9	23.0	24.3	24.0
catS	13.2	23.0	12.8	13.9	22.3	15.7	16.9	14.1	21.1	19.4	22.6
faXa	2.6	8.2	4.5	2.9	7.2	4.6	4.6	6.2	10.2	9.6	10.2
thr	7.0	23.2	11.7	7.5	14.2	11.7	27.0	13.8	28.5	26.1	28.5
try	13.8	44.0	23.6	14.9	22.6	24.7	22.1	34.0	48.5	44.0	48.5
average	20.2	29.0	23.9	21.4	26.5	23.9	26.4	27.7	32.4	31.0	31.8
TGD											
cal1	22.4	31.2	26.8	24.4	31.2	25.9	28.1	28.1	30.5	30.2	30.7
cal2	19.6	46.8	20.9	20.5	50.0	23.2	26.4	21.8	39.8	24.8	50.2
cas1	19.4	21.9	22.5	20.6	30.6	23.1	26.9	36.9	38.8	36.9	38.8
cas3	19.0	29.3	22.8	19.6	27.1	22.3	24.9	33.6	36.0	32.9	36.0
catB	34.2	37.5	32.5	33.3	35.8	33.3	39.2	34.2	35.8	36.7	35.8
catK	10.1	12.3	10.9	10.4	11.8	10.4	11.6	12.6	13.8	12.6	12.2
catL	10.0	11.0	10.0	10.0	11.4	10.0	9.9	10.6	11.2	11.0	10.7
catS	3.0	3.0	3.1	2.9	3.1	2.8	3.4	2.6	3.3	3.0	3.0
faXa	6.6	7.5	7.0	6.7	7.4	7.0	7.4	7.3	7.9	7.7	7.9
thr	28.2	30.5	29.8	28.5	29.9	29.8	31.9	30.3	31.3	31.0	31.3
try	41.7	48.7	46.0	42.5	45.5	46.2	44.9	48.1	50.9	50.0	50.9
average	19.5	25.4	21.1	19.9	25.8	21.3	23.1	24.2	27.2	25.2	28.0

Recovery rates (in %) for all targets in data set 2 are reported for selection sets of 100 compounds averaged over 10 independent trials per target.

Table S4. Search results for ligand prediction using SVM TLK (set 2).

	uniform	needle	water	PROFEAT	spectrum	SSEA	GO	cleavage	SCOP	Topmatch	MEROPS
MACCS											
cal1	38.1	49.5	52.0	50.0	50.5	51.7	49.5	53.2	52.2	53.2	51.0
cal2	32.7	65.7	55.5	44.1	63.0	50.9	52.7	58.9	65.7	63.0	67.1
cas1	6.9	10.0	12.5	10.6	16.3	15.6	19.4	20.0	16.9	20.0	18.8
cas3	6.6	19.5	12.3	7.4	15.6	12.4	9.1	21.4	24.4	21.2	24.0
catB	37.5	35.8	40.0	35.8	39.2	34.2	30.0	39.2	25.0	29.2	23.3
catK	13.7	18.5	16.7	16.3	19.4	16.1	17.8	17.3	16.6	17.3	17.7
catL	22.7	24.4	24.1	23.6	26.0	25.9	23.2	24.3	24.0	24.8	23.4
catS	11.8	24.1	11.0	14.0	23.6	17.6	16.3	14.5	21.9	21.3	23.9
faXa	3.6	9.7	7.7	4.7	8.4	7.9	8.2	9.2	10.4	10.2	10.5
thr	7.3	25.9	19.9	8.8	15.6	19.2	24.7	21.0	28.3	25.9	29.0
try	16.0	44.3	33.0	19.6	23.8	35.1	29.8	41.9	45.1	40.6	47.4
average	17.9	29.8	25.9	21.4	27.4	26.0	25.5	29.2	30.0	29.7	30.5
TGD											
cal1	21.0	30.7	27.3	22.7	31.0	25.1	28.1	27.3	31.2	30.7	29.8
cal2	14.3	44.3	19.8	16.4	47.7	20.7	20.5	20.9	40.0	26.1	48.4
cas1	16.3	20.0	19.4	16.3	29.4	21.3	24.4	33.8	31.9	35.0	33.8
cas3	18.0	30.1	24.4	19.0	27.3	24.3	22.7	34.9	35.9	33.5	35.9
catB	35.0	35.8	34.2	35.0	33.3	33.3	37.5	34.2	29.2	35.0	31.7
catK	8.3	11.8	9.9	8.8	11.6	9.0	9.4	10.8	11.5	11.3	11.7
catL	9.3	10.3	8.9	9.5	10.8	9.5	9.7	10.3	10.4	10.6	9.9
catS	3.9	3.7	3.8	3.7	3.7	3.4	3.6	3.2	3.6	3.4	3.7
faXa	7.4	7.7	7.6	7.3	7.6	7.7	7.8	7.7	8.1	7.9	7.9
thr	28.0	30.4	30.0	28.6	29.8	29.8	31.6	30.4	31.1	30.7	31.2
try	47.4	50.9	50.8	47.9	47.2	50.6	50.9	53.0	52.8	50.9	52.8
average	19.0	25.1	21.5	19.5	25.4	21.3	22.4	24.2	26.0	25.0	27.0

Recovery rates (in %) for all targets in data set 2 are reported for selection sets of 100 compounds averaged over 10 independent trials per target.

Table S5. Search results for ligand prediction using simple SVM (set 2).

reference targets	MACCS				TGD			
	nearest neighbor target		all subfamily members		nearest neighbor target		all subfamily members	
set size	100	1000	100	1000	100	1000	100	1000
cal1	46.8	62.2	identical	identical	29.8	40.2	identical	identical
cal2	61.8	76.8	identical	identical	49.3	65.5	identical	identical
cas1	21.3	51.9	identical	identical	36.9	69.4	identical	identical
cas3	21.6	41.2	identical	identical	36.0	54.2	identical	identical
catB	10.0	39.2	13.3	63.3	13.3	35.8	21.7	49.2
catK	14.0	42.9	16.7	47.7	8.0	22.4	8.9	33.2
catL	14.5	36.4	22.2	64.3	7.0	26.3	9.5	26.6
catS	14.0	41.5	24.8	60.7	3.5	18.0	4.3	32.6
faXa	8.7	31.7	10.5	39.0	7.6	48.7	8.1	53.0
thr	9.6	33.3	28.1	71.6	23.6	70.9	31.2	82.0
try	12.5	29.1	44.7	67.6	34.0	72.3	51.9	87.7
average	21.3	44.2	28.3	58.8	22.6	47.6	26.1	54.0

Recovery rates (in %) are reported for all targets in data set 2 averaged over 10 independent trials per target. Searches are either carried out using only the ligands of the nearest neighbor target as positive training class or using the pooled ligands of all members of the orphanized target's subfamily as positive training class.

Table S6. Search results for ligand prediction using SVM TLK and LC in the absence of nearest neighbor information.

	MACCS				TGD			
strategy	LC		TLK		LC		TLK	
set size	100	1000	100	1000	100	1000	100	1000
ace2	4.4	17.8	4.4	15.7	14.4	36.5	13.5	40.0
cal2	3.2	49.6	2.7	41.8	13.0	32.5	10.7	30.0
cas3	7.0	35.8	7.5	36.3	21.3	48.5	20.8	48.1
catD	15.9	54.3	20.6	60.8	28.6	65.5	27.5	64.2
catL	7.5	26.3	8.5	25.8	5.8	12.6	5.8	11.5
mgcp	0.0	1.1	0.0	1.1	0.0	0.0	0.0	0.0
metap2	0.8	7.3	0.9	6.1	11.1	38.8	12.2	40.8
mmp2	0.4	5.6	0.3	4.7	4.6	23.1	4.7	23.9
mmp8	0.0	6.4	0.0	6.4	5.5	24.6	6.4	23.6
ren	1.3	14.2	1.2	16.9	4.9	44.2	4.5	44.5
thr	4.5	29.7	4.1	29.2	14.5	43.3	13.8	41.1
try	5.0	25.4	4.6	24.4	9.8	44.8	10.8	49.2
average	4.2	22.8	4.6	22.4	11.1	34.5	10.9	34.7

Recovery rates (in %) are reported for all targets in data set 1 averaged over 10 independent search trials per target. For all orphan targets, the nearest neighbor target and its ligands were excluded from training. All searches were carried out with the MEROPS kernel.

Figure S1.

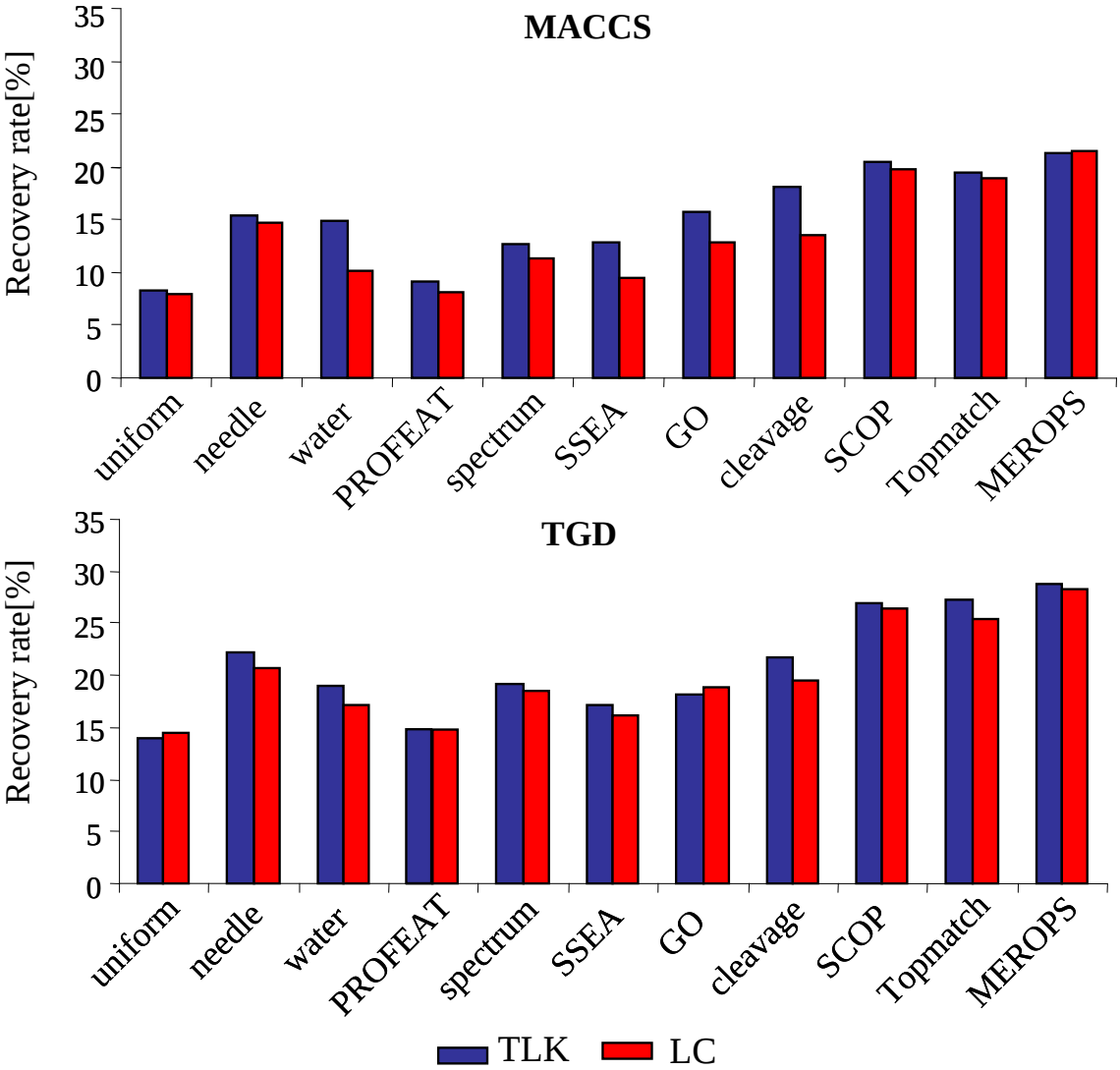


Figure S2.

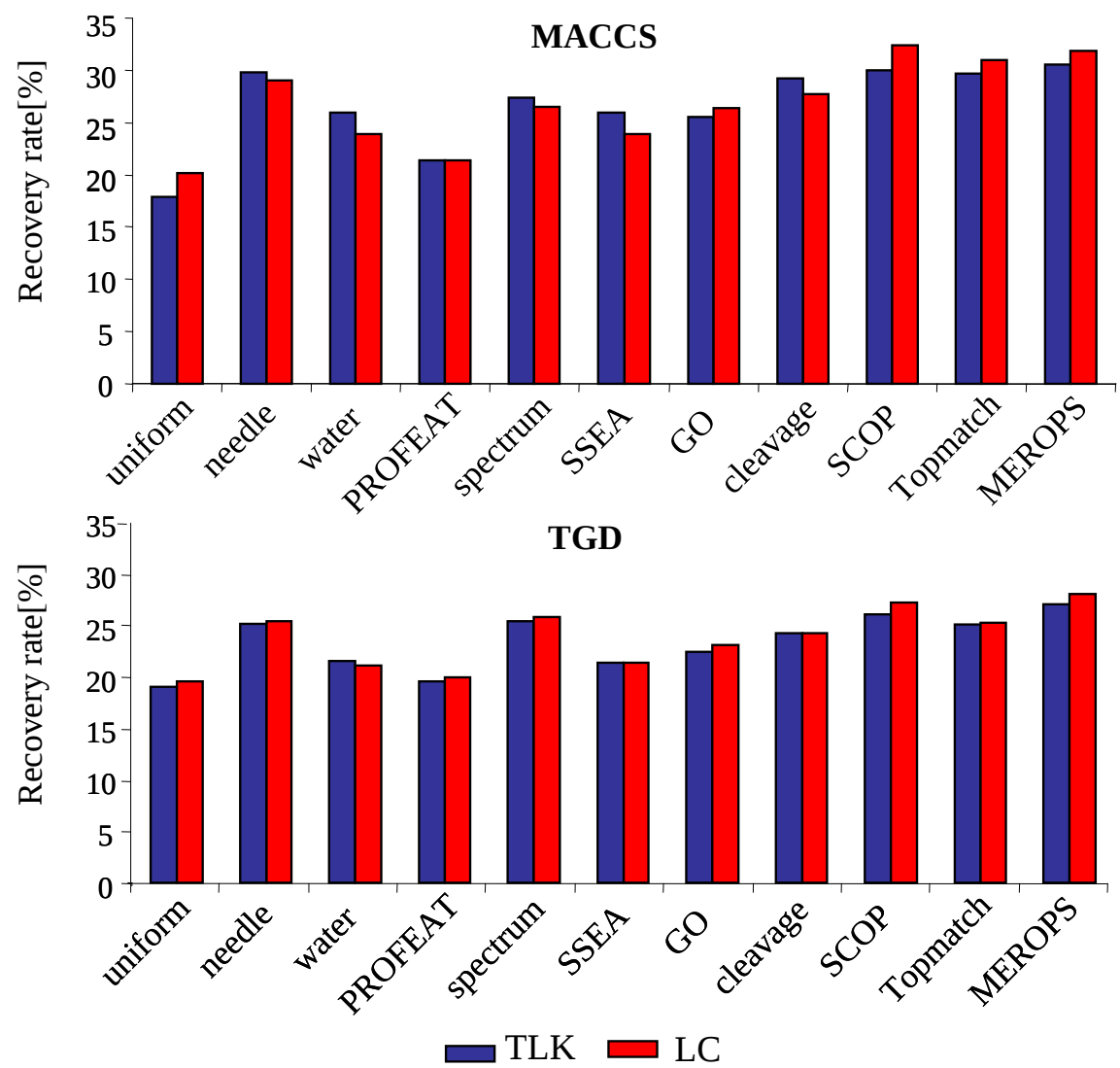


Figure S3.

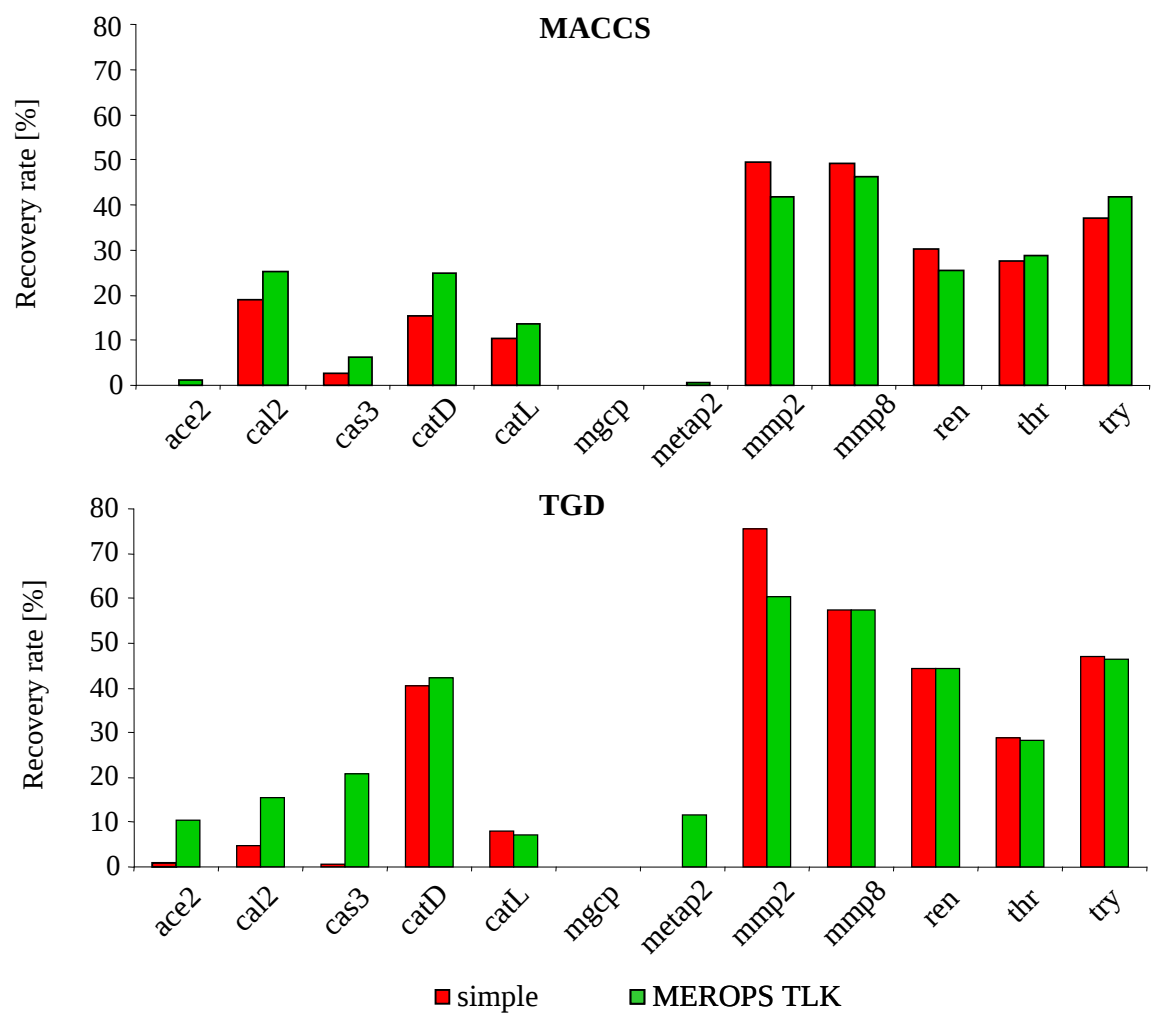


Figure S4.

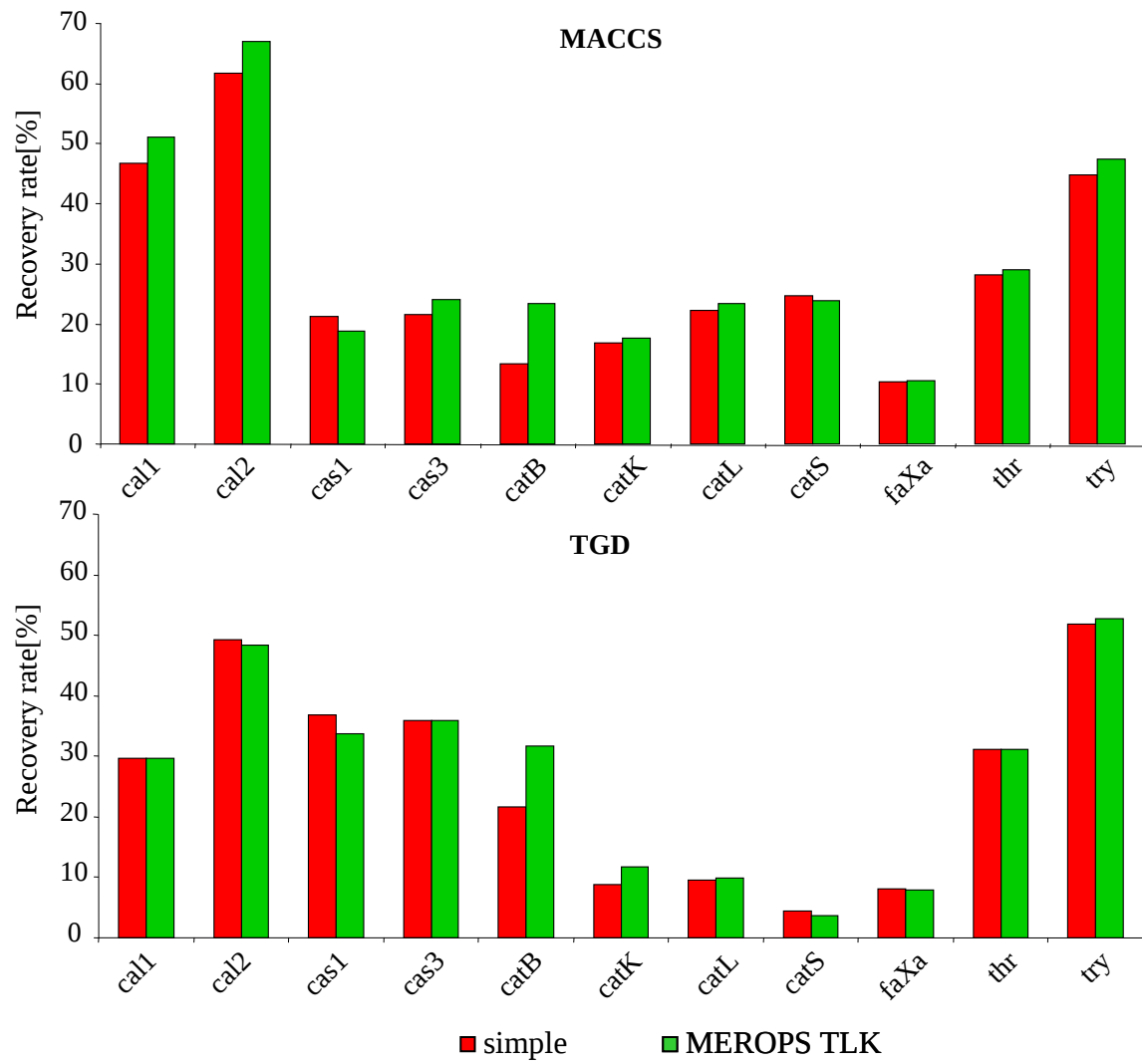


Figure S5.

