BEFORE DOING EXPERIMENTS

Qi Chen
Qi.Chen@ecs.vuw.ac.nz
Bing Xue
Bing.Xue@ecs.vuw.ac.nz
Harith Al-Sahaf
Harith.Al-Sahaf@ecs.vuw.ac.nz

Outline

- Experiment Settings
- Parameter Settings
- Results
- Codes and Programs
- ~~Grid Computing~~

Experiment Settings

- Baseline and Benchmark methods: select appropriate models for comparison
  - Task relevant: feature learning, classification, regression, clustering, …
  - Baseline methods: simple, well-established methods, provide a reference point
  - Original method: to demonstrate the improvement over the predecessor
  - State-of-the-art methods: currently achieve the best performance on a specific task, also consider
    - ✓time of publication: more likely to be more recent publications, a general rule of thumb, 2-3 years old can be considered "old"
    - ✓publication venue, top-tier journal and conference

Experiment Settings

- Benchmark Datasets: use datasets commonly used for benchmarking, number of datasets: generally more than 10
- Training and Test Sets
  - Which method to create training and test sets
    - Hold-out/training-test split: with a splitting ratio of 80:20 or 70:30
    - K-fold cross-validation (sampling without replacement)
    - Bootstrap (sampling with replacement)
  - Splitting seed
  - Stratified splitting to maintain class-ratio

Experiment Settings (continue)

- Number of runs:
  - at least **30** runs (using 30 different seeds) => WHY?
  - or 50 runs unless your experiment is too time-consuming.
- Performance evaluation:
  - Measure: error rate, accuracy.
  - What to report if you have (10-fold cross-validation x 30 runs) results?

Random Seeds For EC methods

- DO record your random seeds to make sure you can re-produce the same results later if needed
  - Do NOT use clock time as the random seed

- Use the same random seeds to compare two different versions of the same approach:
  - E.g. two different GP algorithms: GP1 and GP2, run both of them for 50 times. Please make sure you use the same 50 random seeds for GP1 and GP2 to let them have the same starting points for fair comparisons.
  - It is not necessary to use the same random seeds if you compare GP with PSO.

Parameter Settings

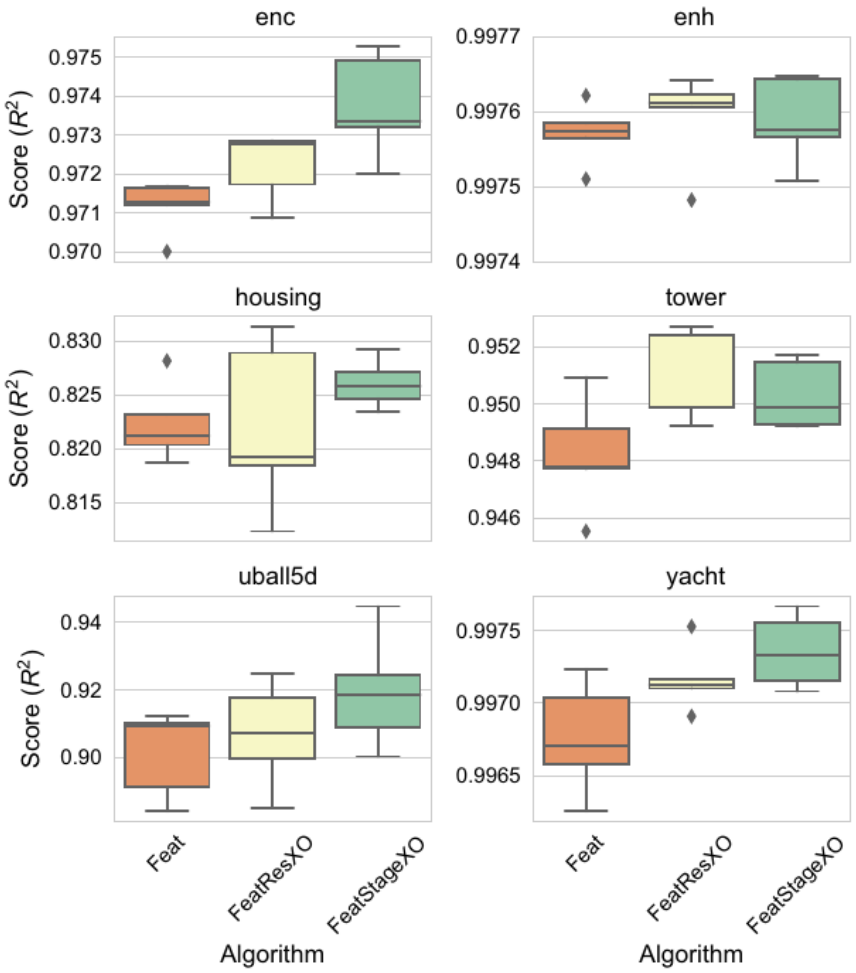
- Start with commonly used parameter settings from the literature, or settings recommended by good papers.
  - Do not randomly pick up some parameters values unless you have good reasons to use them.
- Figure out what the parameter values exactly mean
- Parameter tuning: one aspect at a time

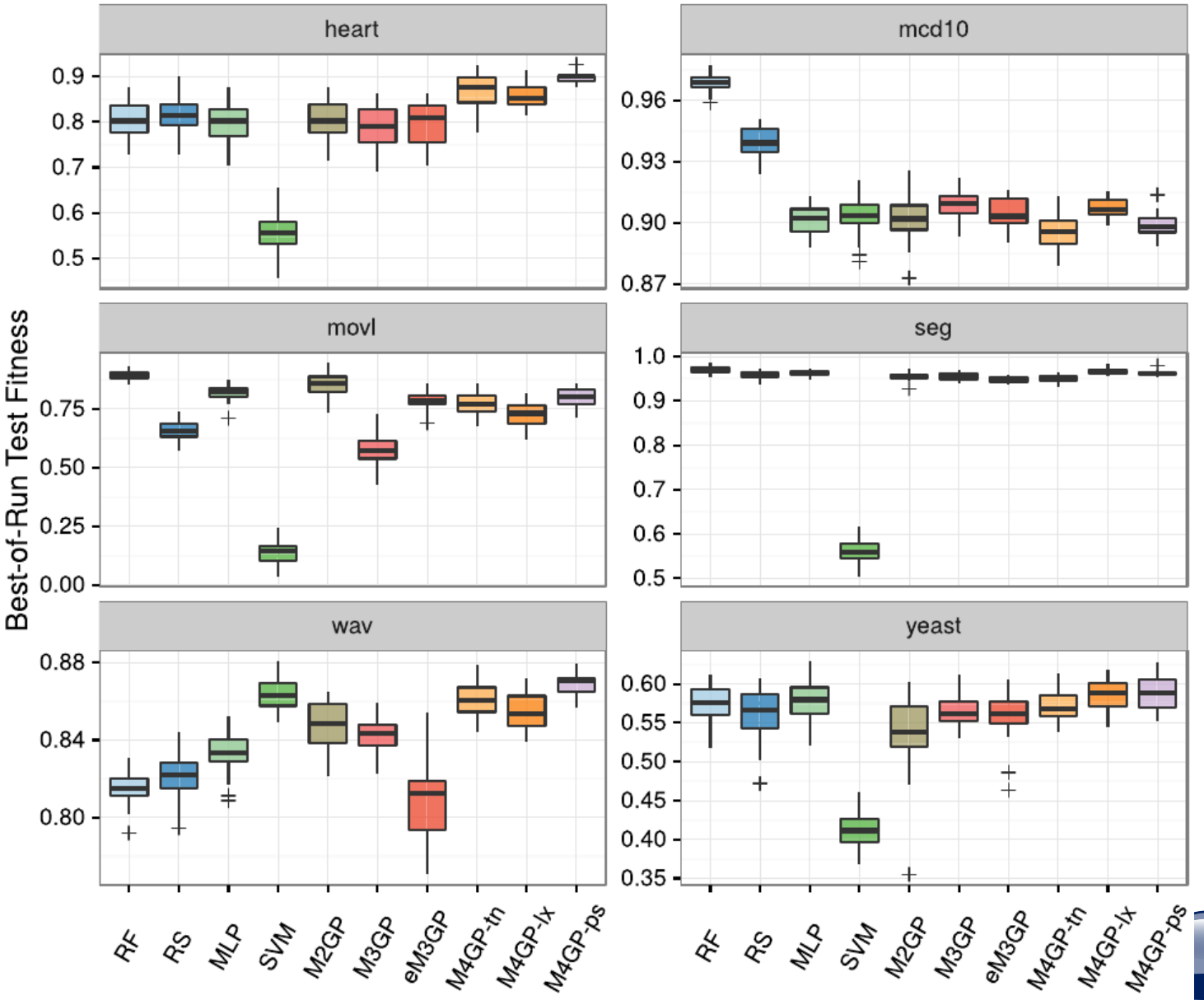
Results

- Please record all useful results
  - Eg, the *gbest* in PSO, the best program from GP, the training, testing performances in each run (you may further check with your supervisors)
  - Evolutionary plots on the training and test sets: performance of the best individuals at every generation
  - Computational (training) time of each run: first generation to the last generation — not include test process
  - Model Size, #feature ..
  - Evolved Models

Table 3
Comparison of best-of-run median test accuracy for the benchmark problems. The best result is highlighted in bold. Significant ($p < 0.01$ according to a pairwise Wilcoxon rank-sum test with Holm correction) improvements with respect to each method is denoted by $a - j$ according to the method labels.

Method	Heart	IM-3	IM-10	Movl	Seg	Vowel	Wav	Yeast
[a]RF	[d]80.2	94.8	[bcde fghij]**96.9**	[bcde fghij]**89.4**	[bcde fgij]**97.3**	[bcde g]89.4	81.5	[de]57.5
[b]RS	[d]81.5	92.8	[cde fghij]93.9	[d f]65.7	[d gj]96.0	[g]82.8	82.2	[d]56.6
[c]MLP	[d]80.2	95.9	90.2	[bd fghj]82.5	[de gj]96.3	[g]82.5	[ab g]83.3	[de]58.0
[d]SVM	55.6	93.8	90.4	14.4	55.8	81.8	[abce fgh]86.3	41.1
[e]M2GP	[d]80.2	93.8	90.2	[bcd fghij]85.9	[d g]95.6	[cd g]85.9	[abc g]84.9	[d]53.8
[f]M3GP	[d]79.0	95.4	[c ij]91.0	[d]57.1	[d]95.6	[abcde g]93.8	[ab g]84.3	[d]56.2
[g]eM3GP	[d]80.9	93.3	[j]90.3	[bd fh]78.6	[d]94.7	78.6	81.2	[d]56.2
[h]M4GP-lx	[abcde fg]85.2	[abcde fg]**97.9**	[ij]90.7	[bd f]73.1	[bde fgj]96.6	[abcde fg]95.6	[abc fg]85.3	[de g]**58.9**
[i]M4GP-ps	[abcde fghj]**90.1**	[abcde fg]97.9	89.8	[bd fh]80.1	[de gj]96.1	[abcde fghj]**97.5**	[abce fghj]**87.1**	[de fg]58.9
[j]M4GP-tn	[abcde fg]87.7	[abcde fg]97.9	89.6	[bd fh]76.9	[d]95.1	[abcde fg]96.0	[abce fg]86.0	[d]56.8


enc
enh
housing
tower
uball5d
yacht
Score (R^2)
Feat
FeatResXO
FeatStageXO
Algorithm


heart
mcd10
movl
seg
wav
yeast
Best-of-Run Test Fitness
RF
RS
MLP
SVM
M2GP
M3GP
eM3GP
M4GP-tn
M4GP-lx
M4GP-ps

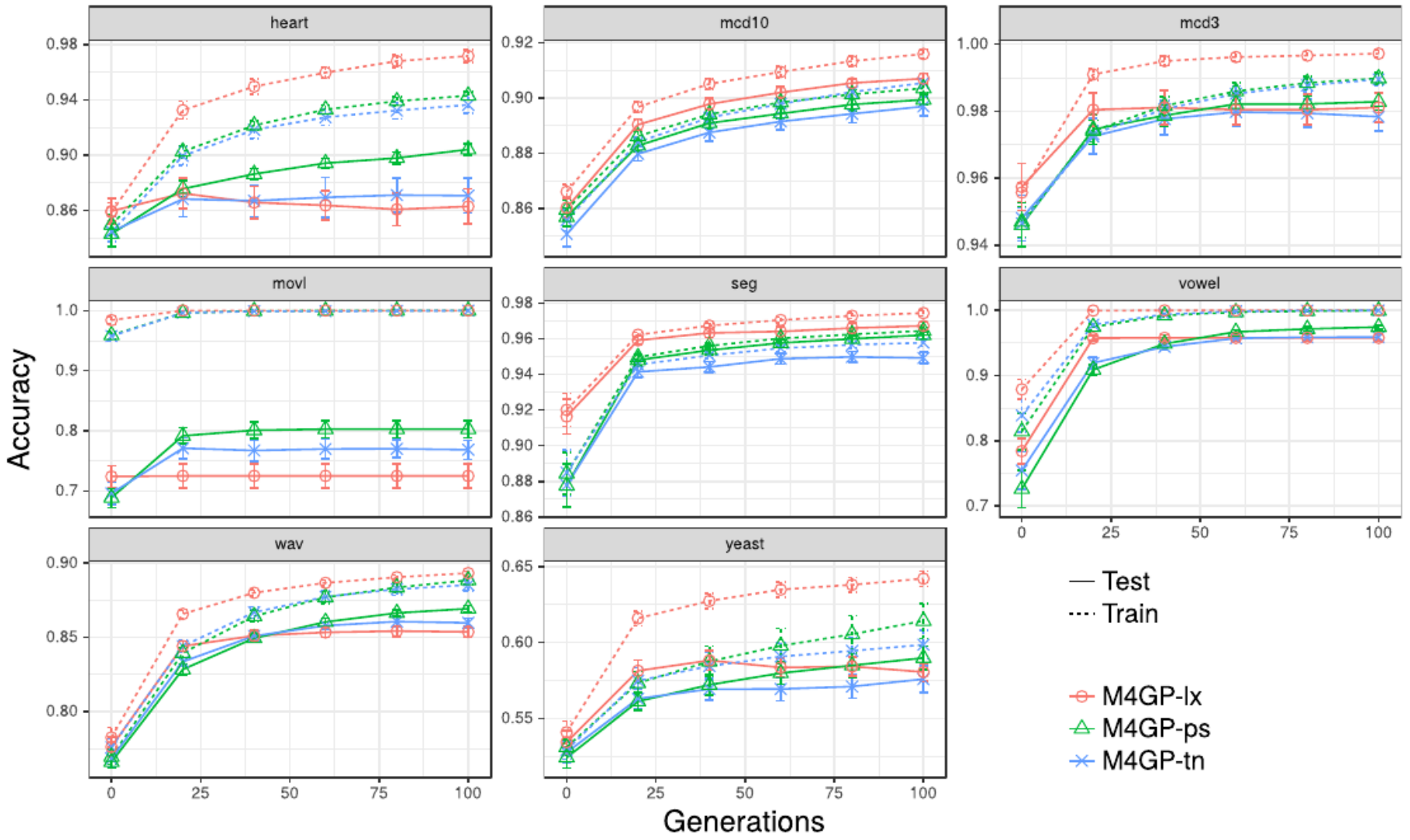
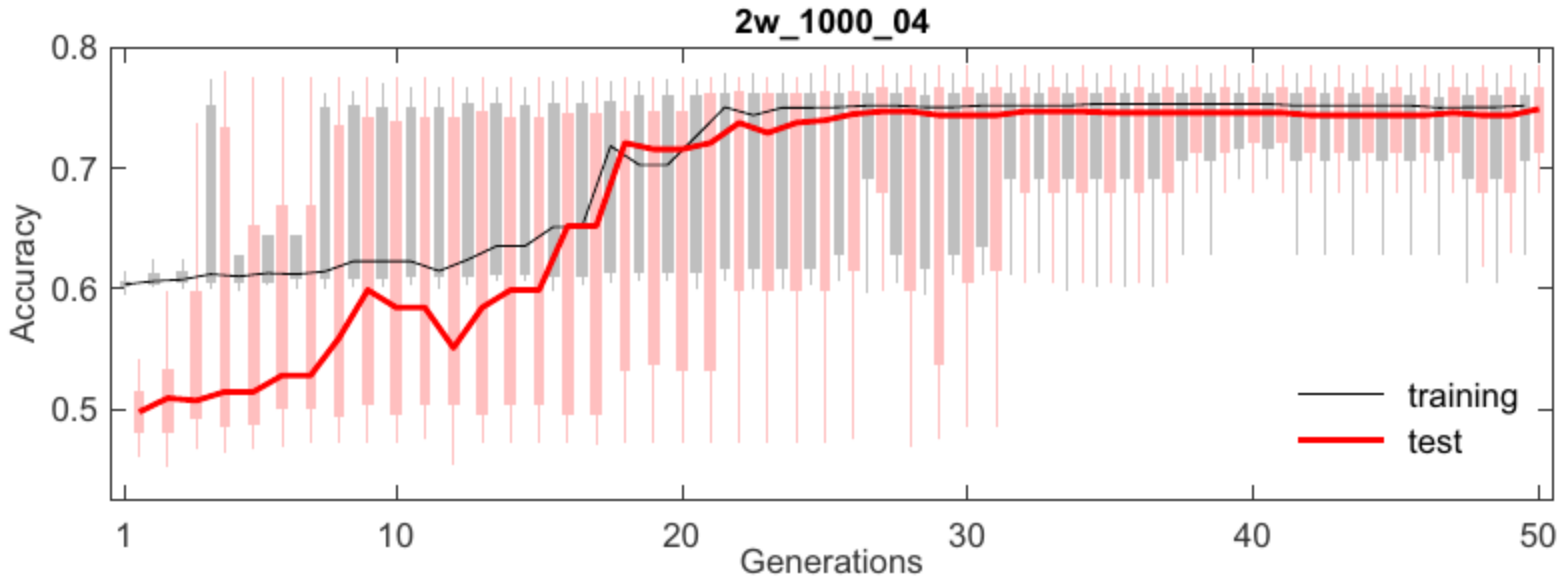


Fig. 5. Mean convergence characteristics of M4GP on the training set (dotted lines) and test set (solid lines) over 100 generations. Shapes indicate different selection methods. Error bars denote the confidence intervals.


2w_1000_04
0.8
0.7
0.6
0.5
Accuracy
training
test
1
10
20
30
40
50
Generations

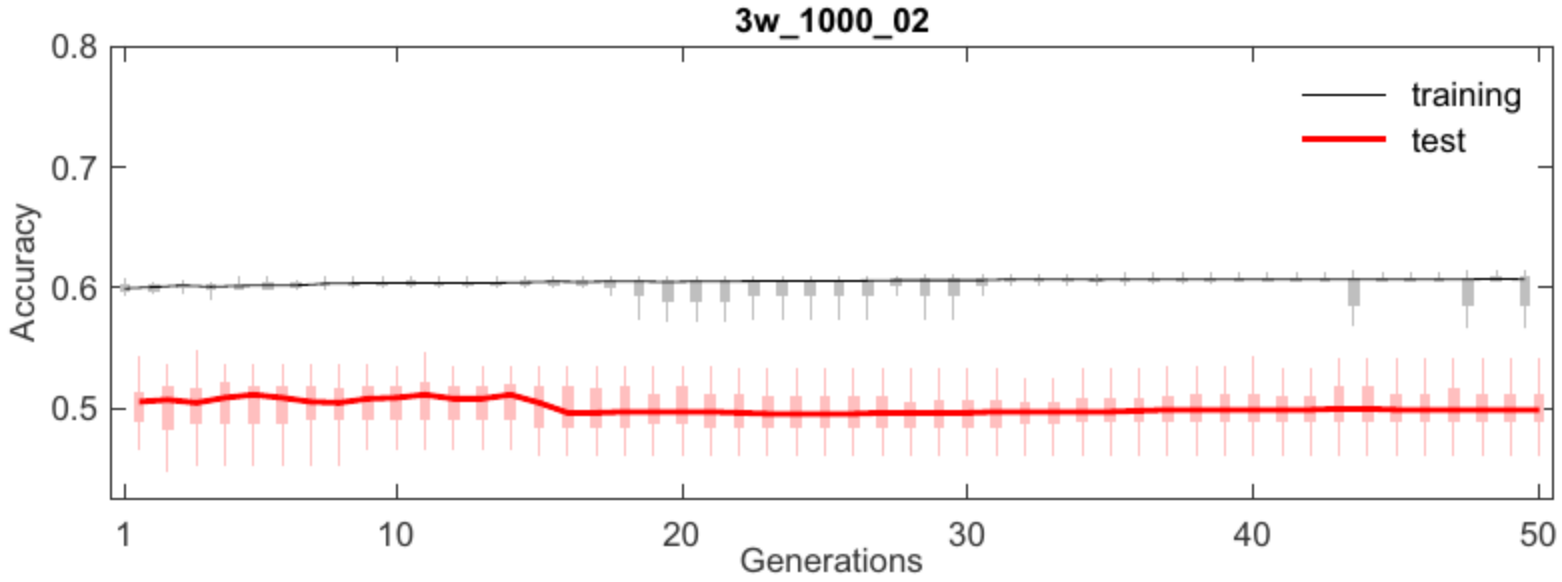
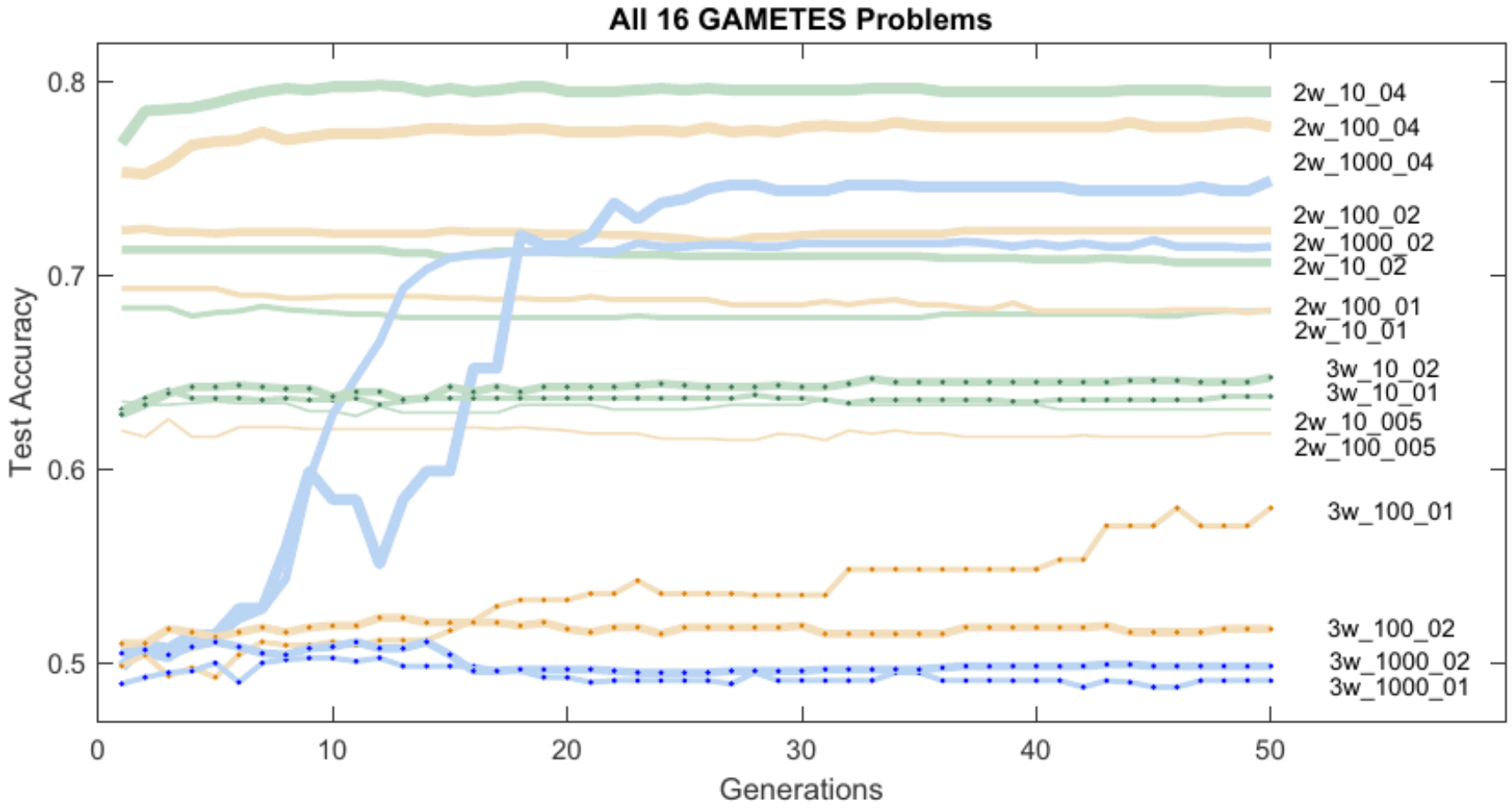
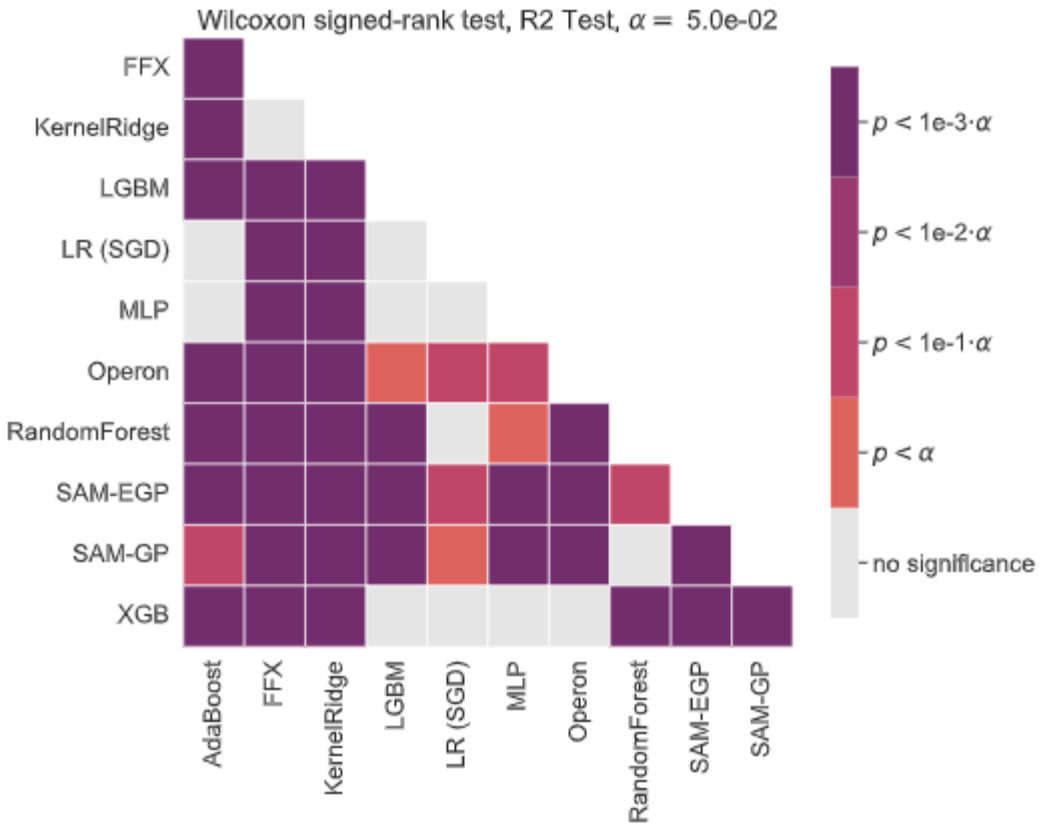

3w_1000_02
0.8
0.7
0.6
0.5
Accuracy
training
test
1
10
20
30
40
50
Generations

Fig. 8 Evolution of test accuracy on all the GAMETES problems. Lines are green/yellow/blue for 10/100/1000-feature problems; lines are thicker for higher signal-to-noise ratios; lines are dotted for the three-way problems. The names of the problems on the right appear by the same order as the lines



Statistical significance test results

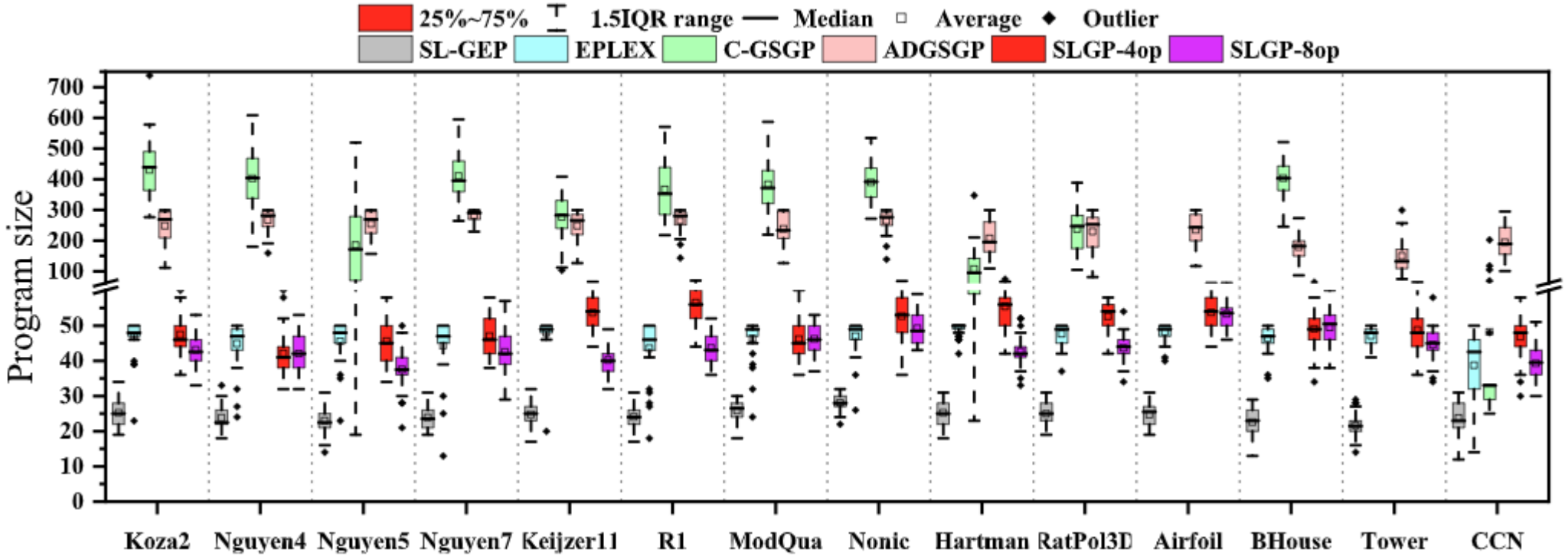
	SGP_E	IGP	IGP_E	FIGP	FIGP_E	SGP_E	IGP	IGP_E	FIGP	FIGP_E
	547_no2					1193_BNG_lowbwt				
SGP	3.08e-1	9.03e-1	3.71e-1	**4.60e-4**	**8.86e-5**	7.15e-1	**1.85e-2**	3.93e-1	3.81e-1	**3.73e-9**
SGP_E		3.18e-1	6.70e-2	**4.97e-5**	**1.22e-5**		**2.09e-2**	7.61e-1	1.98e-1	**8.01e-8**
IGP			2.29e-1	**1.37e-2**	**5.01e-3**			**2.02e-3**	1.98e-1	**1.19e-6**
IGP_E				**1.64e-2**	**3.74e-3**				**2.62e-2**	**9.31e-9**
FIGP					3.49e-1					**1.30e-8**
	503_wind					195_auto_price				
SGP	2.89e-1	**4.18e-4**	6.85e-1	**7.61e-3**	**3.85e-7**	**5.59e-9**	**1.72e-3**	3.09e-1	**9.31e-9**	**1.99e-6**
SGP_E		**2.37e-3**	1.91e-1	**3.64e-2**	**1.86e-9**		**2.05e-7**	**1.68e-6**	**1.86e-9**	**1.86e-9**
IGP			**7.98e-4**	8.24e-1	**1.37e-4**			**2.56e-3**	**1.30e-8**	**8.33e-7**
IGP_E				**4.60e-4**	**1.86e-9**				**1.30e-8**	**2.32e-4**
FIGP					**4.67e-3**					**3.86e-7**
	583_fri_c1_1000_50					588_fri_c4_1000_100				
SGP	1.84e-1	6.36e-2	9.19e-1	**3.22e-3**	**4.18e-4**	**2.62e-2**	2.53e-1	4.40e-1	**2.02e-3**	3.09e-1
SGP_E		**4.84e-4**	**3.64e-2**	**8.86e-5**	**2.76e-6**		**2.83e-4**	1.29e-1	**1.68e-6**	**2.62e-4**
IGP			1.14e-1	2.13e-1	**2.37e-3**			**2.34e-2**	**2.62e-2**	5.03e-1
IGP_E				**5.01e-3**	**1.40e-5**				**1.60e-5**	9.20e-2
FIGP					2.21e-1					**9.30e-3**
	4544_GeographicalOriginalofMusic					505_tecator				
SGP	1.24e-1	2.67e-1	1.29e-1	**3.73e-9**	**1.70e-4**	**1.30e-7**	**1.30e-2**	9.61e-2	**8.00e-3**	**9.31e-9**
SGP_E		6.70e-1	9.68e-1	**1.86e-8**	5.01e-3		3.82e-1	**6.20e-3**	1.19e-1	5.77e-2
IGP			8.71e-1	**3.73e-9**	**1.72e-3**			3.39e-1	7.77e-1	5.49e-2
IGP_E				**1.86e-9**	**1.11e-4**				5.43e-1	**1.70e-4**
FIGP					**5.15e-6**					**1.21e-2**



Model Size, #feature …

AVERAGE FEATURE SUBSET SIZE OBTAINED BY VGS-MOEA AND ITS TWO VARIANTS

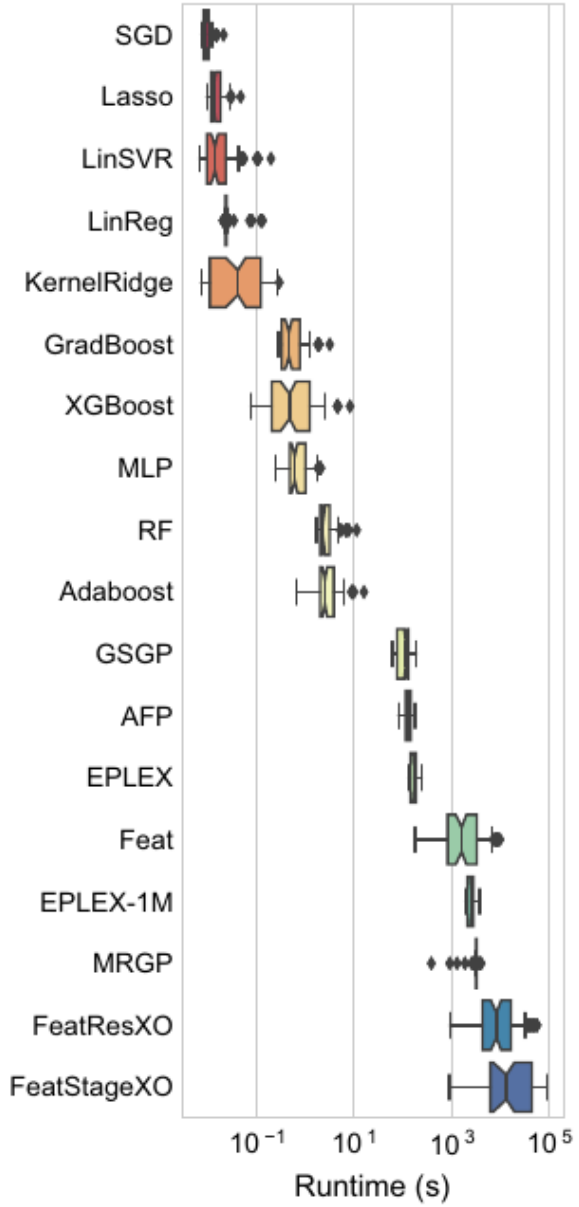
Data Set	VGS-MOEA-NAR	VGS-MOEA-NGO	VGS-MOEA
COIL20	94.40 ± 36.32(+)	160.20 ± 44.34(+)	**59.03** ± 25.91
SRBCT	155.13 ± 25.56(+)	200.07 ± 41.58(+)	**29.33** ± 21.57
PCMAC	368.13 ± 138.28(+)	343.23 ± 109.86(+)	**50.33** ± 10.67
lymphoma	326.33 ± 70.82(+)	407.10 ± 70.59(+)	**70.70** ± 47.53
GLIOMA	412.80 ± 100.16(+)	558.87 ± 117.29(+)	**23.97** ± 26.06
BASEHOCK	465.20 ± 151.87(+)	475.87 ± 153.12(+)	**65.57** ± 19.56
TOX_171	570.53 ± 166.54(+)	776.63 ± 208.77(+)	**102.07** ± 72.61
Brain1	495.63 ± 204.95(+)	685.07 ± 138.21(+)	**36.47** ± 31.56
leukemia	436.30 ± 68.84(+)	552.97 ± 132.53(+)	**24.17** ± 18.69



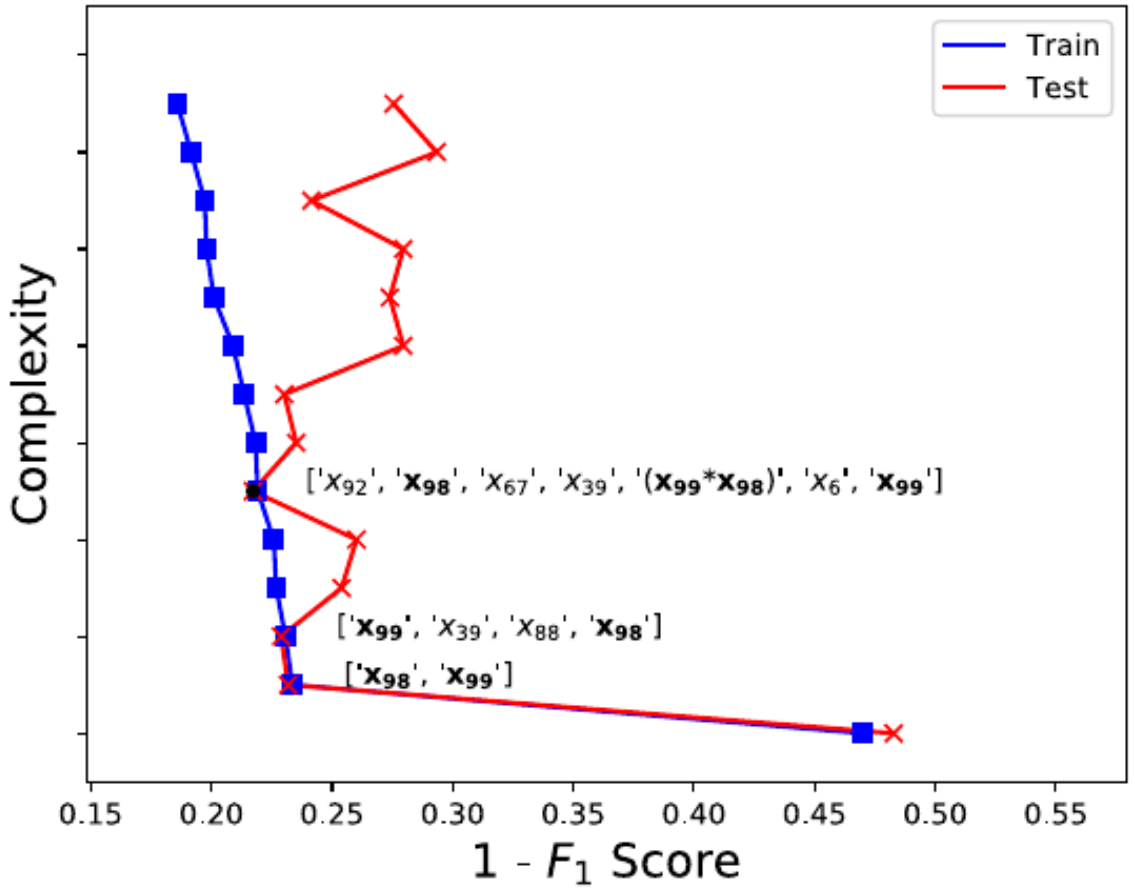
Computational Time

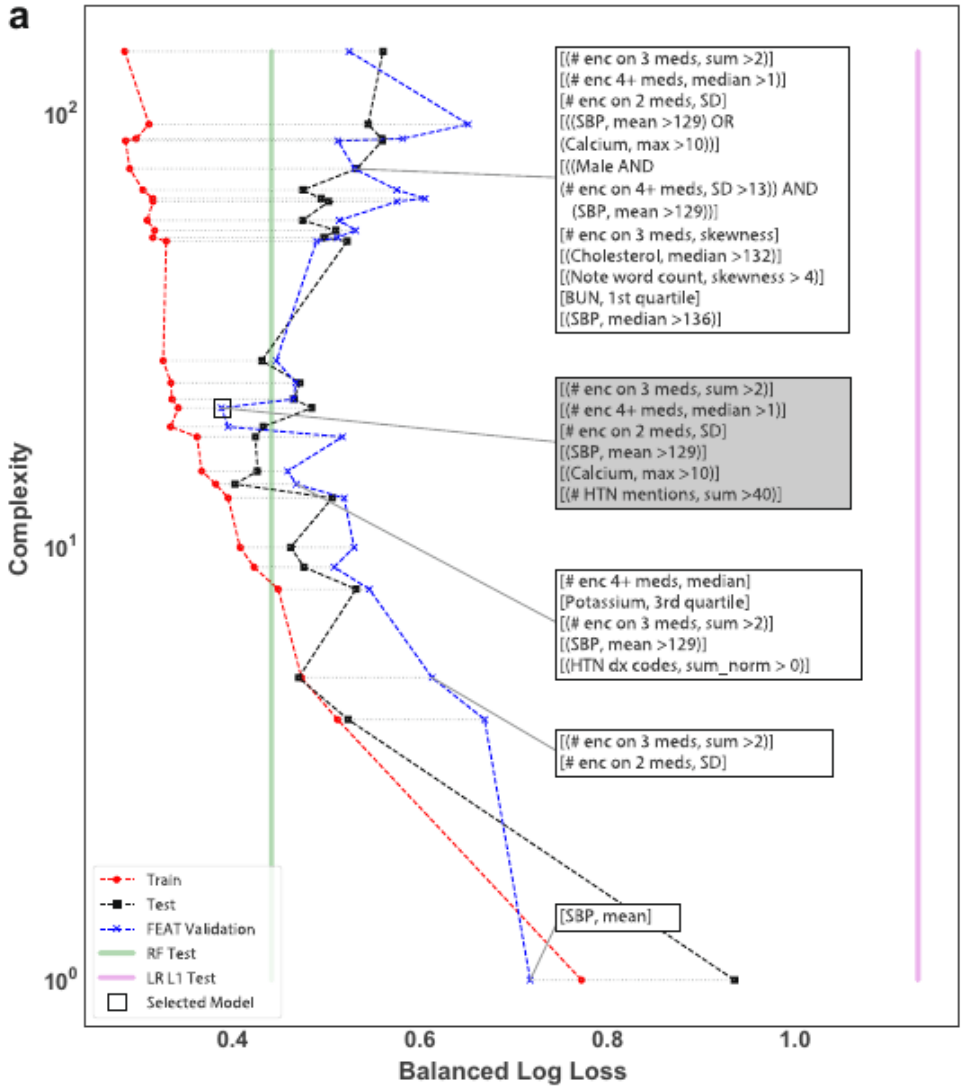
TABLE V
COMPUTATION TIME (IN MINUTES)

Dataset	CSO	CCSO
Vehicle	**3.44**	4.41
WallRobot	92.12	**88.08**
German	**4.07**	4.66
GuesterPhase	375.01	**303.71**
Ionosphere	**1.21**	1.90
Chess	71.28	**37.68**
Movementlibras	**1.11**	1.42
Hillvalley	5.04	**4.78**
Musk1	1.81	**1.75**
USPS	240.33	**220.13**
Madelon	26.59	**25.35**
Isolet	**11.92**	15.80
MultipleFeatures	**18.80**	20.43
Gametes	14.93	**13.58**
QAR	**14.64**	22.51
QOT	308.07	**270.56**
COIL20	12.05	**11.20**
ORL	2.03	**1.94**
Bioresponse	**68.61**	77.73
RELATHE	**28.34**	29.86
GLIOMA	**0.51**	0.96
BASEHOCK	**55.81**	58.67
Gisette	**464.45**	487.06
Brain1	**0.93**	1.71



Fronts


Train
Test
Complexity
['x_{92}', '**x98**', 'x_{67}', 'x_{39}', '(**x99***x98)', 'x_6', '**x99**']
['**x99**', 'x_{39}', 'x_{88}', '**x98**']
['**x98**', '**x99**']
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
1 - F_1 Score


a
Complexity
10^2
10^1
10^0
[(# enc on 3 meds, sum >2)]
[(# enc 4+ meds, median >1)]
[# enc on 2 meds, SD]
[((SBP, mean >129) OR
(Calcium, max >10))]
[((Male AND
(# enc on 4+ meds, SD >13)) AND
(SBP, mean >129))]
[# enc on 3 meds, skewness]
[(Cholesterol, median >132)]
[(Note word count, skewness > 4)]
[BUN, 1st quartile]
[(SBP, median >136)]
[(# enc on 3 meds, sum >2)]
[(# enc 4+ meds, median >1)]
[# enc on 2 meds, SD]
[(SBP, mean >129)]
[(Calcium, max >10)]
[(# HTN mentions, sum >40)]
[# enc 4+ meds, median]
[Potassium, 3rd quartile]
[(# enc on 3 meds, sum >2)]
[(SBP, mean >129)]
[(HTN dx codes, sum_norm > 0)]
[(# enc on 3 meds, sum >2)]
[# enc on 2 meds, SD]
[SBP, mean]
Train
Test
FEAT Validation
RF Test
LR L1 Test
Selected Model
0.4
0.6
0.8
1.0
Balanced Log Loss

Results

- Keep the results of the standard algorithm
- Do keep all the original results from each run.
- Perform statistical significance tests: T-test; Wilcoxon test; Friedman test
- Do NOT delete results unless they use too much memory, or they are wrong

Codes and Programs

- Backup different versions of your codes using
  - Version control: Gitlab, bitbucket
- Make clear documentation of your codes
- Make clear README documentation
- Organise your files: a single directory for a project, containing all of the data, code, and results for your project
- Following best programming practices: choose file/method/variable name carefully, avoid hard coding, …

THANK

Questions?