Project C

Raw data

Readme results folder

Folder structure:

results/{dry, wet}/{crf}/{segments}/{sample}/{Algo}

Then in every folder you will find:

solution.csv → containing the parameters of the best solution for each trial.

purity.csv → containing three columns with the peak purity, time of first eluting compound and time of last eluting compound respectively, for each trial.

runtime.csv → contains a line for each trial, with the cumulative runtime for each iteration on each line. For the total runtime of the trial, take the last element.

score.csv → contains a line for each trial with the score at each iteration. Note: CMA has negative vaues in score.csv (it assumes minimization and the negative values were stored).

best_solution_chrom_x.png → contains a plot of the chromatogram of the best solution for each trial.

Note that GridSearch looks slightly different, as it is run on a grid, it only contains the total runtime, single best score, single best purity/min_time/max_time.

Information regarding number of iterations

Every method has been run for 10 trials with different random seeds.

Bayesian optimization → Ran for 200 iterations + 10 initial datapoints

Random search → Ran for 10010 iterations/function evaluations

GenAlgo/DiffEvo/CMA → Ran for 1001 iterations, but use a population of 10, so 10010 function evaluations, runtime and score as saved per iteration. runtime without the first random initialization. So scores.csv has length 1001, runtime.csv has length 1000.

CMA → ran for 10000 function evaluations, first 10 of population not plotted

GridSearch ran for:

- 1 segments → N=10 (10 ** (2 + 2)) = 10000 function evaluations
- 2 segments → N =5 (5 ** (4 +2)) = 15625 function evaluations
- 3 segments → N = 4 (4 ** (6 +2)) = 65536 function evaluations
- 4 segments → N = 3 (3 ** (8 +2)) = 59049 function evaluations

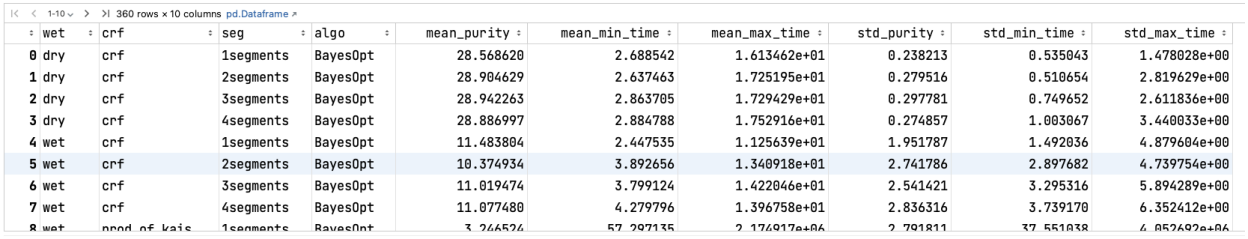
Grid search is restricted to (N ** (2*segments + 2), so it was either way too big or too small compared to other budgets

Processed data

Purity files

This data should be easier to digest and easier to make plots with.

purity_mintime_maxtime_averaged_over_trials.csv



1-10 360 rows × 10 columns pd.Dataframe

	wet	crf	seg	algo	mean_purity	mean_min_time	mean_max_time	std_purity	std_min_time	std_max_time
0	dry	crf	1segments	BayesOpt	28.568620	2.688542	1.613462e+01	0.238213	0.535043	1.478028e+00
1	dry	crf	2segments	BayesOpt	28.904629	2.637463	1.725195e+01	0.279516	0.510654	2.819629e+00
2	dry	crf	3segments	BayesOpt	28.942263	2.863705	1.729429e+01	0.297781	0.749652	2.611836e+00
3	dry	crf	4segments	BayesOpt	28.886997	2.884788	1.752916e+01	0.274857	1.003067	3.440033e+00
4	wet	crf	1segments	BayesOpt	11.483804	2.447535	1.125639e+01	1.951787	1.492036	4.879604e+00
5	wet	crf	2segments	BayesOpt	10.374934	3.892656	1.340918e+01	2.741786	2.897682	4.739754e+00
6	wet	crf	3segments	BayesOpt	11.019474	3.799124	1.422046e+01	2.541421	3.295316	5.894289e+00
7	wet	crf	4segments	BayesOpt	11.077480	4.279796	1.396758e+01	2.836316	3.739170	6.352412e+00
8	wet	nnod_of_kais	1segments	BayesOpt	3.246524	57.297135	2.174917e+06	2.791811	37.551038	4.052692e+06

Here the mean and std over the columns of purity.csv is taken over all of the 10 trials for each combination.

purity_mintime_maxtime_averaged_over_trials_and_samples.csv

Here an additional mean and std is taken over all of the samples.

Scores and runtime files

Within the folder structure:

results/{dry, wet}/{crf}/{segments}/{sample}/{Algo}

In every folder you will find:

best_so_far.csv

These files are the same as *score.csv,* but are now made strictly increasing, and are interpolated so they all have the same length, **so way easier to make plots with**.

runtime_ext.csv

Same files as runtime, but now all agorithms with the same length for easier plotting.

runtime_mean.csv, mean runtime over the 10 trials

runtime_std.csv, std of runtime over the 10 trials.

Notebooks

data_preprocessing.ipynb contains the code that was used do make the files discussed in the Processed section.

paper_figures.ipynb **contains the code to make some example plots, this probably is a good starting point for our discussion/analysis.**