Jožef Stefan Institute

JOŽEF STEFAN INSTITUTE
COMPUTER SYSTEMS

Trustworthy Benchmarking for Black-Box Single-Objective Optimization

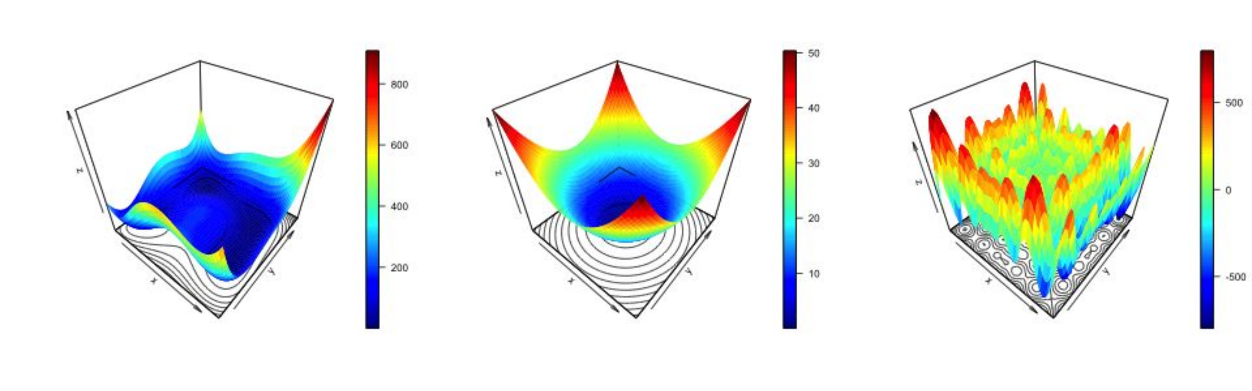
Tome Eftimov
tome.eftimov@ijs.si

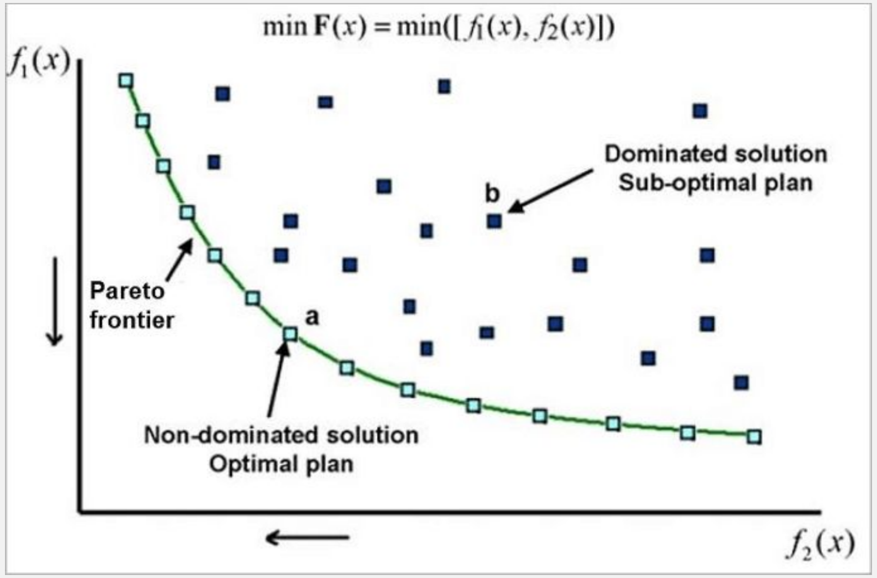
16th International Joint Conference on Computational Intelligence (**IJCCI 2024**)
1st International Conference on Explainable AI for Neural and Symbolic Methods (**EXPLAINS 2024**)
20-22 November, 2024
Porto, Portugal

Outline

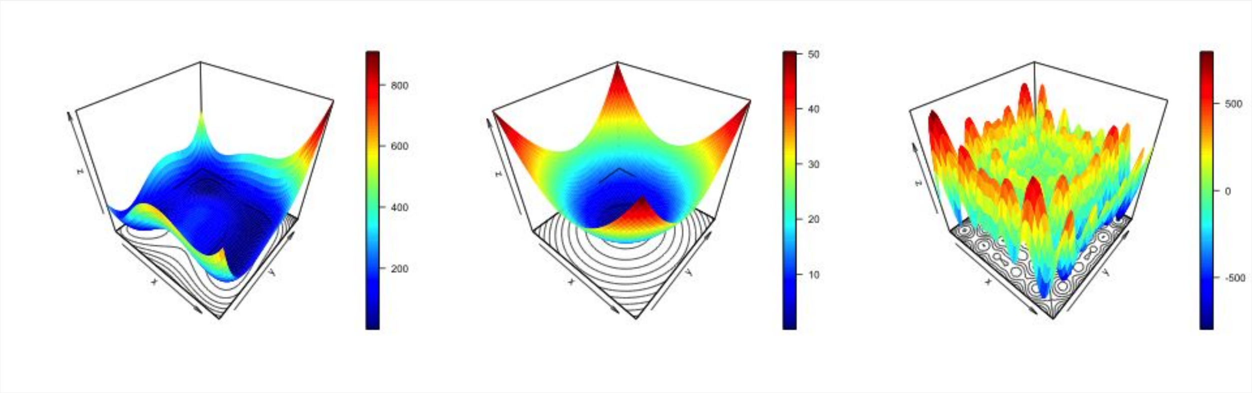
- Benchmarking
- Learning for optimization/ Meta-learning
- Algorithm footprints
- Take home messages

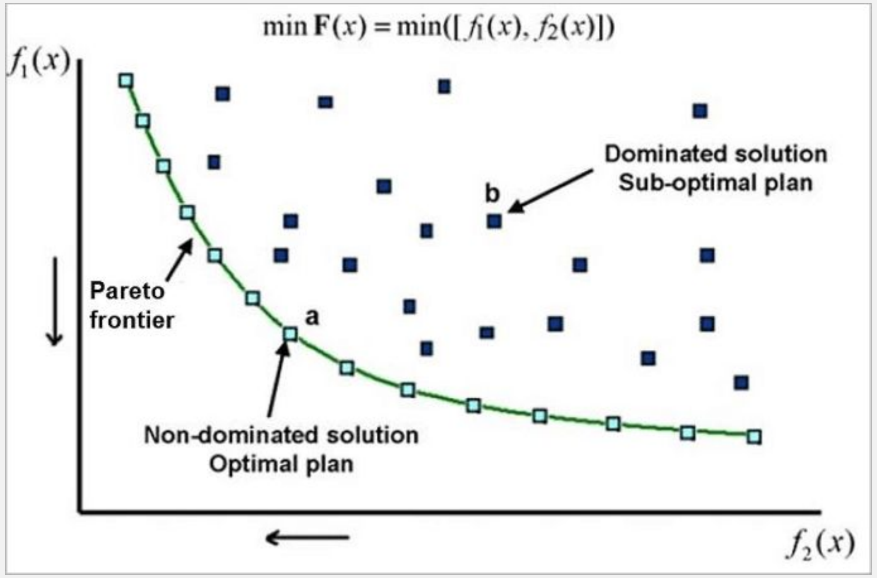
Single-Objective vs. Multi-Objective Optimization





Single-Objective vs. Multi-Objective Optimization





Benchmarking

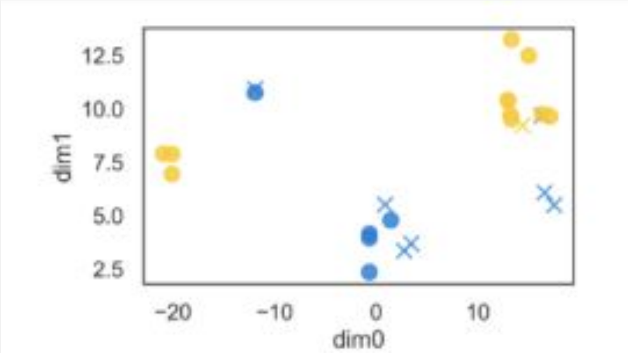
- Which **problems to be selected**?
- Which **algorithms to be selected**?
- Fair experimental design
  - hyperparameter tuning
- Which **statistical analyses** to be applied?

OPTION

- Benchmarking optimization ontology
- Performance data
- Problem landscape features
- Different benchmark suites

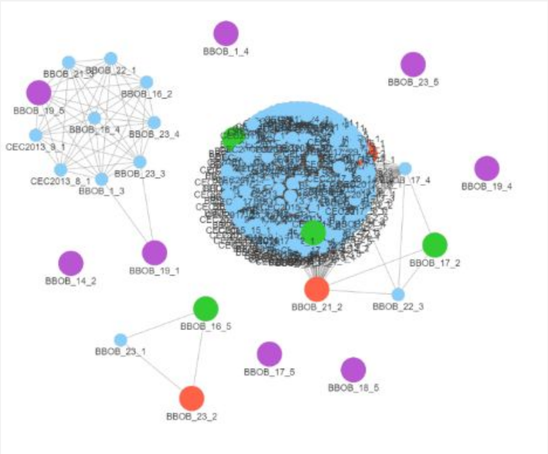
Algorithm Footprint

- Explainable performance
- Set of easily and challenging solvable problem instances



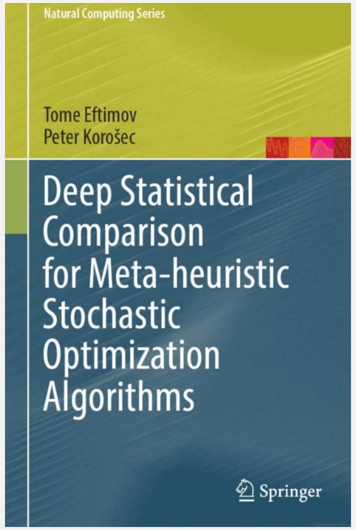
SELECTOR

- Selecting diverse and unbiased problem instances
- Based on problem landscape features



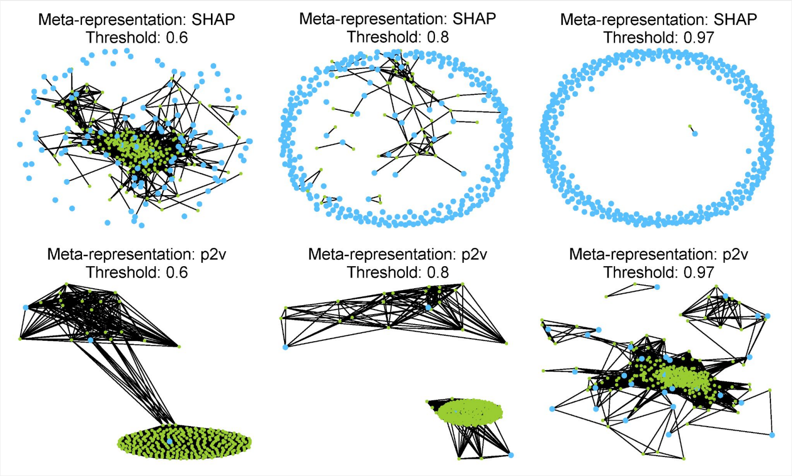
DSC - Deep Statistical Comparison

- Robust statistical comparisons
- Based on results distribution
- Available also for multi-objective



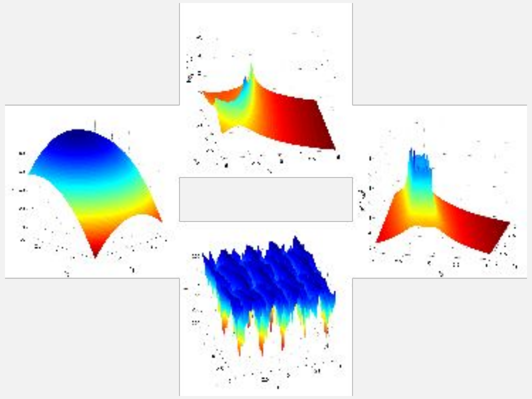
PS-ASS

- Algorithm portfolio selection
- Based on algorithm features



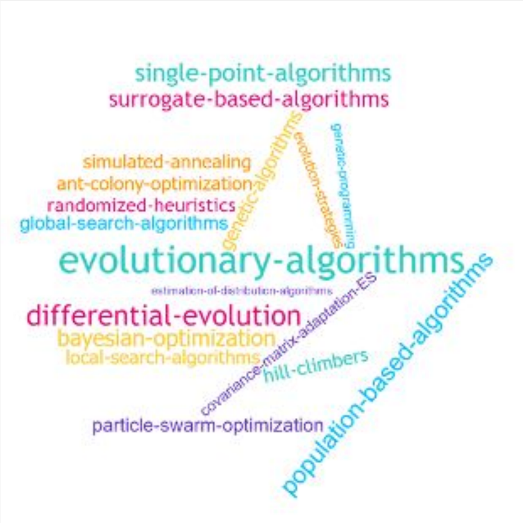
Benchmarking

Selection of a problem portfolio



What types of optimization problems will be included in our problem portfolio?

Selection of an algorithm portfolio



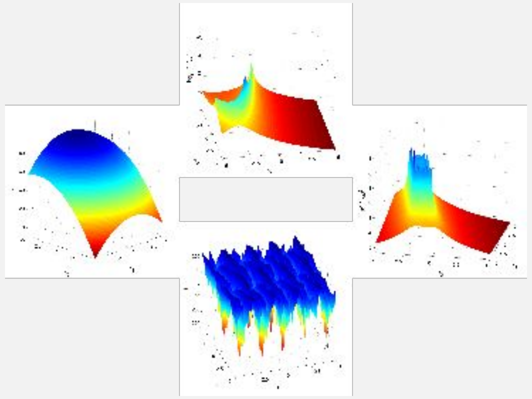
Which algorithms will be incorporated into our algorithm portfolio?

Statistical comparison

Which statistical approach to be utilized?

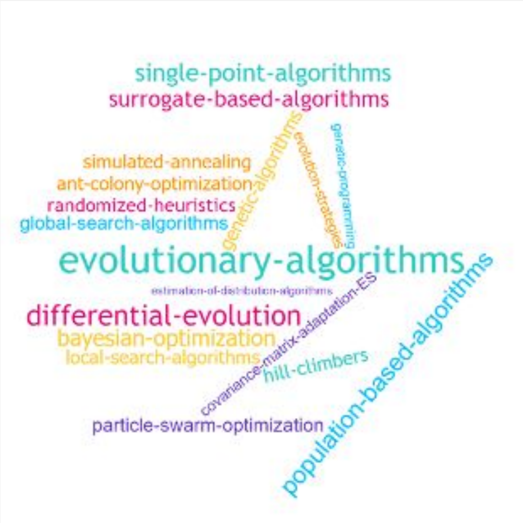
Benchmarking

Selection of a problem portfolio



What types of optimization problems will be included in our problem portfolio?

Selection of an algorithm portfolio

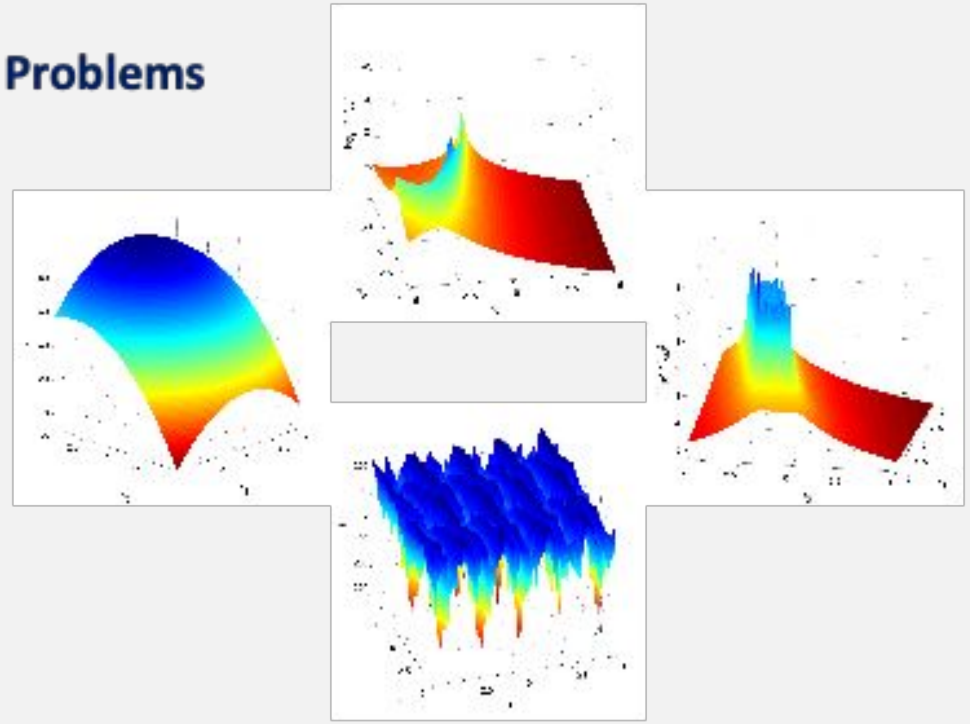


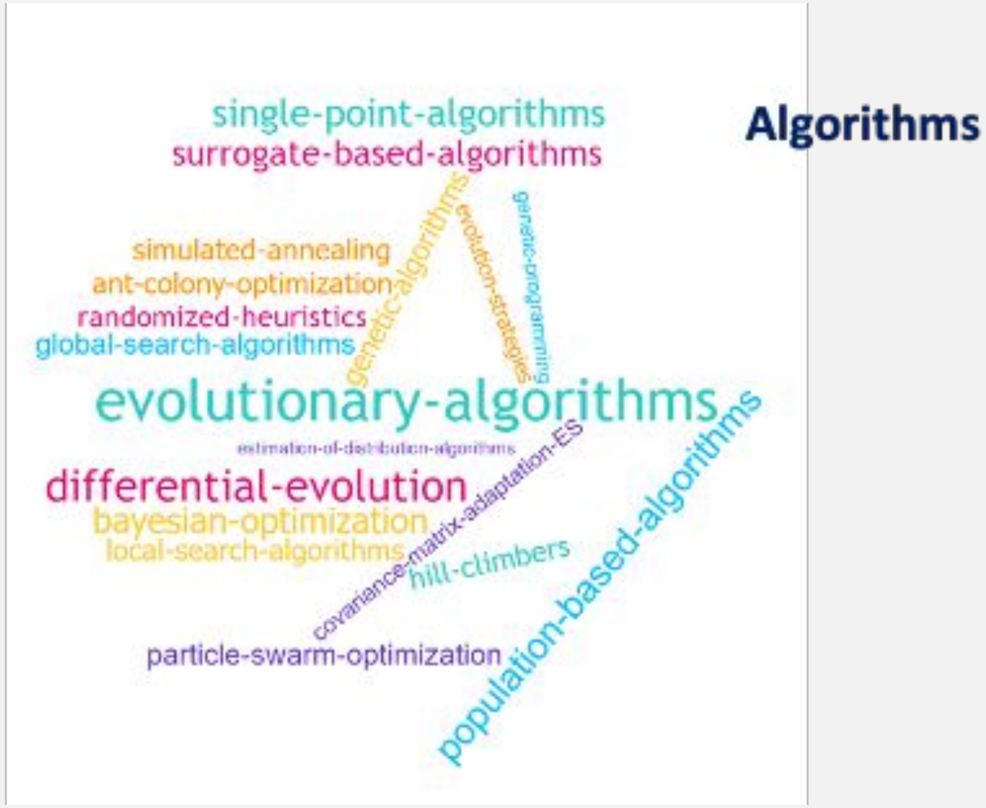
Which algorithms will be incorporated into our algorithm portfolio?

Statistical comparison

Which statistical approach to be utilized?

Statistical comparison





Which algorithm statistically outperforms the others?

State-Of-The-Art before 2017

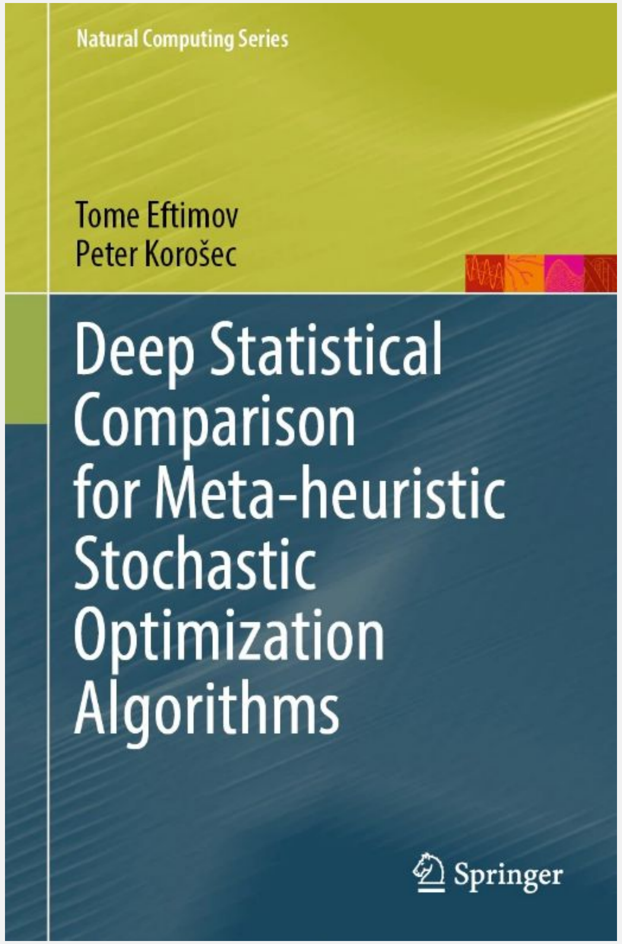
- **Machine Learning**
  - Demšar, J. (2006). **Statistical comparisons of classifiers over multiple data sets**. Journal of Machine learning research, 7(Jan), 1-30.

- **Evolutionary Computation**
  - Derrac, J., Garcia, S., Molina, D., & Herrera, F. (2011). **A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms**. Swarm and Evolutionary Computation, 1(1), 3-18.

  - Garcia, S., Molina, D., Lozano, M., & Herrera, F. (2009). **A study on the use of non-parametric tests for analyzing the evolutionary algorithms' behavior: a case study on the CEC'2005 special session on real parameter optimization**. Journal of Heuristics, 15(6), 617.

Deep Statistical Comparison

- Two steps:
  - A novel ranking scheme based on comparing distribution
  - Use an appropriate statistical test



Statistical comparison of three algorithms

Friedman ranking scheme (medians)

F	BSifeg	BSrr	Srr
f_1	1.50	1.50	3.00
f_2	2.00	1.00	3.00
f_3	2.00	1.00	3.00
f_4	3.00	2.00	1.00
f_5	2.00	2.00	2.00
f_6	2.00	3.00	1.00
f_7	3.00	2.00	1.00
f_8	2.00	1.00	3.00
f_9	3.00	2.00	1.00
f_{10}	2.00	3.00	1.00
f_{11}	2.00	3.00	1.00
f_{12}	3.00	1.00	2.00
f_{13}	2.00	3.00	1.00
f_{14}	3.00	2.00	1.00
f_{15}	3.00	2.00	1.00
f_{16}	2.00	3.00	1.00
f_{17}	3.00	2.00	1.00
f_{18}	2.00	3.00	1.00
f_{19}	2.00	3.00	1.00
f_{20}	3.00	1.00	2.00
f_{21}	3.00	2.00	1.00
f_{22}	3.00	2.00	1.00

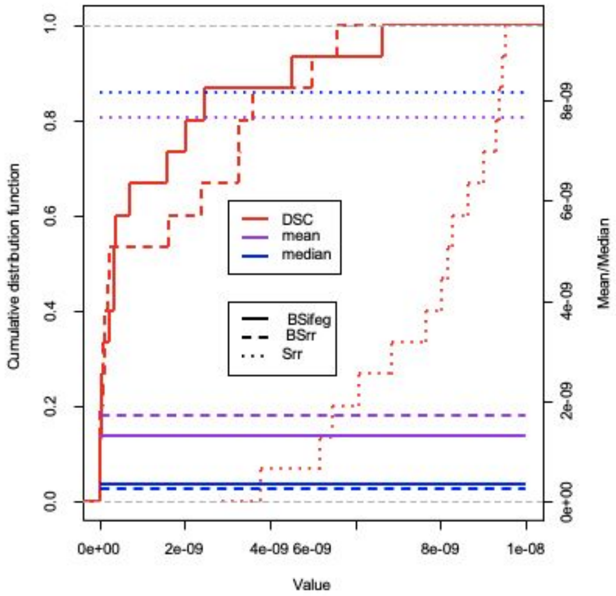
Friedman ranking scheme (averages)

F	BSifeg	BSrr	Srr
f_1	1.50	1.50	3.00
f_2	2.00	1.00	3.00
f_3	1.00	2.00	3.00
f_4	3.00	2.00	1.00
f_5	2.00	2.00	2.00
f_6	2.00	3.00	1.00
f_7	3.00	2.00	1.00
f_8	1.00	2.00	3.00
f_9	3.00	2.00	1.00
f_{10}	3.00	2.00	1.00
f_{11}	1.00	3.00	2.00
f_{12}	2.00	1.00	3.00
f_{13}	2.00	3.00	1.00
f_{14}	3.00	2.00	1.00
f_{15}	3.00	2.00	1.00
f_{16}	2.00	3.00	1.00
f_{17}	3.00	2.00	1.00
f_{18}	2.00	3.00	1.00
f_{19}	3.00	2.00	1.00
f_{20}	3.00	2.00	1.00
f_{21}	3.00	2.00	1.00
f_{22}	2.00	3.00	1.00

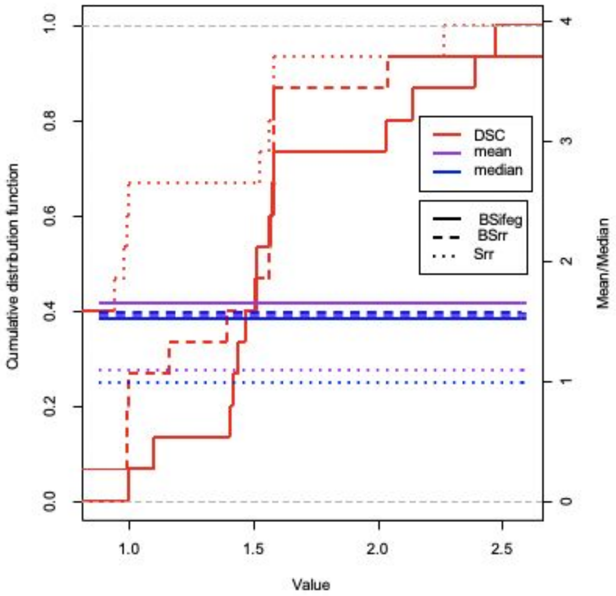
DSC ranking scheme

F	BSifeg	BSrr	Srr
f_1	1.50	1.50	3.00
f_2	1.50	1.50	3.00
f_3	1.50	1.50	3.00
f_4	2.00	2.00	2.00
f_5	2.00	2.00	2.00
f_6	2.00	3.00	1.00
f_7	2.00	2.00	2.00
f_8	2.00	2.00	2.00
f_9	2.00	2.00	2.00
f_{10}	2.00	2.00	2.00
f_{11}	2.00	2.00	2.00
f_{12}	2.00	2.00	2.00
f_{13}	2.00	3.00	1.00
f_{14}	2.00	2.00	2.00
f_{15}	2.00	2.00	2.00
f_{16}	2.00	2.00	2.00
f_{17}	2.00	2.00	2.00
f_{18}	2.00	2.00	2.00
f_{19}	3.00	2.00	1.00
f_{20}	2.00	2.00	2.00
f_{21}	2.00	2.00	2.00
f_{22}	2.00	2.00	2.00

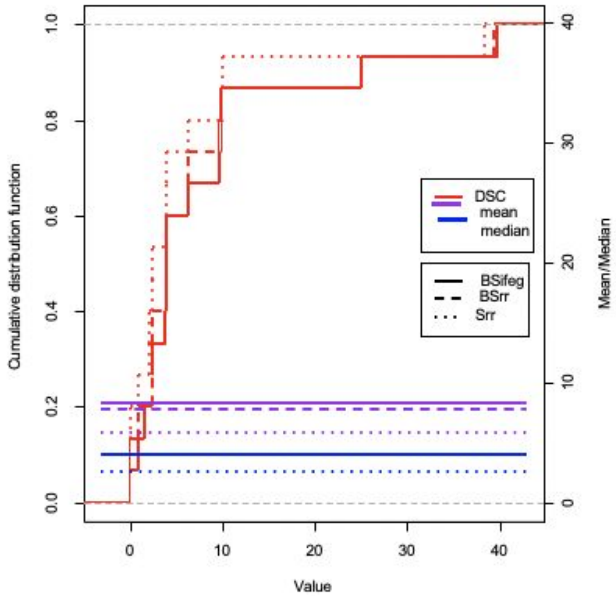
Comparison on a single problem



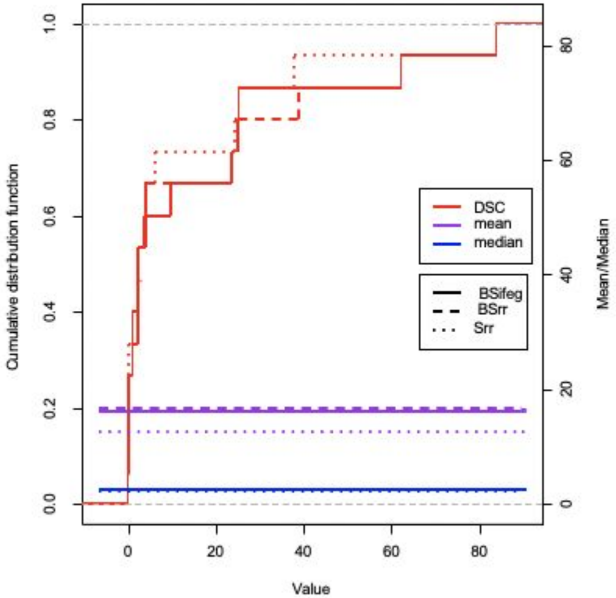
f_3



f_{19}



f_{21}



f_{22}

DSC tutorials

IJCCI 2018

10th International Joint Conference on Computational Intelligence

SSCI 2019

PPSN 2022

GECCO 2020 @ Cancun

GECCO 2021 @ Lille

GECCO 2022 @ Boston

GECCO 2024 @ Melbourne

2021 IEEE

CONGRESS ON EVOLUTIONARY COMPUTATION

28.06–1.07.2021 • Kraków • POLAND

IEEE 2023 Congress on Evolutionary Computation

WCCI2022

IEEE WORLD CONGRESS ON COMPUTATIONAL INTELLIGENCE

a u t o ML

AutoML-Conf 2022

Swarm Intelligence (2022) 16:1–6
https://doi.org/10.1007/s11721-021-00202-9

Metaphor-based metaheuristics, a call for action: the elephant in the room

Claus Aranha[1] · Christian L. Camacho Villalón[2] · Felipe Campelo[3] · Marco Dorigo[2] · Rubén Ruiz[4] · Marc Sevaux[5] · Kenneth Sörensen[6] · Thomas Stützle[2]

Received: 29 July 2021 / Accepted: 28 August 2021 / Published online: 30 November 2021

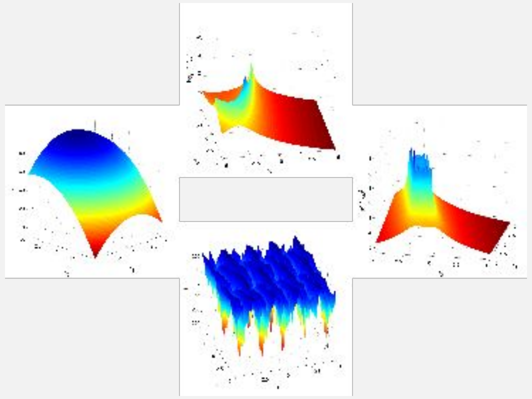
Taking inspiration from natural behaviors to devise new optimization algorithms has played an important role in the history of the field of metaheuristics (Sörensen et al. 2017). Unfortunately, in the last two decades we have been witnessing a new trend by which dozens of metaphor-based metaheuristics based on the most diverse possible set of natural, artificial, social, and sometimes even supernatural phenomena and behaviors are proposed, without a clear motivation beyond the desire of their authors to publish their papers.

Confronting the Elephant in the Room

- We often lack a clear understanding of an AI algorithm strengths and weaknesses.
- Why does an algorithm outperform others?
- **Understanding how algorithms and optimization problems interact could help identify factors that make certain problems easier or more difficult for specific algorithms!!!**

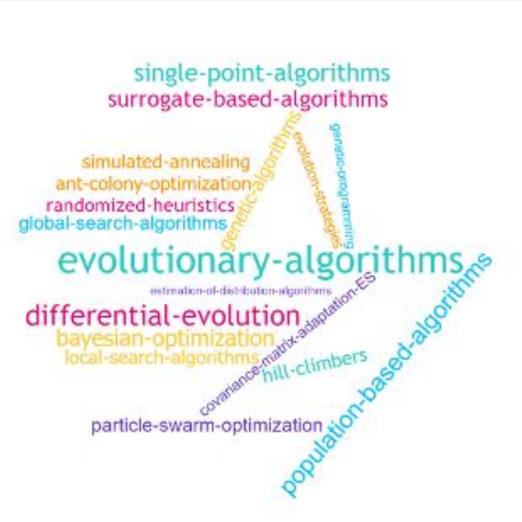
Learning for optimization/Meta-learning

Selection of a problem portfolio



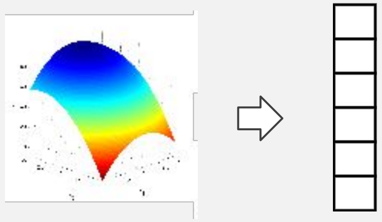
What types of optimization problems will be included in our problem portfolio?

Selection of an algorithm portfolio



Which algorithms will be incorporated into our algorithm portfolio?

Feature Representation



- Problem features
- Algorithm features
- Problem-algorithm trajectory features

How do we represent optimization problems and algorithms in vector form?

Algorithm selector (AS)

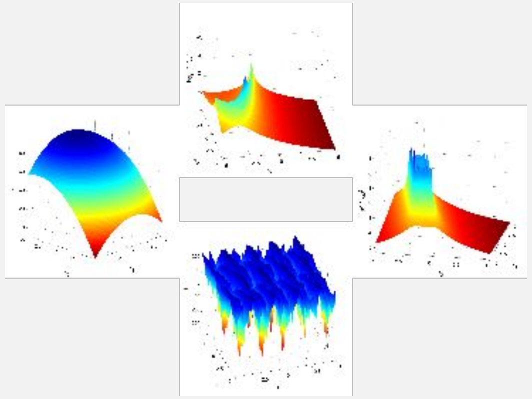
AS approaches:
- (Pairwise-)regression
- (Pairwise-)classification
- …

ML methods:
- RandomForest
- XGBoost
- TabPFN
- FTTransformer
- …

No significant difference in performance of different ML models and AS approaches for BBOB!!!

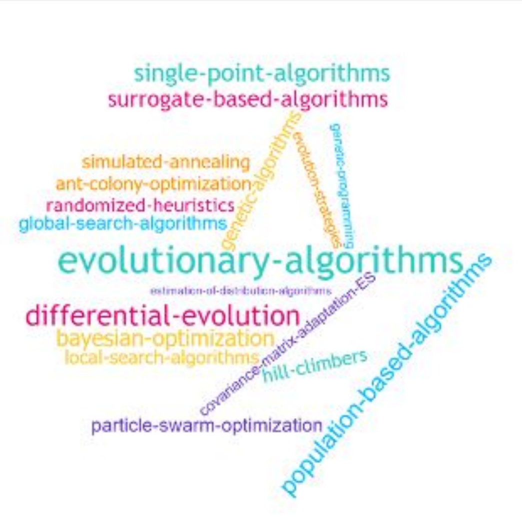
Learning for optimization/Meta-learning

Selection of a problem portfolio



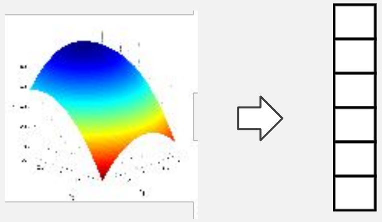
What types of optimization problems will be included in our problem portfolio?

Selection of an algorithm portfolio



Which algorithms will be incorporated into our algorithm portfolio?

Feature Representation



- Problem features
- Algorithm features
- Problem-algorithm trajectory features

How do we represent optimization problems and algorithms in vector form?

Algorithm selector (AS)

AS approaches:
- (Pairwise-)regression
- (Pairwise-)classification
- …

ML methods:
- RandomForest
- XGBoost
- TabPFN
- FTTransformer
- …

No significant difference in performance of different ML models and AS approaches for BBOB!!!

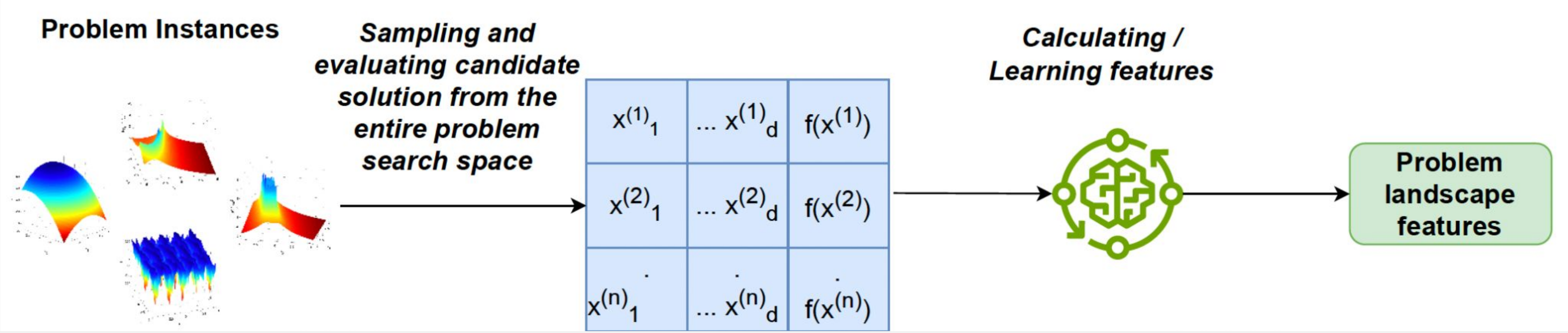
Feature representations

- **Problem features**
  - static features that describe characteristics of an optimization problem
  - Use cases: *complementarity between benchmark suite, selection of a representative learning/benchmarking data*

- **Algorithm features**
  - describe the algorithm characteristics
  - Use case: *selection of complementary algorithm portfolio*

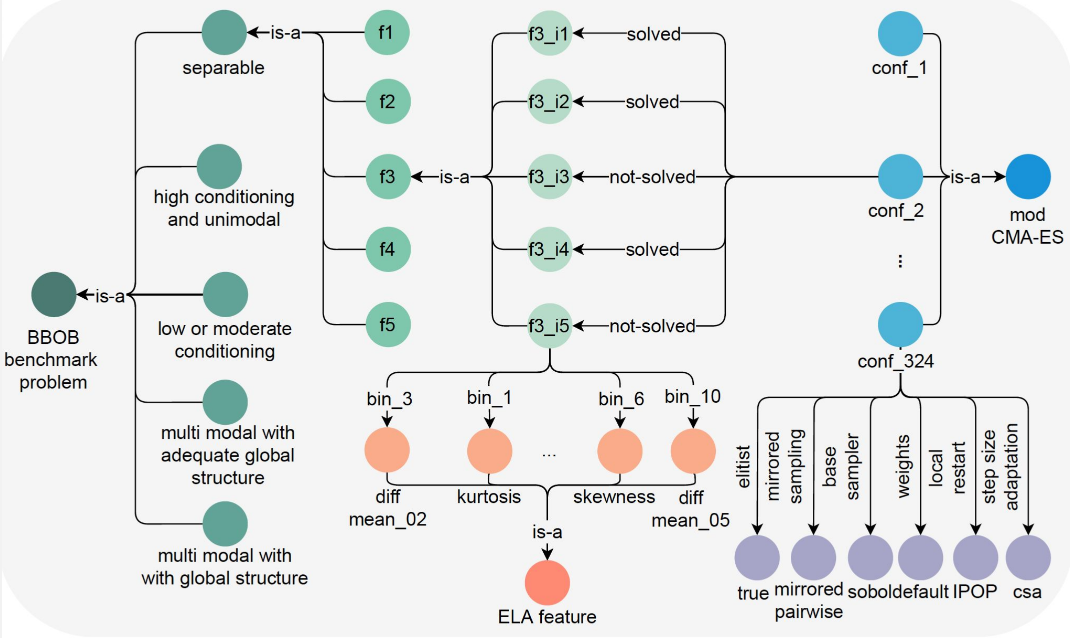
- **Problem-algorithm trajectory features**
  - describe the interactions or the optimization process trajectory when an algorithm is run on a specific problem instances
  - Use cases: *per-run algorithm selection, understanding algorithm behaviour*

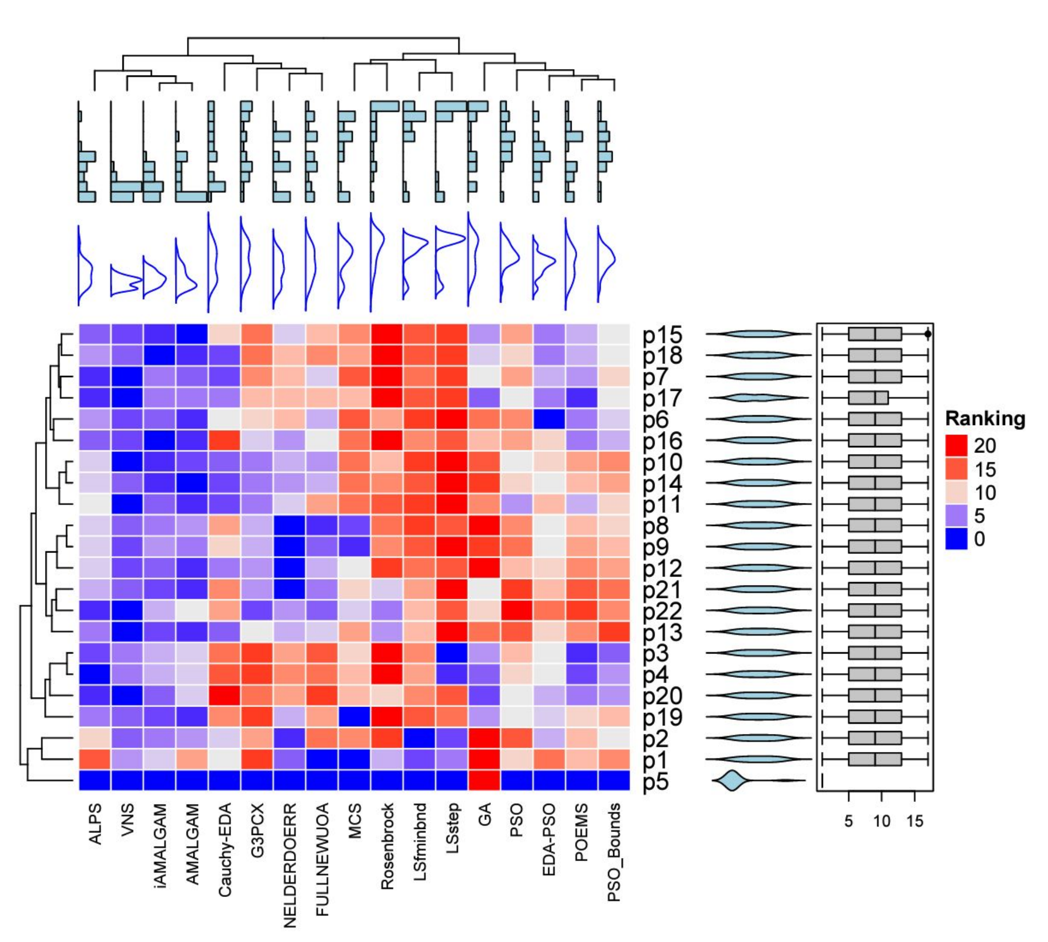
Problem Features



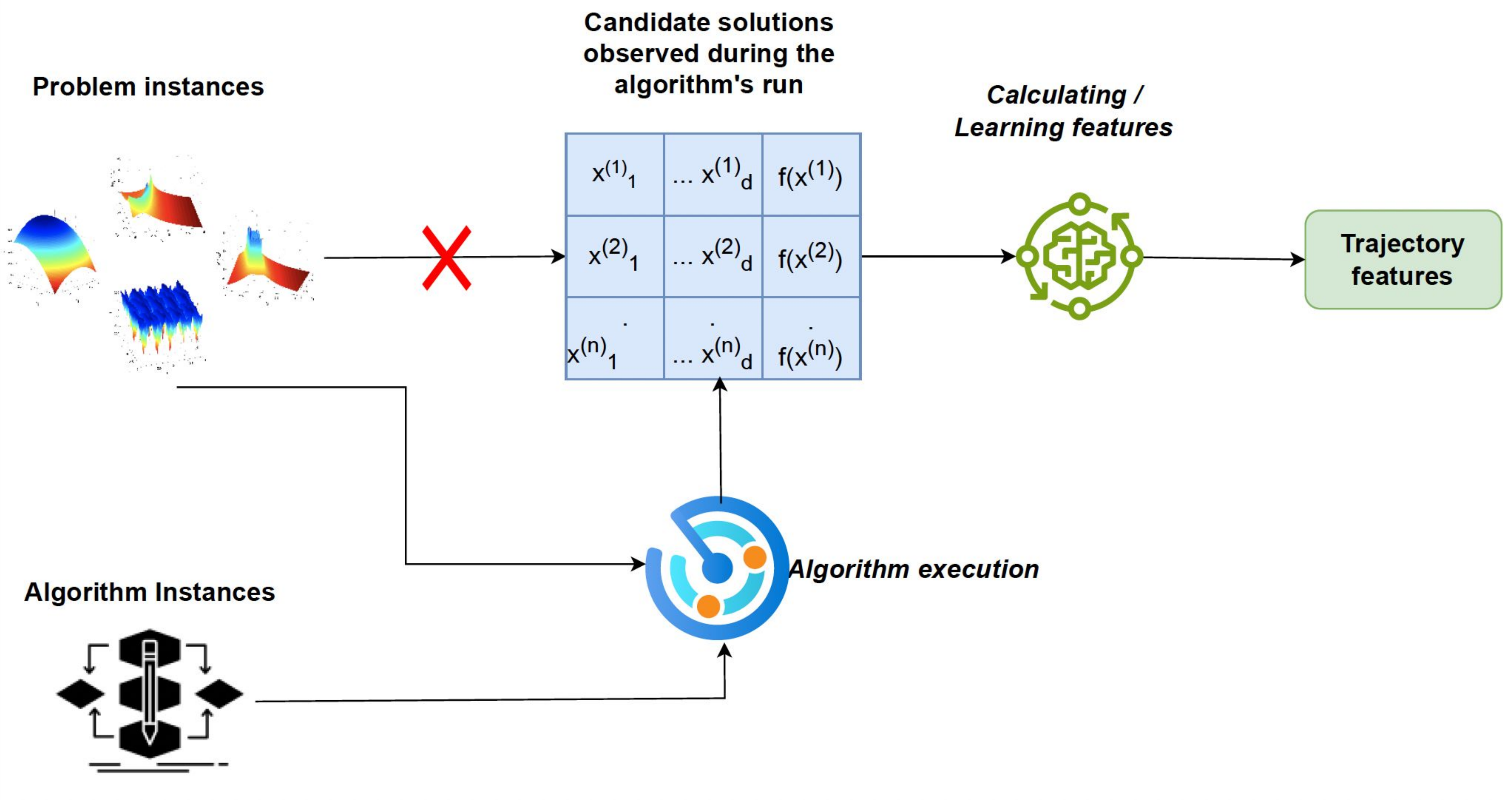
Algorithm Features

- Based on source code
- Based on performance (performance2vec)
- Based on Shapley values of performance predictive model
- Via Knowledge Graph





Problem-Algorithm Trajectory Features

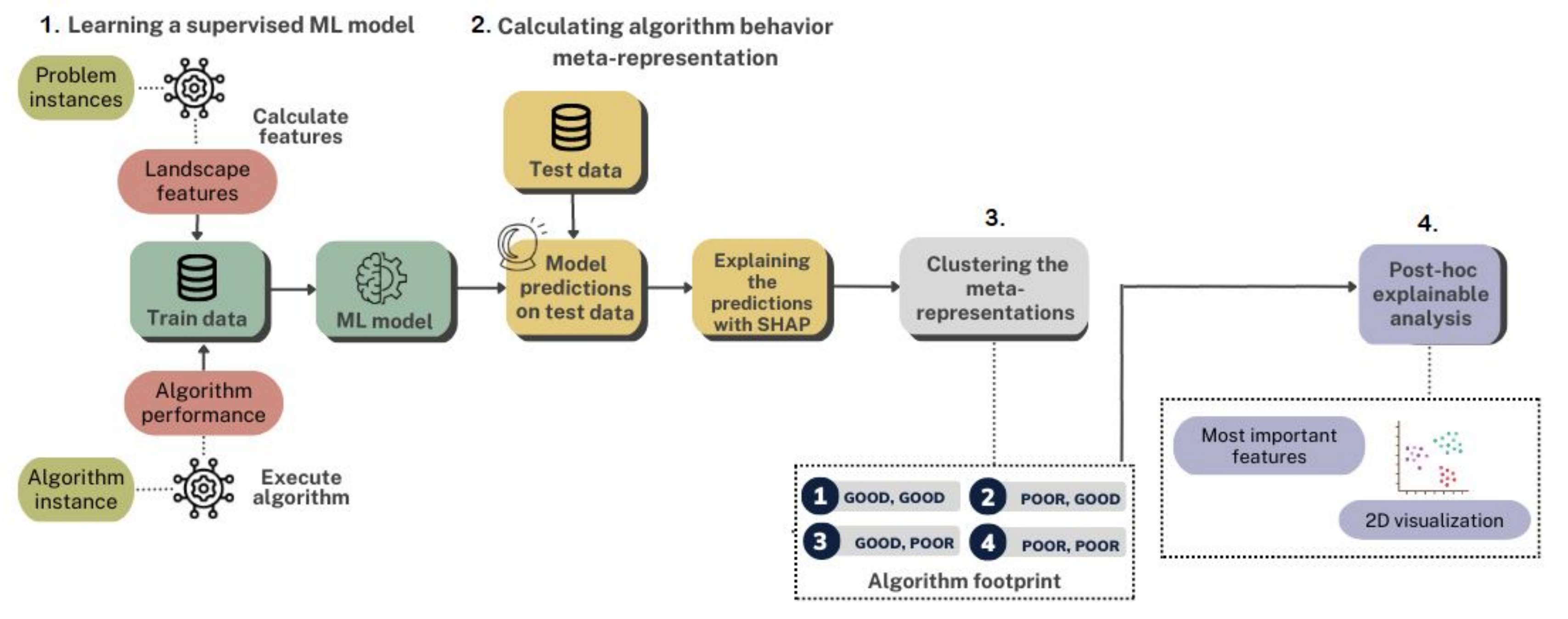


Algorithm Footprint

The term "**algorithm instance footprint**" refers to the regions (i.e., sets of problem instances) where an algorithm instance performs well or poorly, with accompanying **identification of the problem landscape properties** and <u>**their interactions**</u> that contribute to this performance variation.

Algorithm Footprint

- **Train a supervised ML model** to predict algorithm performance.
- **Use SHAP to explain** each feature's contribution to the prediction.
- Create **meta-representations** embedding landscape properties and algorithm performance.
- **Cluster meta-representations** to identify performance regions.
- **Analyze cluster properties** to identify factors affecting algorithm performance.



ML results

Problem Instances	Landscape Features
• The BBOB benchmark suite, of 120 noise-free, single-objective optimization problem instances; • in 10 dimensions; • 5 instances per problem.	• The problem instances are described using 64 features derived by Exploratory Landscape Analysis (ELA); • feature selection.
Algorithm Performance	**ML Algorithms**
• The performance of 3 randomly selected Differential Evolution (DE) configurations is predicted; • 30 runs; • solution precision after a fixed budget of function evaluations; • log10 transformation on the target.	• Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbours (KNN) are used as predictive models; • 5-fold cross-validation; • report Mean Absolute Error (MAE) on the test set.

ML model performance:

- 5-fold cross-validation
- Mean Absolute Error (MAE) and the R2 score

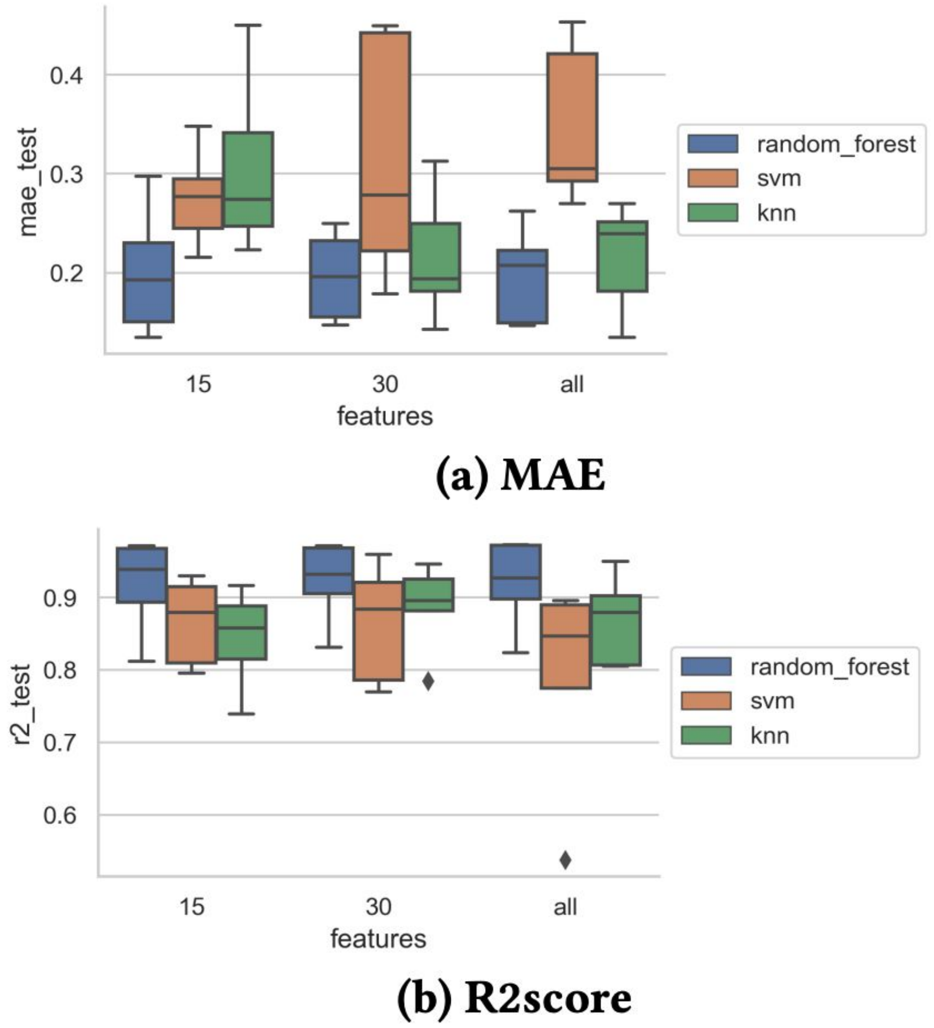


Figure: Performance of the algorithm performance models when predicting the performance of DE1, over the test portion of the 5 folds: (a) MAE, (b) R2 score, for different feature portfolios.

DE Footprint

Four deterministic clusters:

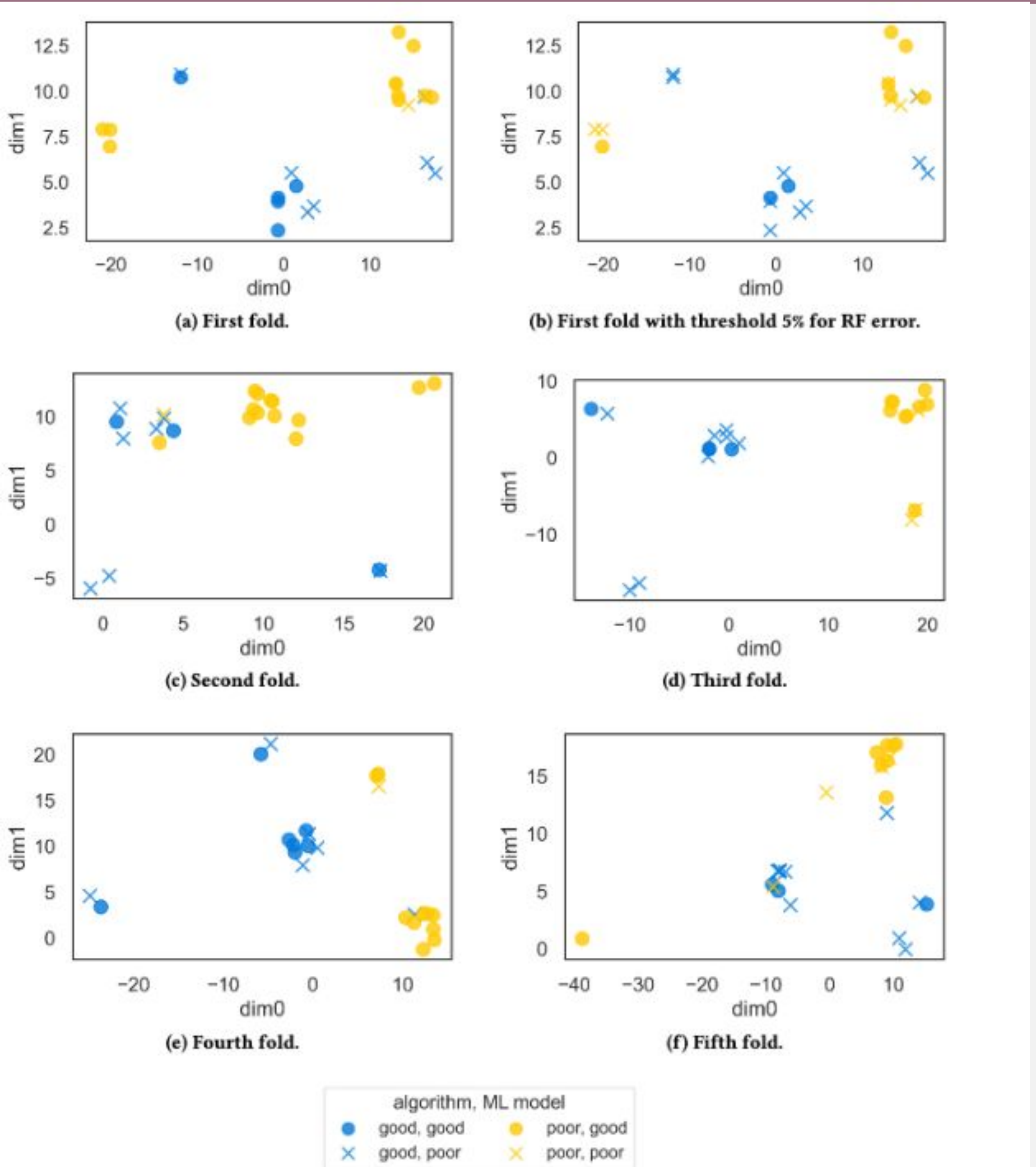
- poor or good *i)* algorithm performance and *ii)* prediction error.

The clustering is done based on apriori set thresholds:

- *t* = the median algorithm performance

- In the (*, *good*) scenario the ML model successfully detects the algorithm behavior.
- In the case of (*, *poor*) the ML model cannot predict the algorithm performance (*good*) or (*poor*), within the specified error.

- There is a distinction between good vs. poor algorithm instance performance (i.e., placing (*good*, *good*) to (*good*, *poor*) problem instances together vs. (*poor*, *good*) to (*poor*, *poor*) together.
- The second dimension, which is the ML model performance, only guarantees confidence in providing further explanations for problem instances that are predicted in the tolerance error.



(a) First fold.

(b) First fold with threshold 5% for RF error.

(c) Second fold.

(d) Third fold.

(e) Fourth fold.

(f) Fifth fold.

DE footprint

model	fold number	(good, good)	(good, poor)	(poor, good)	(poor, poor)
RF	1	16, 19, 20, 21, 22	1, 2, 5, 14, 17, 18, 23	3, 4, 6, 7, 8, 9, 10, 11, 12, 15, 24	13
KNN	1	2, 16, 18, 19, 20, 21, 23	1, 5, 14, 17, 22	4, 6, 7, 8, 9, 10, 11, 12, 15, 24	3, 13
SVM	1	16, 19, 20, 21, 22	1, 2, 5, 14, 17, 18, 23	3, 4, 6, 7, 8, 9, 10, 11, 12, 24	13, 15
RF	2	19, 20, 21	1, 2, 5, 14, 17, 22, 23	3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 24	18
KNN	2	5, 17, 19, 20, 23	1, 2, 14, 21, 22	3, 4, 6, 7, 8, 9, 10, 11, 12, 16, 24	13, 15, 18
SVM	2	19, 21	1, 2, 5, 14, 17, 20, 22, 23	6, 8, 9, 12, 15, 16, 18, 24	3, 4, 7, 10, 11, 13
RF	3	19, 20, 21, 22	1, 2, 5, 14, 16, 17, 18, 23	3, 4, 6, 8, 9, 12, 13, 15, 24	7, 10, 11
KNN	3	1, 16, 18, 19, 20, 21, 22, 23	2, 5, 14, 17	6, 8, 9, 10, 12, 13, 24	3, 4, 7, 11, 15
SVM	3	6, 22	1, 2, 5, 14, 17, 18, 19, 20, 21, 23	9, 12, 13, 24	3, 4, 6, 7, 8, 10, 11, 15
RF	4	5, 16, 18, 19, 20, 21, 22	1, 2, 7, 14, 17, 23	3, 4, 6, 8, 9, 10, 12, 13, 15, 24	11
KNN	4	1, 7, 16, 19, 20, 21, 22, 23	2, 5, 14, 17, 18	4, 6, 8, 9, 10, 12, 24	3, 11, 13, 15
SVM	4	16, 20, 21	1, 2, 5, 7, 14, 17, 18, 19, 22, 23	6, 9, 11, 13, 24	3, 4, 8, 10, 12, 15
RF	5	19, 20, 21	1, 2, 5, 7, 14, 16, 17, 22, 23	6, 8, 9, 11, 12, 13, 15, 24	3, 4, 10, 18
KNN	5	1, 14, 16, 19, 20, 21, 22, 23	2, 5, 7, 17	4, 6, 8, 9, 11, 12, 15, 24	3, 10, 13, 18
SVM	5	5, 16, 19, 21, 22	1, 2, 7, 14, 17, 20, 23	3, 8, 9, 10, 11, 12, 18, 24	4, 6, 13, 15

- **DE1 has stable performance on the 19, 20, and 21 BBOB problem classes.** No matter the different transformations (e.g., shifting, scaling) that are applied the algorithm instance is able to find a solution with the specified target.
- For **the 6th, 8th, 9th, 12th, 15th, and 24th BBOB problem classes,** the algorithm instance is not able to solve them within the specified target.
- The problem instances of **the 7th and the 18th problem classes** are distributed across all of the clusters, thus the algorithm instance does not have stable performance on them.

Post-hoc analysis

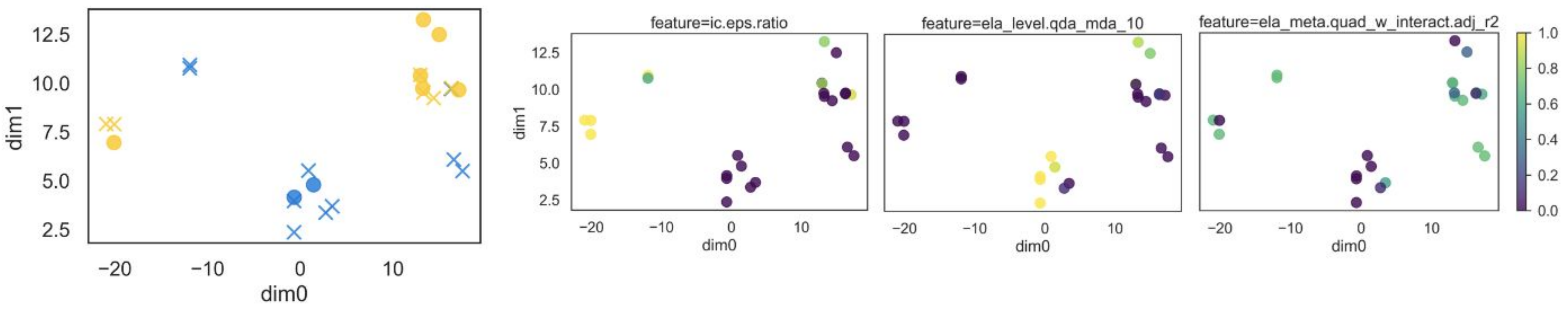
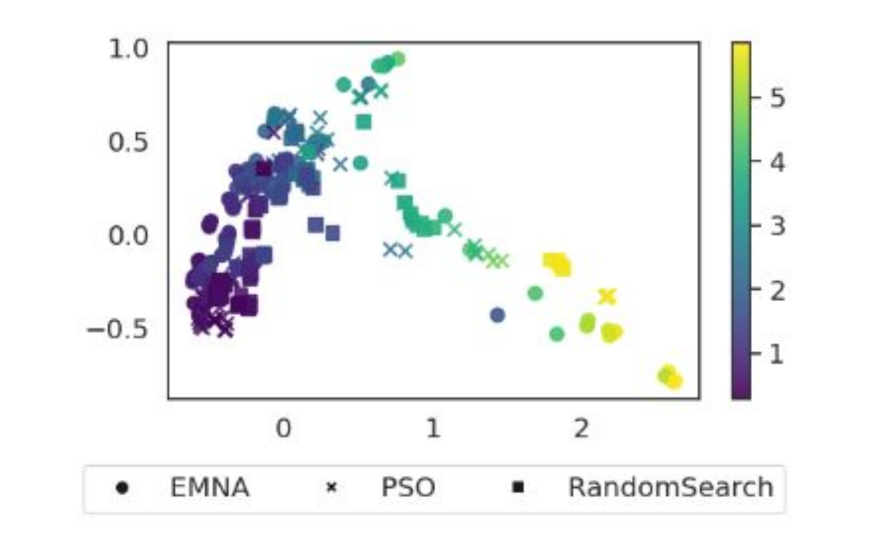


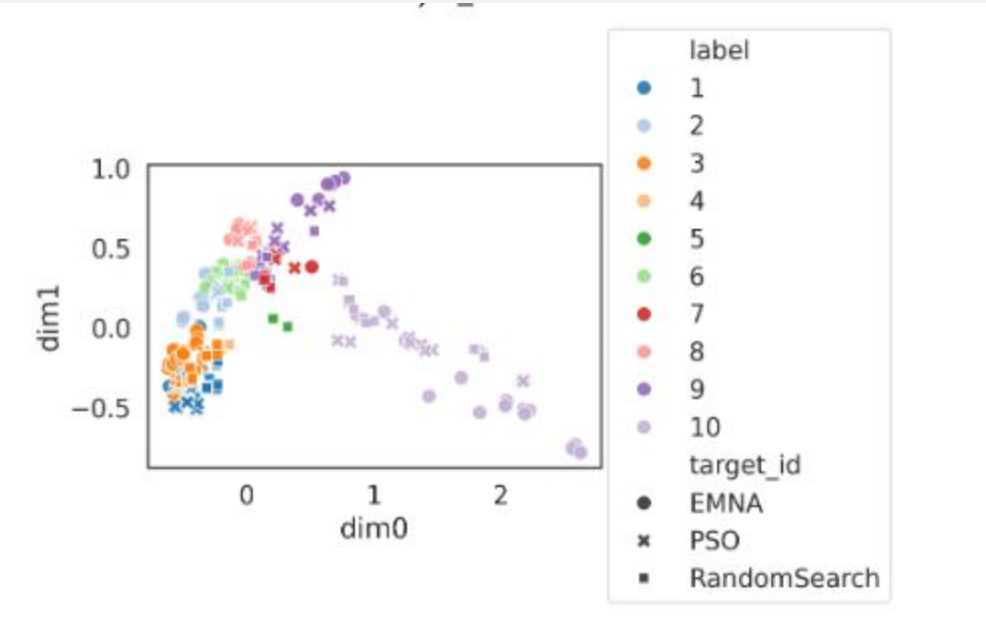
Figure: The distribution of ELA features across the algorithm instance footprintt.

Benchmarking Algorithm Footprint

- Use a multi-target regression model for automated algorithm performance prediction
- Use case:
  - Three algorithms
    - Particle Swarm Optimization (PSO)
    - Random Search (RS)
    - Estimation of Multivariate Normal Algorithm (EMNA)

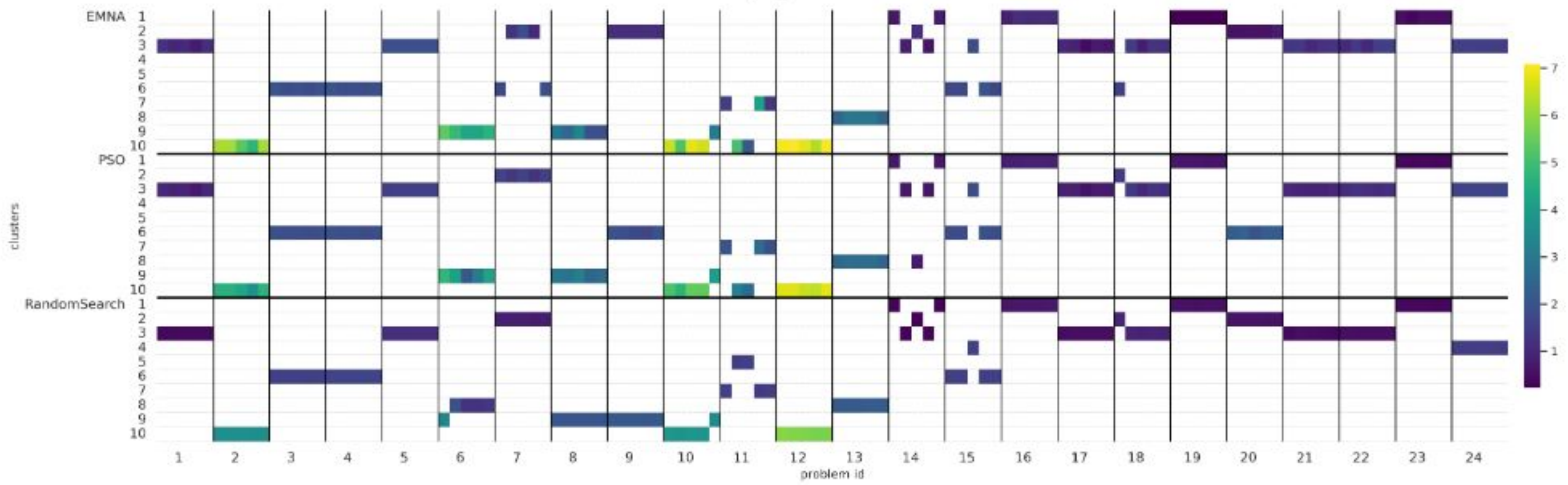
Algorithm	Model	MAE	R2
EMNA	mean	1.338104	-0.000754
PSO		1.096115	-0.000665
RS		0.922459	-0.000302
EMNA	random_forest	**0.149445**	**0.973351**
PSO		**0.123431**	**0.973023**
RS		**0.061197**	**0.980432**
EMNA	neural_network	0.385019	0.912184
PSO		0.263178	0.928873
RS		0.330922	0.844444
EMNA	multitask_elastic_net	0.432810	0.891677
PSO		0.276216	0.932986
RS		0.346470	0.860945





b) n_clusters=10

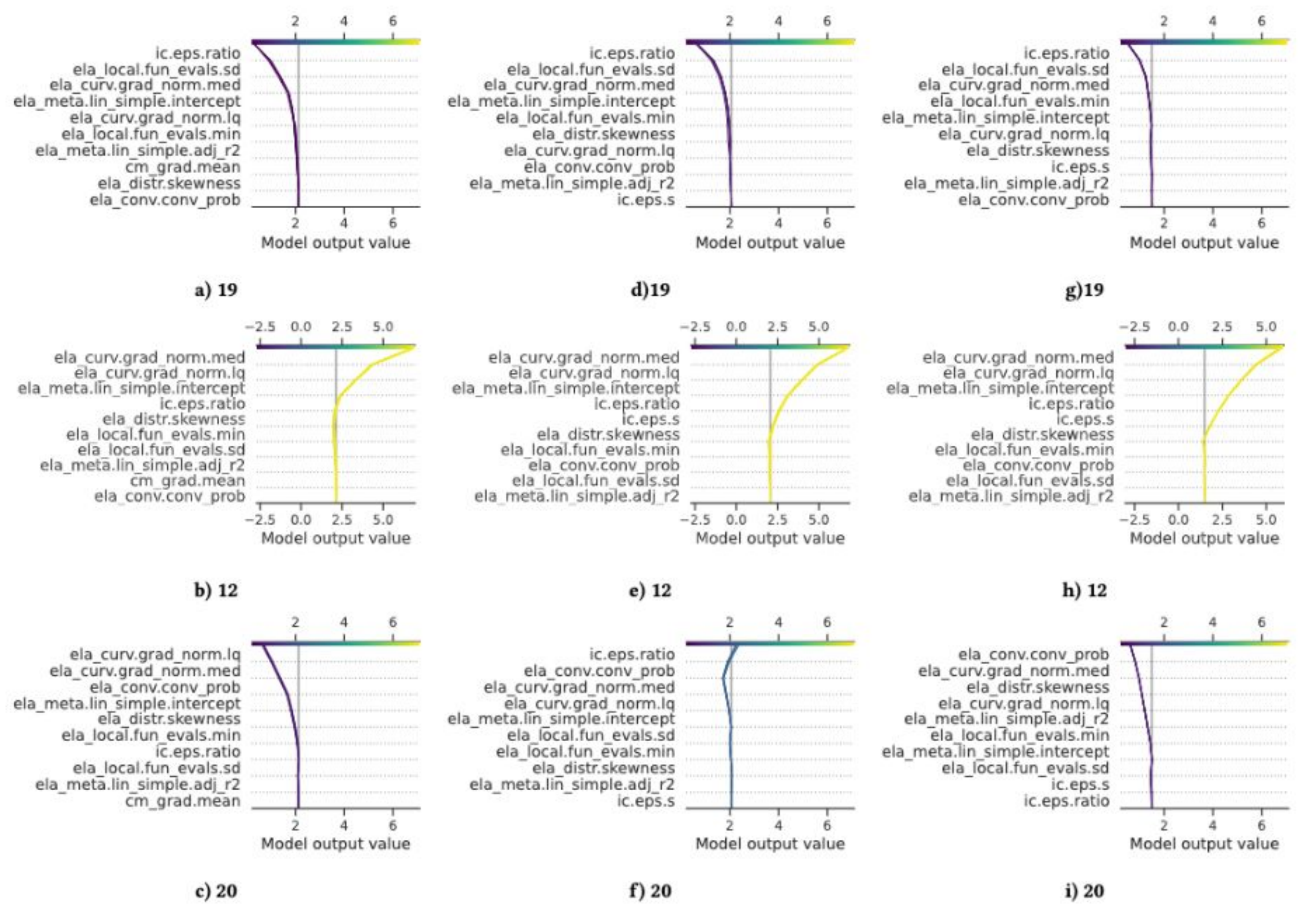
Benchmarking Algorithm Footprint



b) n_clusters=10

Coverage matrix of the distribution of the meta-representations in the clusters, with the subfigures illustrating the results for a) 3, and b) 10 clusters, respectively.

Post-hoc Analysis



Visualization of the contribution patterns of the 10 most important features in the algorithm performance prediction, with subfigures illustrating the results for problem a-c) EMNA, d-f) PSO, and g-i) Random Search for the corresponding problem as indicated in the subfigure caption.

Take Home Messages

- Use approaches to **understand what are the strengths and weaknesses** of a new algorithm instead of looking into its average performance!

Team

Gjorgjina Cenikj
Ph.D. Student

Ana Nikolikj
Ph.D. Student

Ana Kostovska
Ph.D. Candidate

Jan Drole
Master Student

Ana Gjorgjevikj, Ph.D.
Postdoc

Tome Eftimov, Ph.D.
Assistant Professor

Huge thanks to …

Anja Jankovic, Ph.D.
Postdoctoral researcher
RWTH Aachen University
Germany

Diederick Vermetten
Ph.D. Candidate
Leiden Institute of Advanced Computer Science (LIACS)
The Netherlands

Carola Doerr, Ph.D.
CNRS Research director
Sorbonne Université
France

Funding & Support

Slovenian Research Agency

- **Slovenian Research and Innovation Agency**
  - Mr-BEC (2019 - 2021)
  - RESPONSE (2022 - 2024)
  - AutoOpt (2022 - 2025)
  - AI for Science (2024 - 2027)
- **Horizon Europe**
  - AutoLearn-SI (2025 - 2030)
  - MSCA COFUND - SQUASH (2025 - 2028)

Jožef Stefan Institute

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA SLOVENIJA LJUBLJANA 2004

Stanford University

UCSF University of California San Francisco

Faculty of Computer Science and Engineering – University Sts. Cyril and Methodius - Skopje

AutoLearn-SI

- HE ERA-Chair
- 2.5 million EUR
- Starting date: 01.03.2025
- Scope: Automated Machine Learning and Optimization techniques
- **3 Ph.D. positions** starting at 01.10.2025
- **2 Postdoc Positions** starting at 01.07.2026

- Hosting **Asst. Prof. Eva Tuba** as an ERA Chair
- Trinity University, San Antonio, TX
- 2024 ACM W Rising Star Award

MSCA COFUND - SQUASH

- **One Postdoc Position** starting at 01.10.2025
- Scope: Landscape analysis of quantum optimization algorithms

Selection of representative learning data

Deep Statistical Comparison of three algorithms across different benchmark suites

- Three algorithms selected from the Nevergrad (Facebook) framework
- Statistical comparison on already established benchmark suites
  - BBOB/COCO, CEC 2013, CEC 2014, CEC 2015, CEC 2017

	BBOB		CEC 2013		CEC 2014		CEC 2015		CEC 2017		All	
	DE	RSPSO	DE	RSPSO	DE	RSPSO	DE	RSPSO	DE	RSPSO	DE	RSPSO
RSPSO	0.00/0		0.95/1		0.29/1		0.02/0		0.00/0		0.00/0	
CMA	0.33/1	0.00/0	0.47/1	0.65/1	0.75/1	0.07/1	0.04/0	0.98/1	0.97/1	0.00/0	0.06/1	0.00/0

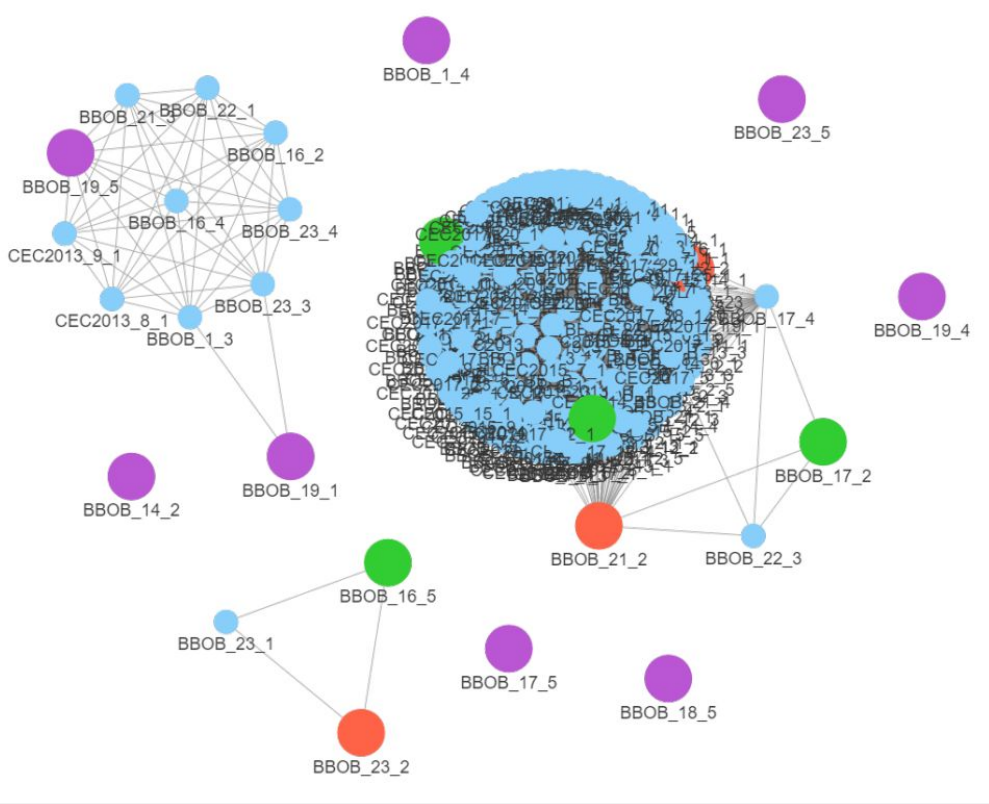
*0 - statistically significant difference in performance found
*1 - no statistically significant difference in performance found

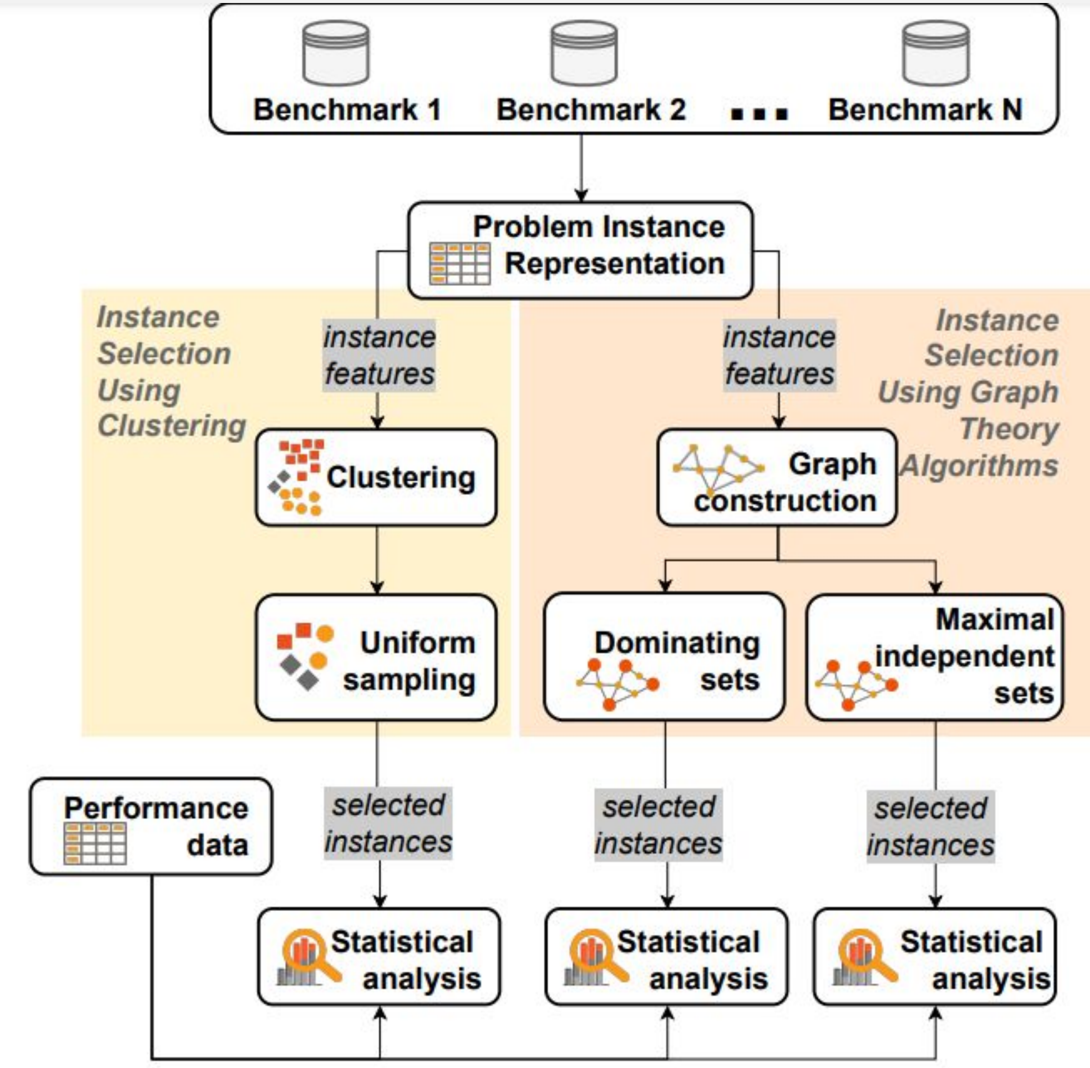
Problem features

- Exploratory Landscape Analysis (ELA)
  - 64 features

- BBOB/COCO *(24 problems x 5 instances)*,
- CEC 2013 *(28 problems)*,
- CEC 2014 *(30 problems)*,
- CEC 2015 *(15 problems)*,
- CEC 2017 *(29 problems)*
  - 10D

SELECTOR - Selection of diverse benchmark problem instances





Cenikj, G., Lang, R. D., Engelbrecht, A. P., Doerr, C., Korošec, P., & Eftimov, T. (2022, July). Selector: selecting a representative benchmark suite for reproducible statistical comparison. In *Proceedings of The Genetic and Evolutionary Computation Conference* (pp. 620-629).

Comparison using the new selected benchmark suites via clustering

Results of the Friedman test and the Nemenyi post-hoc test for the statistical comparison of the three algorithms using the benchmark suites selected from the 21 and the 26 clusters, respectively.

	21 clusters		26 clusters	
	DE	RSPSO	DE	RSPSO
RSPSO	0.24/1		0.28/1	
CMA	0.48/1	0.02/0	0.51/1	0.02/0

Results of the Friedman test and the Nemenyi post-hoc test for the statistical comparison of the three algorithms using the benchmark suites selected by using different percentage of representatives for the larger clusters.

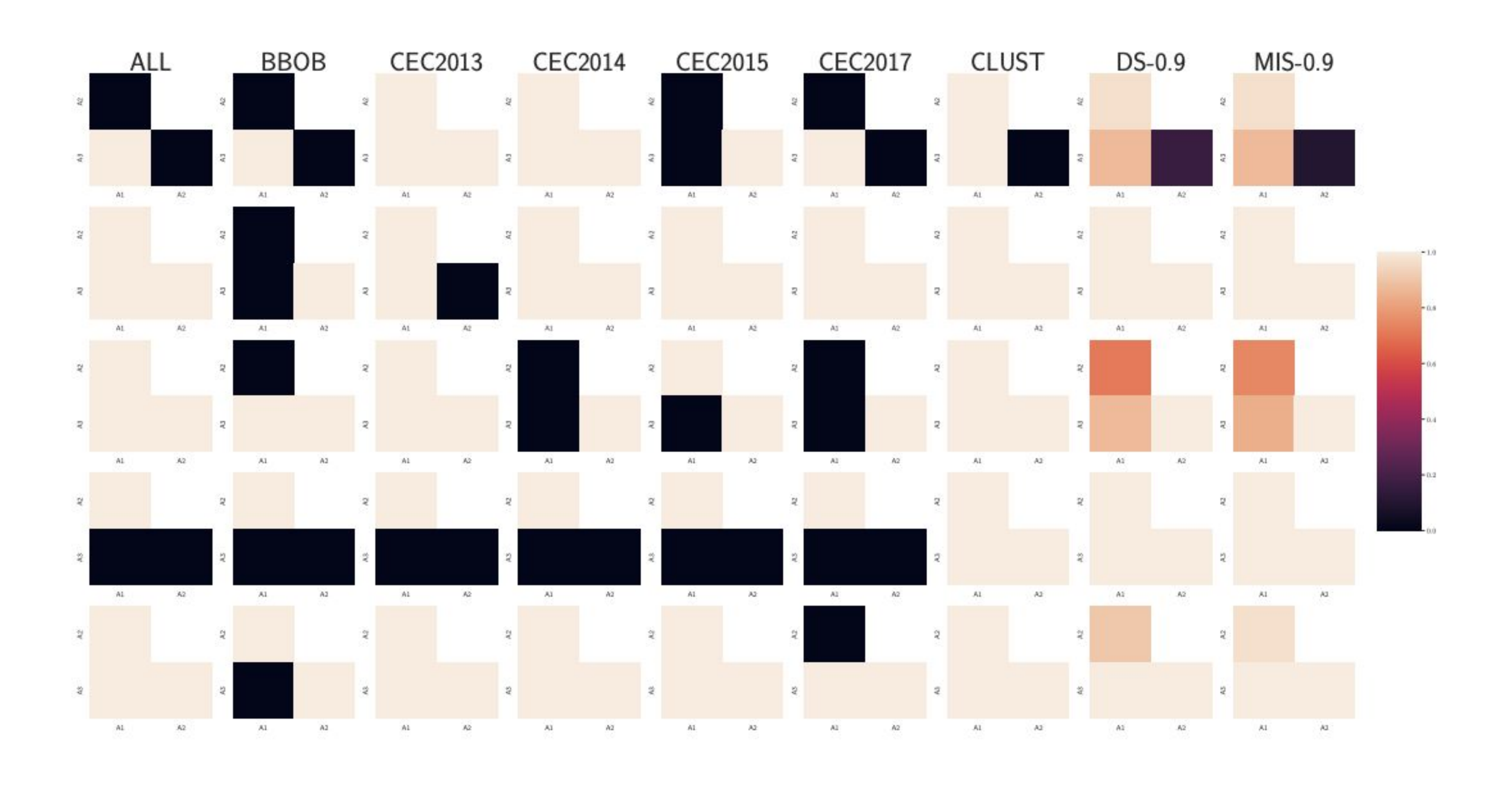
	12.5% repres.		25% repres.	
	DE	RSPSO	DE	RSPSO
RSPSO	15.00		15.00	
CMA	14.00	0.00	14.00	0.00

Comparison using the new selected benchmark suites via graph theory

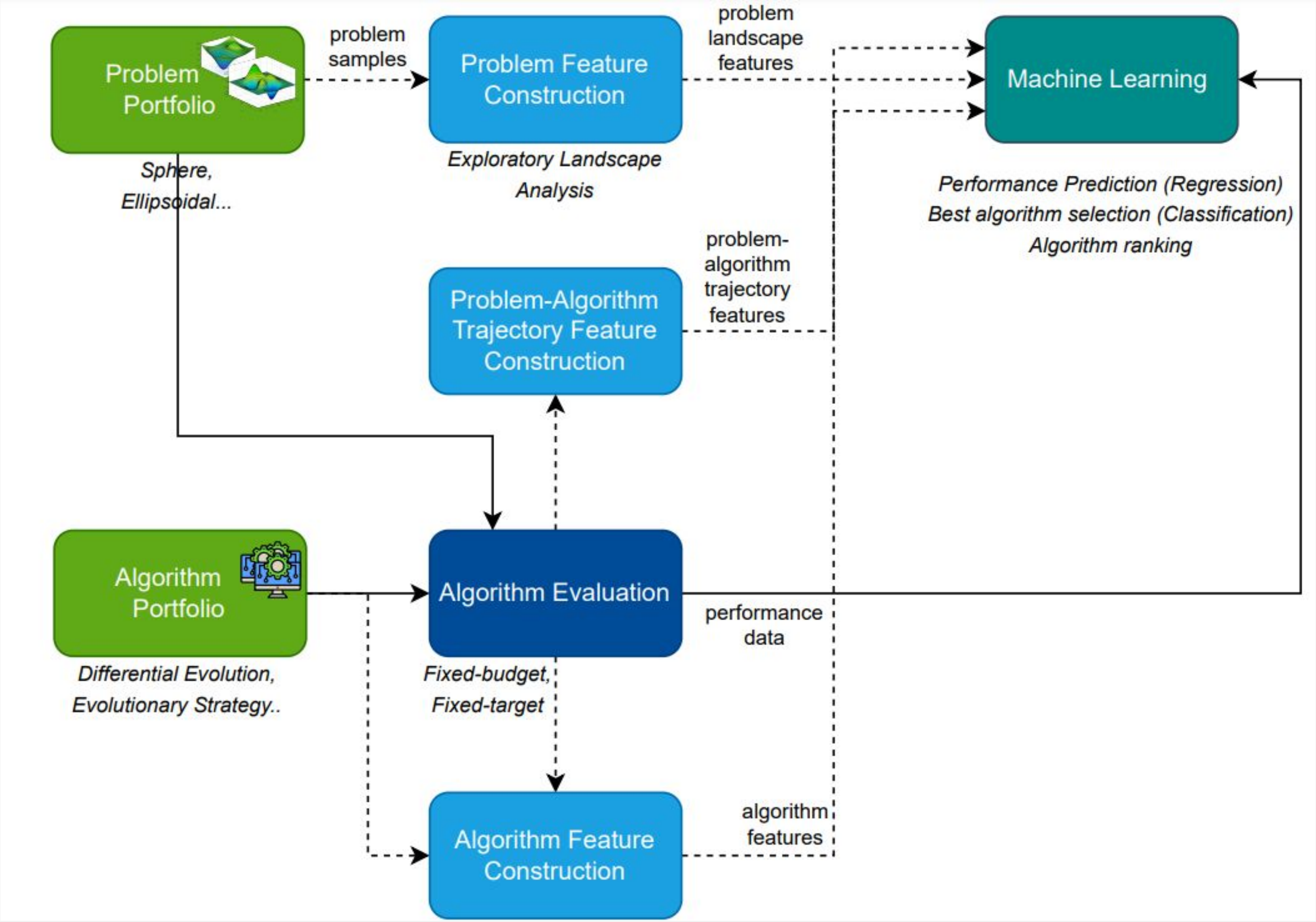
Results of the Friedman test and the Nemenyi post-hoc test for the statistical comparison of the three algorithms using the benchmark suites selected by the MIS and DS graph algorithms, for different cosine similarity measures. The numbers indicate the number of times in which no statistical significance was identified between the performance of a pair of algorithms, out of 30 independent executions of the statistical analysis, on 30 different subsets of instances produced by 30 runs of the algorithms.

	DS 0.9		DS 0.95		DS 0.97	
	DE	RSPSO	DE	RSPSO	DE	RSPSO
RSPSO	30.00		30.00		30.00	
CMA	27.00	5.00	26.00	3.00	22.00	0.00
	MIS 0.9		MIS 0.95		MIS 0.97	
	DE	RSPSO	DE	RSPSO	DE	RSPSO
RSPSO	30.00		30.00		30.00	
CMA	27.00	3.00	30.00	0.00	24.00	0.00

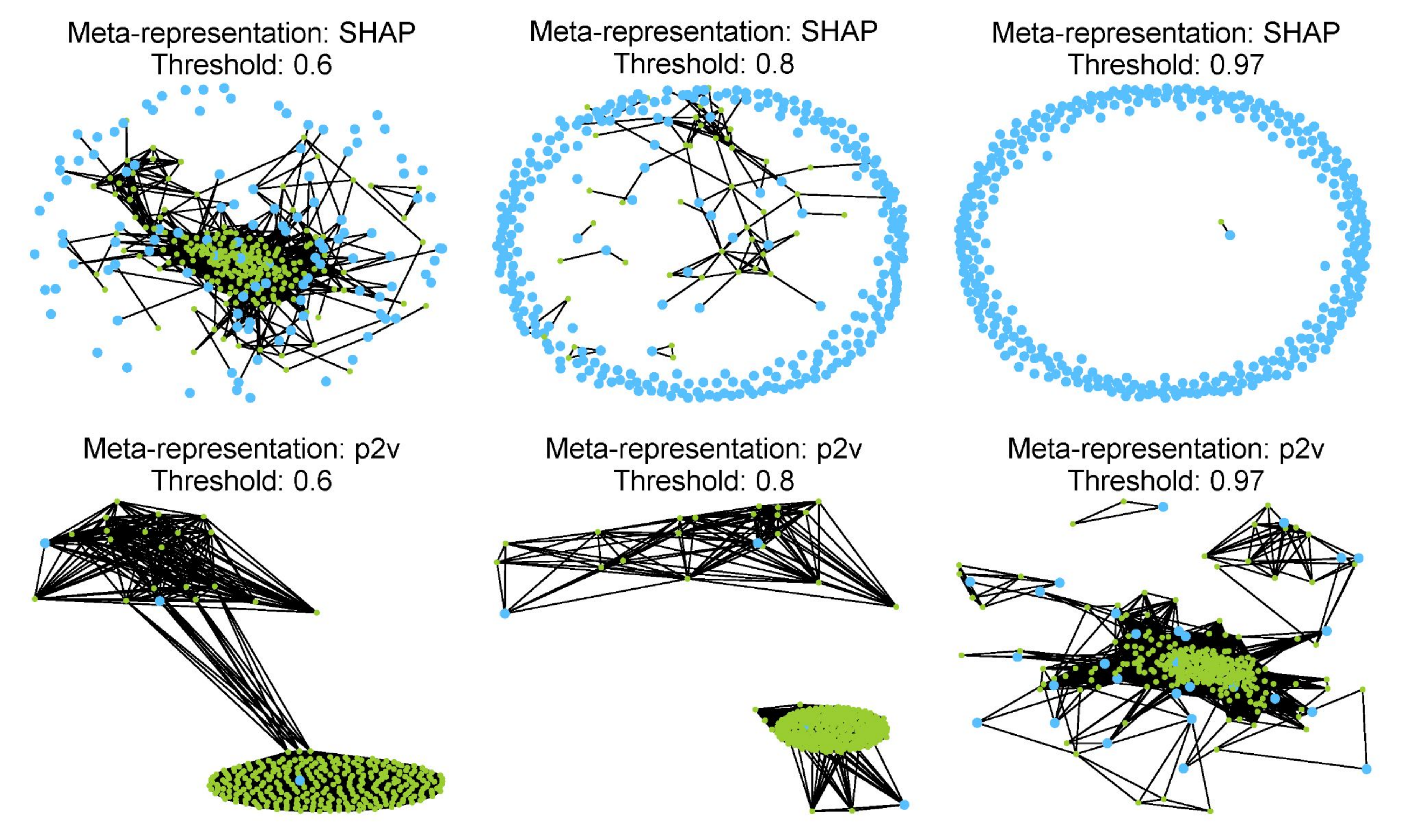
Generalization of the SELECTOR



Automated algorithm selection

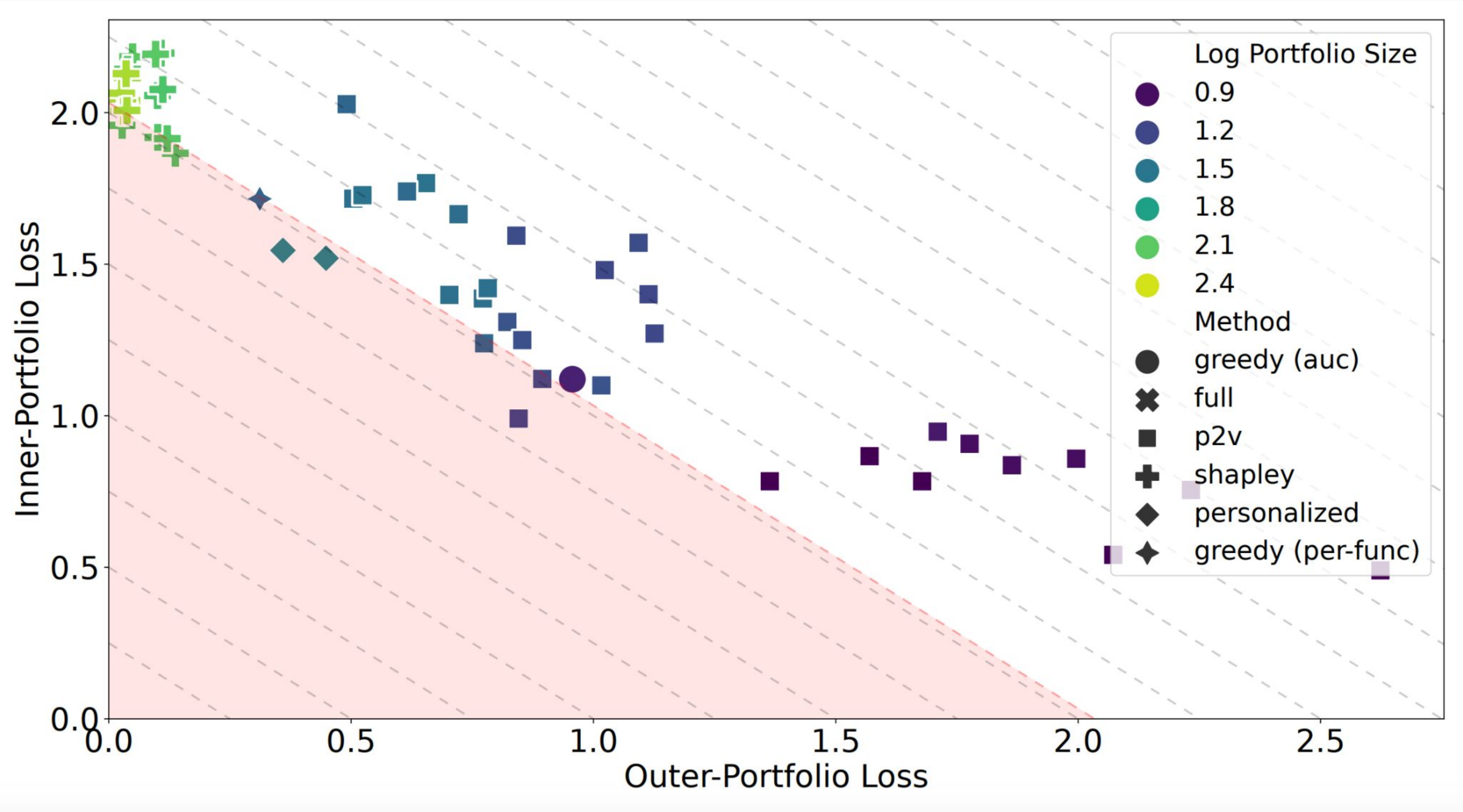


Selection of complementary algorithm portfolio



Kostovska, A., Cenikj, G., Vermetten, D., Jankovic, A., Nikolikj, A., Skvorc, U., ... & Eftimov, T. (2023, December). PS-AAS: Portfolio Selection for Automated Algorithm Selection in Black-Box Optimization. In International Conference on Automated Machine Learning (pp. 11-1). PMLR.

Selection of complementary algorithm portfolio



x-axis: the best possible loss of the portfolio = the difference between the portfolio's VBS and the VBS of the full set of 324 algorithms.
y-axis: the loss of the AS = the difference in performance between the algorithm it selects and the VBS of the portfolio it can choose from.