Genetic and Evolutionary Computation Conference
(GECCO '24)
July 14--18, 2024, Melbourne, Australia

Minimum variance threshold for ϵ-lexicase selection

Guilherme Seidyo Imai Aldeia
Federal University of ABC
Santo Andre, São Paulo, Brazil
guilherme.aldeia@ufabc.edu.br

Fabrício Olivetti de França
Federal University of ABC
Santo Andre, São Paulo, Brazil
folivetti@ufabc.edu.br

William G. La Cava
Boston Children's Hospital
Harvard Medical School
Boston, Massachusetts, USA
william.lacava@childrens.harvard.edu

UFABC

Boston Children's Hospital
Until every child is well™

AVA Lab

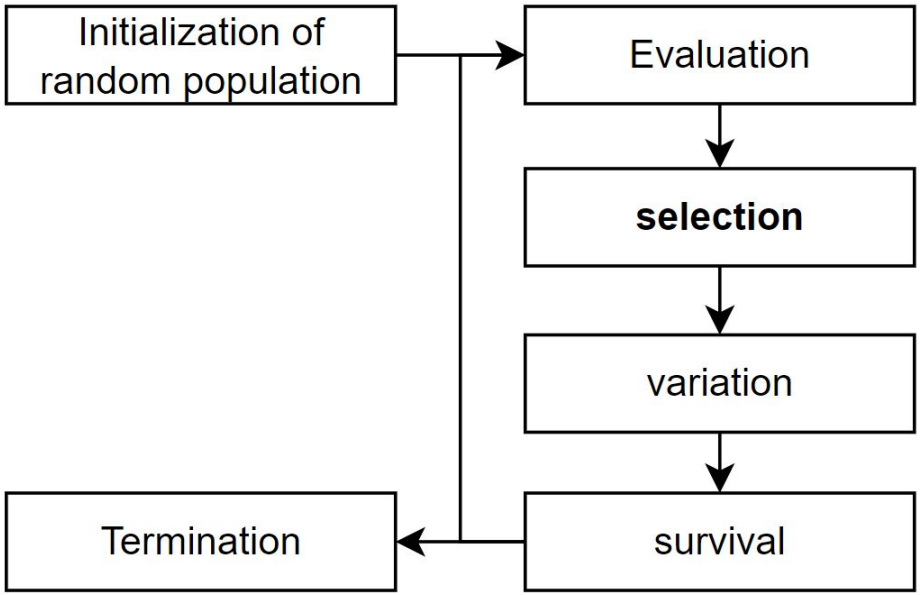
VE RI TAS

chip Computational Health Informatics Program

CMCC
Centro de Matemática, Computação e Cognição

Evolutionary algorithms

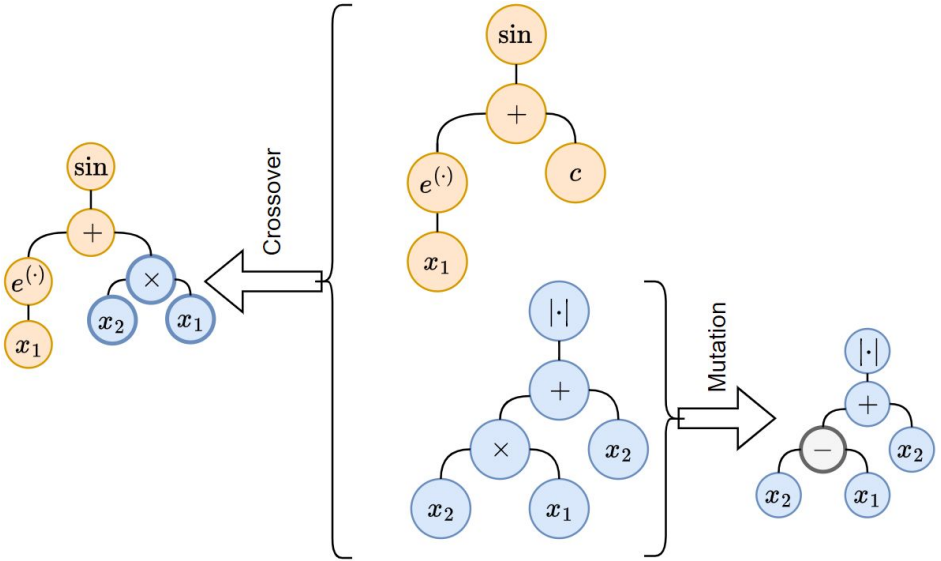
One important step of evolutionary algorithms is the **parent selection**:



Evolutionary algorithms

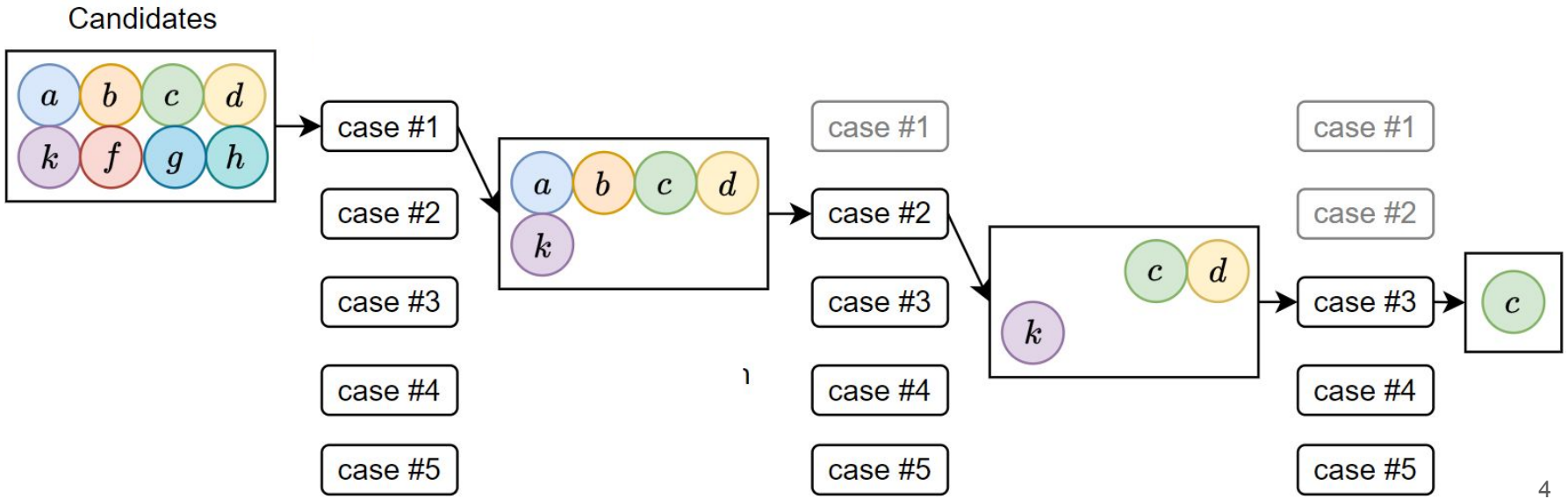
One important step of evolutionary algorithms is the **parent selection**:

The parents are used to **generate** the offspring, and different methods can shift the population **distribution**, affecting **convergence** and final **fitness**.



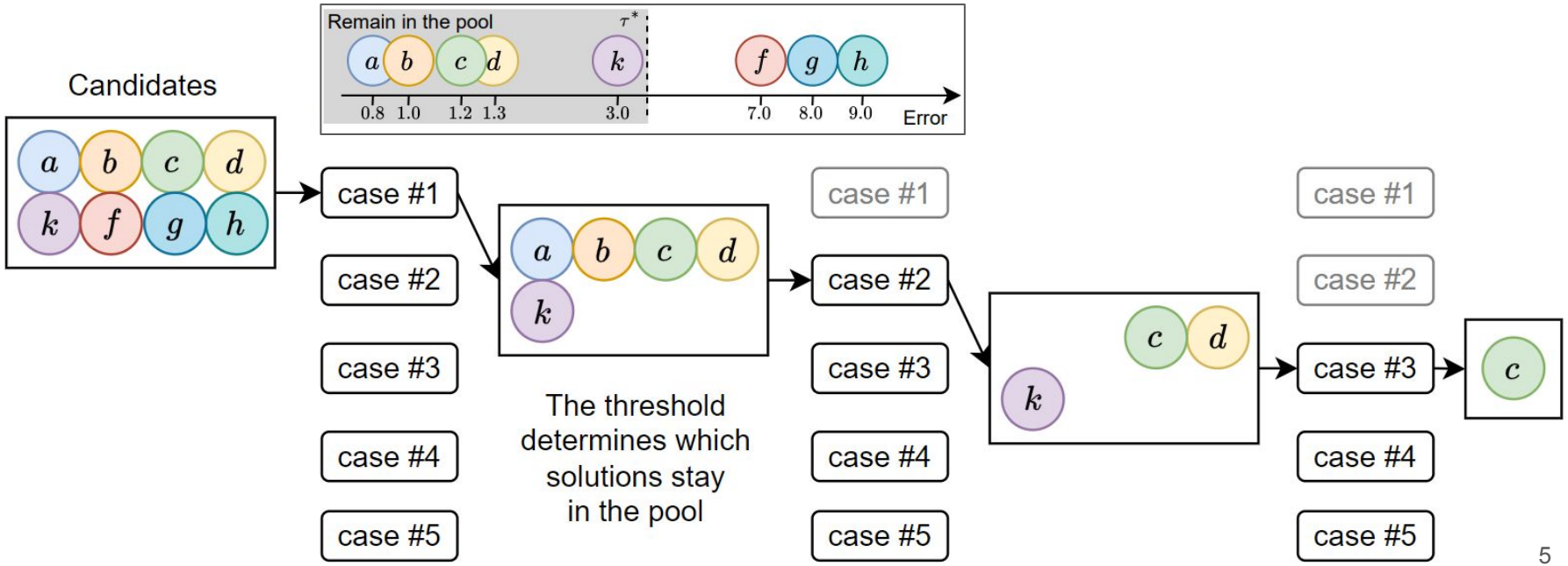
How (ϵ-)lexicase works

Lexicase (for **classification**) uses test cases to shrink a pool of parent candidates. The threshold (ϵ) adapts it for real values (**regression**).



How (ϵ-)lexicase works

Lexicase (for **classification**) uses test cases to shrink a pool of parent candidates. The threshold (ϵ) adapts it for real values (**regression**).



Why is it good?

The selection requires that parents perform well at different data samples.

Aggregating the fitness chose parents that performs well on average while missing individuals that perform well on difficult subsets of the problem.

Can we do it differently?

Our hypothesis is we can get a better performance if we change how we split the pool in ϵ-lexicase by doing like decision trees do.

The idea is to replace **median absolute deviation** (MAD) with the **minimum variance threshold** (MVT). This is connected to information entropy.

Can we do it differently?

MAD can be interpreted as a dispersion measured from the center of the distribution.

$$\epsilon_t = \lambda(\mathbf{e}_t) = \text{median}(|\mathbf{e}_t - \text{median}(\mathbf{e}_t)|)$$

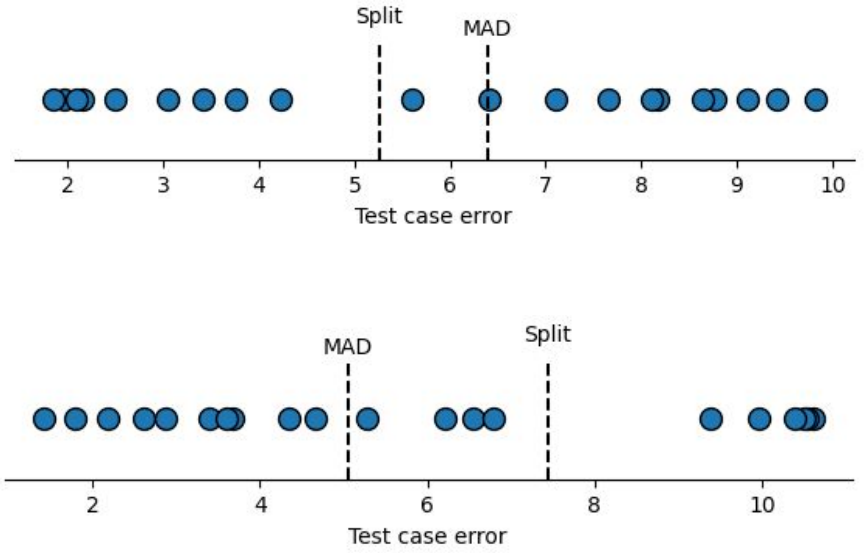
We propose to split the pool in two clusters, minimizing the variance among individuals from the same cluster.

$$\tau^* = \min_{\min(\mathbf{e}_t)<\tau<\max(\mathbf{e}_t)} \left(\frac{\text{Var}(\mathbf{l})}{|\mathbf{l}|} + \frac{\text{Var}(\mathbf{r})}{|\mathbf{r}|} \right)$$

MAD vs split

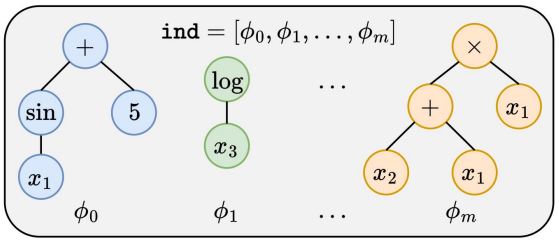
Split is more sensitive to similarities than MAD.

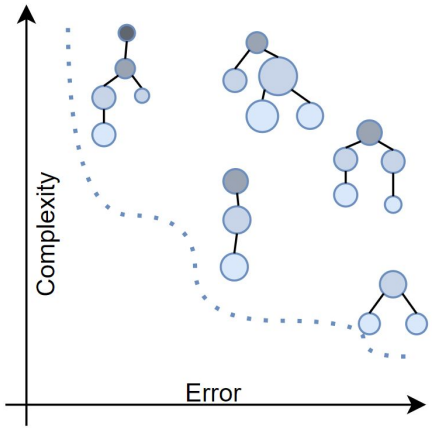
The threshold now is on the biggest gap between the errors, clustering the good-performing and bad-performing individuals.



Experimental design

We used Feature Engineering Automation Tool (FEAT), and changed only the selection method to isolate and test if a different split criteria would change the final results.





Parameter	Value
objectives	["fitness","complexity"]
pop_size	100
gens	100
cross_rate	0.5
ml	Linear ridge regression
max_depth	6
backprop	True
iters	10
validation split	0.25
selection	NSGA2
batch_size	200
functions	[+,−,∗,÷,$(\cdot)^2$,$(\cdot)^3$,$\sqrt{\cdot}$,sin,cos,$e^{(\cdot)}$,log]

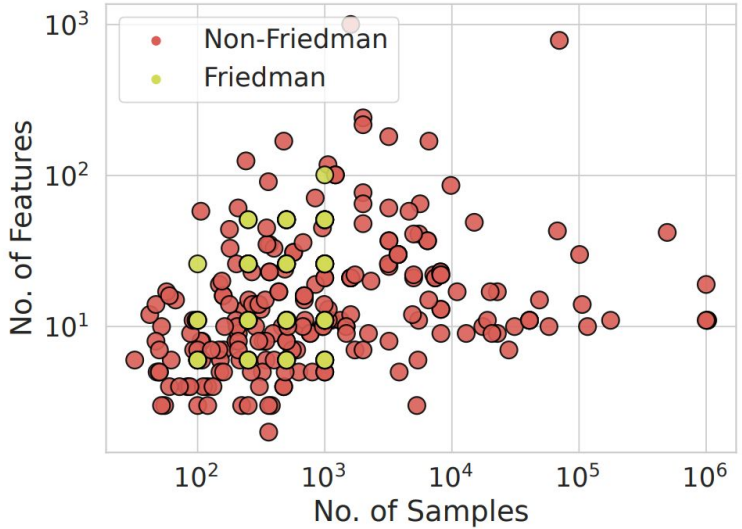
Experimental design

ϵ-lexicase comes in different flavors. We implemented the Static and Dynamic versions for our MVT split threshold (S-split and D-split):

Strategy	**Pool**	**Criteria**
Static	pop ($\mathcal{N}$)	$e_t^* + \epsilon_t$
Semi-dynamic	pop ($\mathcal{N}$)	$\texttt{elite} + \epsilon_t$, where $\epsilon \leftarrow \lambda(\mathbf{e}_t)$ for $t \in \mathcal{T}$
Dynamic	pool ($\mathcal{S}$)	$\texttt{elite} + \epsilon_t$, where $\epsilon \leftarrow \lambda(\mathbf{e}_t(\mathcal{S}))$

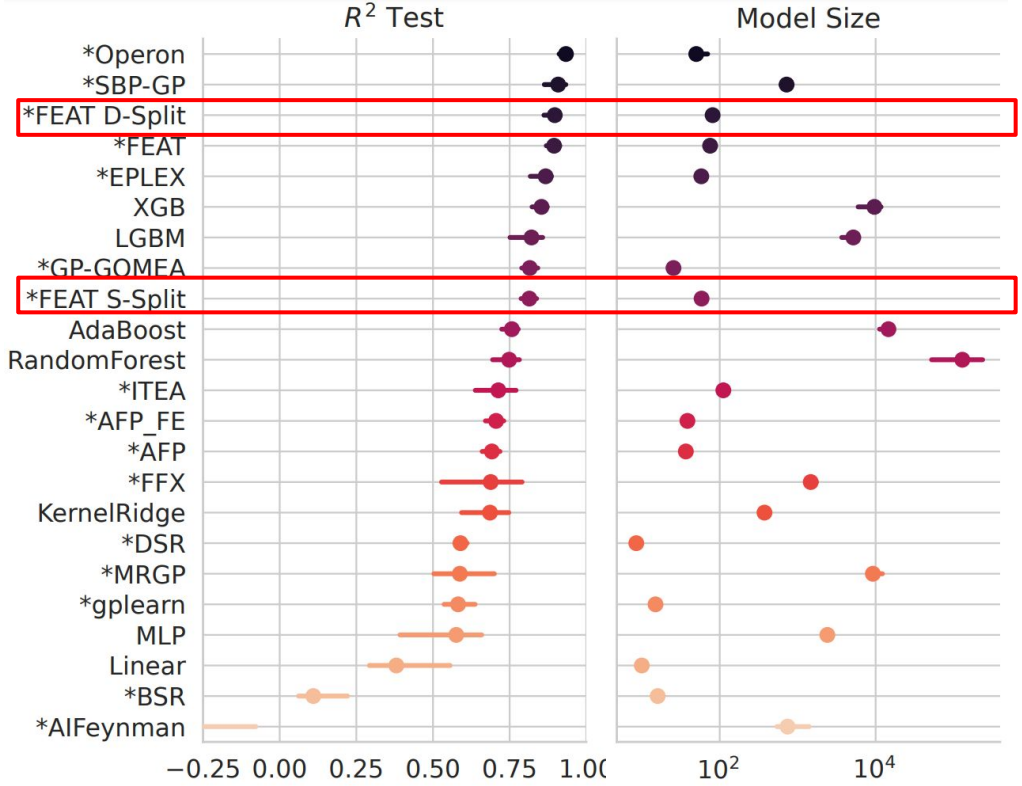
Experimental design

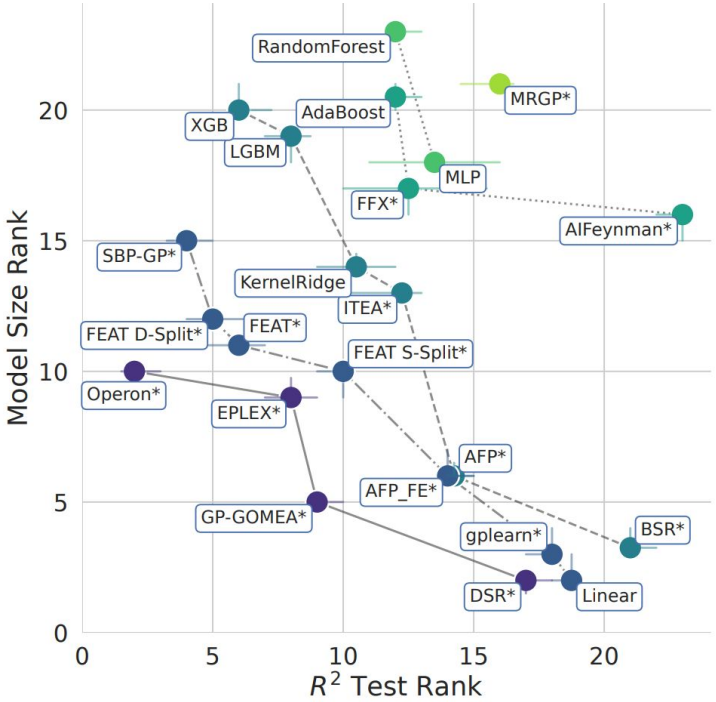
We benchmarked it with the black-box problems from SRBench (https://github.com/cavalab/srbench)



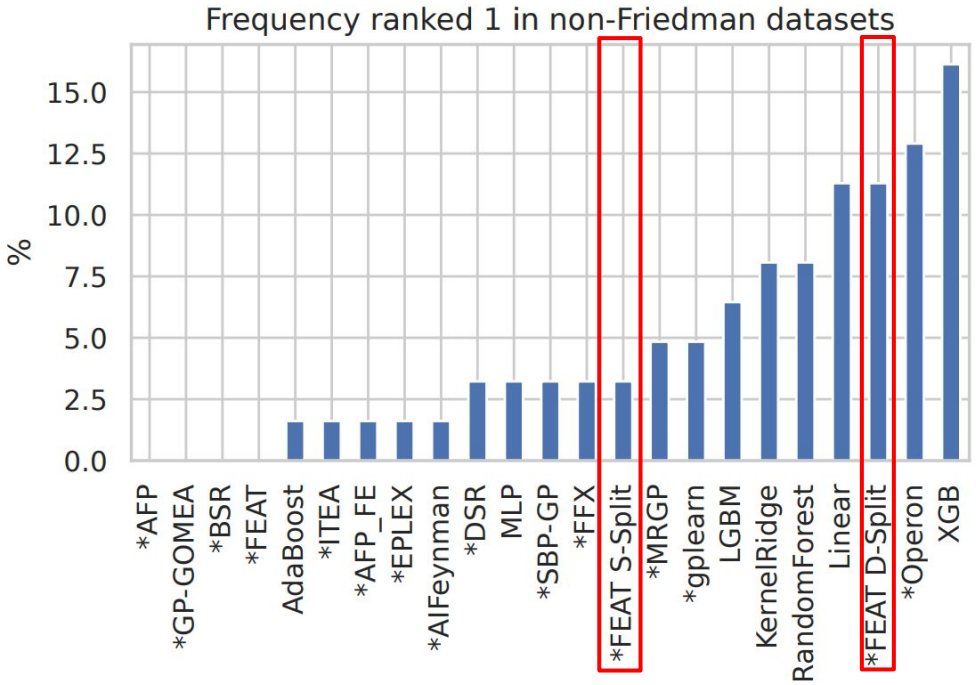
Setting	Black-box Problems
No. of datasets	122
No. of algorithms	21 (14 SR, 7 ML)
No. of trials per dataset	10
Train/test Split	.75/.25
Hyperparameter Tuning	5-fold Halving Grid Search CV
Termination criteria	500k evaluations/train or 48 hours
Levels of target noise	None
Total comparisons	26840
Computing Budget	1.29M core hours

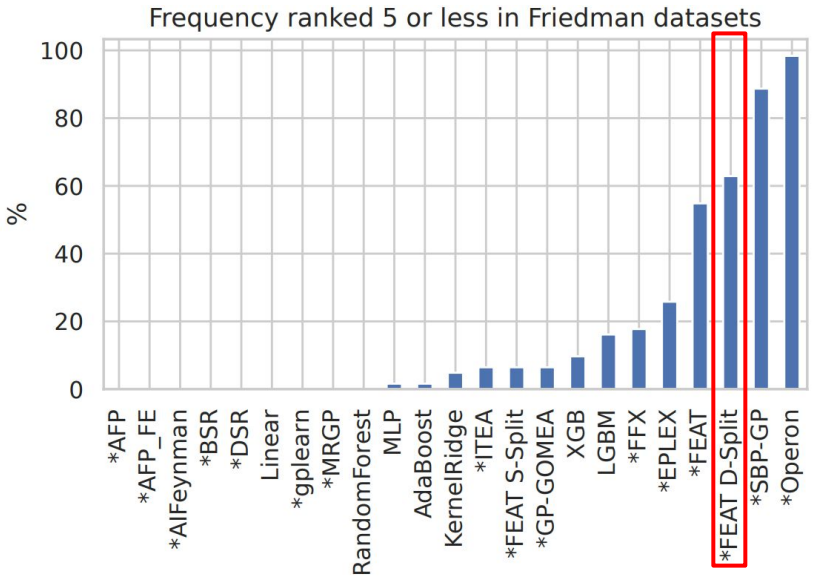
Overall performance on SRBench





Synthetic (Friedman) vs non-synthetic (PMLB)

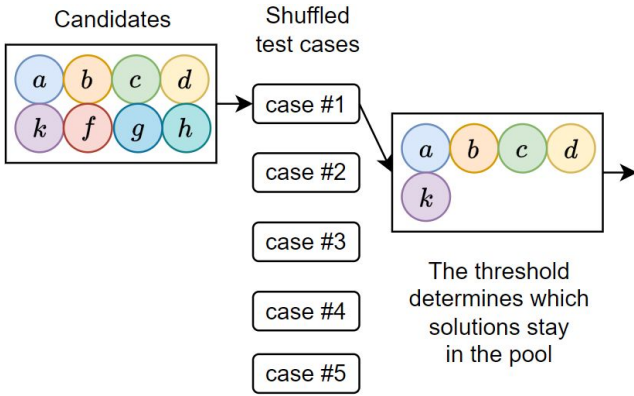




Conclusions

Competitive results with MAD ϵ-lexicase, while improving performance on real-world datasets.

The downside is requiring more test cases to select the parents, leading to higher execution times.

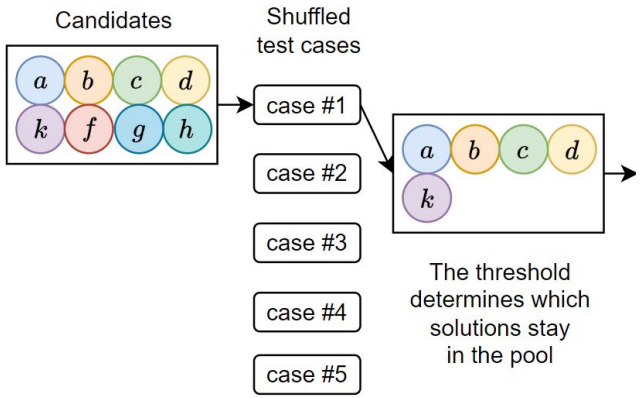


Takeaways

By changing only the selection, you can affect the performance. It plays a crucial role in EAs.

We propose a way of splitting the pool with a more natural interpretation.

We got more accurate solutions, at the cost of a higher runtime (due to larger selection pools).



ϵ-lexicase with minimum variance split

Thank you!

Guilherme Aldeia

guilherme.aldeia@ufabc.edu.br

UFABC

Boston Children's Hospital
Until every child is well™

AVA Lab

VE RI TAS

chip Computational Health Informatics Program

CMCC Centro de Matemática, Computação e Cognição