

Online Diversity Control in Symbolic Regression via a Fast Hash-based Tree Similarity Measure

Bogdan Burlacu, Michael Affenzeller, Gabriel Kronberger, Michael Kommenda

Josef Ressel Centre for Symbolic Regression (JRZ)
Heuristic and Evolutionary Algorithms Laboratory (HEAL)
University of Applied Sciences Upper Austria, Hagenberg

Contact:

Bogdan Burlacu
Heuristic and Evolutionary
Algorithms Lab (HEAL)
Softwarepark 11
4232 Hagenberg, Austria

E-mail:

bogdan.burlacu@fh-hagenberg.at

Web:

<http://heal.heuristiclab.com>

<http://dev.heuristiclab.com>



Diversity in Genetic Programming

☉ Diversity represents an important aspect of Genetic Programming

- Describes state of convergence
- Exploration – exploitation trade-off

☉ Many diversity measures for GP [Črepinšek et al., 2013]

- Semantic – phenotypic
- Structural – genotypic
- History-based
- Distance-based
- Difference-based
- Entropy-based
- ...



GP Tree Distance

Tree distances seldomly used due to high runtime costs

[Bille, P. (2005). A survey on tree edit distance and related problems. *Theoretical computer science*, 337(1-3), 217-239.]

- Generally NP-Hard
- At least quadratic in time and space

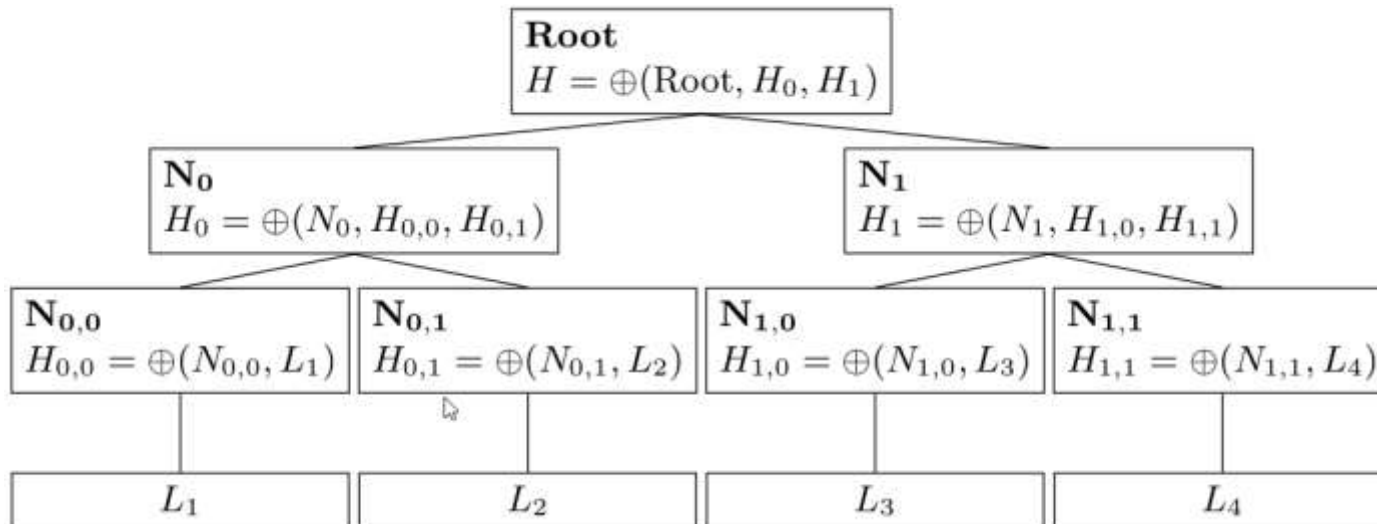
Hash-based tree distance

- Simple algorithm tailored for GP
- Single expensive pass over the population
- Fast calculation of distance matrix
- *Inexact* due to potential hash collisions

Hash-based Tree Distance

Tree hashing algorithm

- Converts a tree to a sequence of hash values
- Same hash value \rightarrow isomorphic subtrees
- Can handle structural and semantic differences
- Parent hash values aggregated from own label + child hash values
- \oplus is a general-purpose hash function (can be cryptographic)



Tree Hashing Algorithm

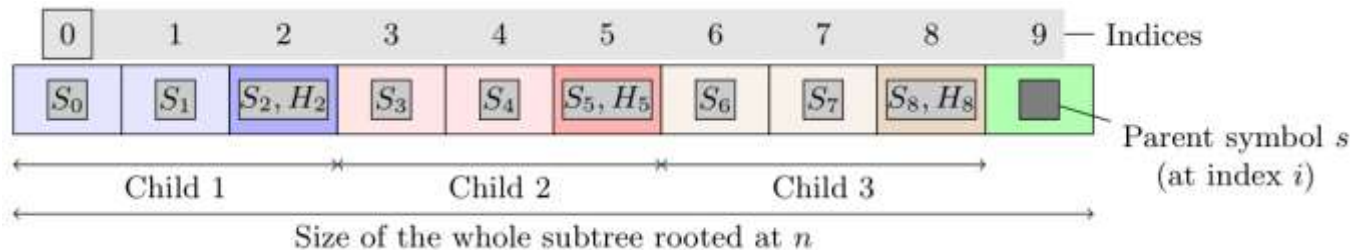
Bottom-up approach

Sort commutative symbols

Hashing mode:

- Structural: only hash node labels (*Constant*, X_1 , Y , etc.)
- Hybrid: hash numerical coefficients (3.14 , $3 \cdot X_1$, $0.6 \cdot Y$, etc.)

Simple array operations





Hash-based Tree Distance

Distance defined as ratio of common hash values

- Sørensen–Dice coefficient returns a value in $[0,1]$
- M is a mapping of isomorphic subtrees (same hash value)
- M computed as the intersection of corresponding hash value sequences H_1, H_2

$$S(T_1, T_2) = \frac{2 \cdot |M|}{|T_1| + |T_2|} \text{ (tree similarity)}$$

$$D(T_1, T_2) = 1 - S(T_1, T_2) \text{ (tree distance)}$$

Single hashing pass over GP population

- Hash each tree into corresponding hash value sequence
- Sort all hash value sequences
- Simple merge-count to calculate tree distance



Hash-based Tree Distance

Runtime performance

- Identical results to bottom-up tree distance [Valiente, 2001]
- Compute distance matrix for 5000 random trees
- Single-mode: hash each tree anew when computing distance matrix
- Batch-mode: hash and sort all trees, use hash value sequences for further computation

TABLE I
RUNTIME PERFORMANCE OF HASH-BASED TREE DISTANCE VS THE
BOTTOM-UP TREE DISTANCE

Tree distance method	Elapsed time (s)	Speed-up
Bottom-up	1225.751	1.0x
Hash-based (single-mode)	297.521	4.1x
Hash-based (batch-mode)	3.677	333.3x

Suitable for online diversity control



Online Diversity Control

Approach

- Compute distance matrix per generation
- Favor individuals that are “farther away”

Implementation

- Standard GA: average distance as penalty to fitness (during selection)
- NSGA-II: maximize average distance as secondary objective

Secondary objectives

- Structural and hybrid distance
- Recursive complexity
- Tree size
- Visitation length
- Number of variables



Performance Ranking

Median rank over all problems

- NSGA-II Hybrid and GA Hybrid overall best
- NSGA-II Hybrid produces higher quality and lower result variance

Algorithm	Training rank	Test rank
NSGA-II Hybrid distance	1.0	1.0
GA Hybrid distance	2.0	3.0
NSGA-II Nested tree size	4.0	4.0
GA Structural distance	5.0	5.0
NSGA-II Tree size	5.0	5.0
GA Standard	6.0	6.0
NSGA-II Tree complexity	6.0	6.0
NSGA-II Structural distance	7.0	6.0

Normalized Mean Squared Error (Train)

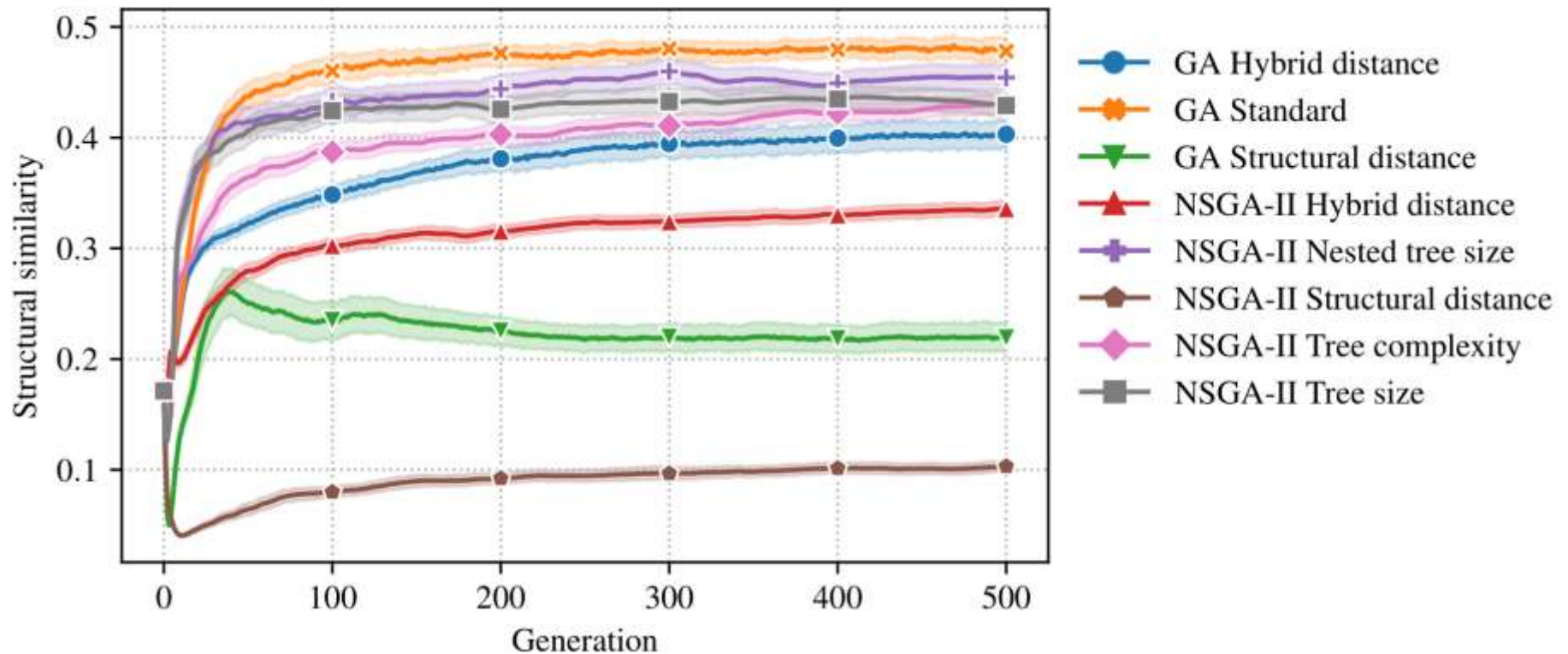
Algorithm	Breiman - I	Friedman - I	Friedman - II	Paglie-1	Poly-10	Vladislavleva-1	Vladislavleva-2	Vladislavleva-3	Vladislavleva-4	Vladislavleva-5	Vladislavleva-6	Vladislavleva-7	Vladislavleva-8
GA	0.129	0.142	0.052	0.003	0.173	0.001	0.012	0.022	0.045	0.001	0.112	0.099	0.181
GA Hybrid Distance	0.115	0.139	0.041	0.003	0.125	0.001	0.004	0.017	0.043	0.001	0.057	0.092	0.033
GA Structural Distance	0.123	0.142	0.041	0.004	0.171	0.002	0.005	0.049	0.043	0.003	0.068	0.105	0.020
NSGA-II Hybrid Distance	0.110	0.137	0.040	0.001	0.128	0.000	0.001	0.004	0.030	0.000	0.000	0.079	0.007
NSGA-II Nested tree length	0.121	0.145	0.086	0.006	0.330	0.001	0.004	0.013	0.030	0.002	0.000	0.093	0.011
NSGA-II Structural Distance	0.149	0.150	0.069	0.003	0.177	0.002	0.008	0.027	0.032	0.003	0.124	0.103	0.039
NSGA-II Tree Complexity	0.113	0.165	0.110	0.007	0.183	0.002	0.011	0.018	0.046	0.022	0.036	0.097	0.013
NSGA-II Tree size	0.122	0.152	0.098	0.008	0.187	0.001	0.004	0.012	0.024	0.002	0.000	0.094	0.030

Normalized Mean Squared Error (Test)

Algorithm	Breiman - I	Friedman - I	Friedman - II	Paglie-1	Poly-10	Vladislavleva-1	Vladislavleva-2	Vladislavleva-3	Vladislavleva-4	Vladislavleva-5	Vladislavleva-6	Vladislavleva-7	Vladislavleva-8
GA	0.134	0.143	0.053	0.074	0.172	0.046	0.015	0.063	0.096	0.013	1.372	0.118	0.642
GA Hybrid Distance	0.120	0.139	0.042	0.010	0.146	0.011	0.009	0.018	0.095	0.004	0.486	0.104	0.520
GA Structural Distance	0.130	0.141	0.042	0.005	0.186	0.028	0.010	0.076	0.108	0.015	0.543	0.126	0.480
NSGA-II Hybrid Distance	0.117	0.137	0.041	0.007	0.147	0.015	0.002	0.008	0.060	0.002	0.000	0.108	0.427
NSGA-II Nested tree length	0.126	0.145	0.090	0.010	0.383	0.017	0.006	0.017	0.067	0.009	0.000	0.106	0.534
NSGA-II Structural Distance	0.154	0.149	0.072	0.005	0.195	0.045	0.013	0.111	0.096	0.140	0.953	0.147	0.488
NSGA-II Tree Complexity	0.117	0.160	0.116	0.007	0.187	0.032	0.021	0.022	0.093	0.109	1.445	0.107	0.759
NSGA-II Tree size	0.127	0.150	0.106	0.009	0.209	0.017	0.008	0.017	0.051	0.009	0.000	0.112	0.744

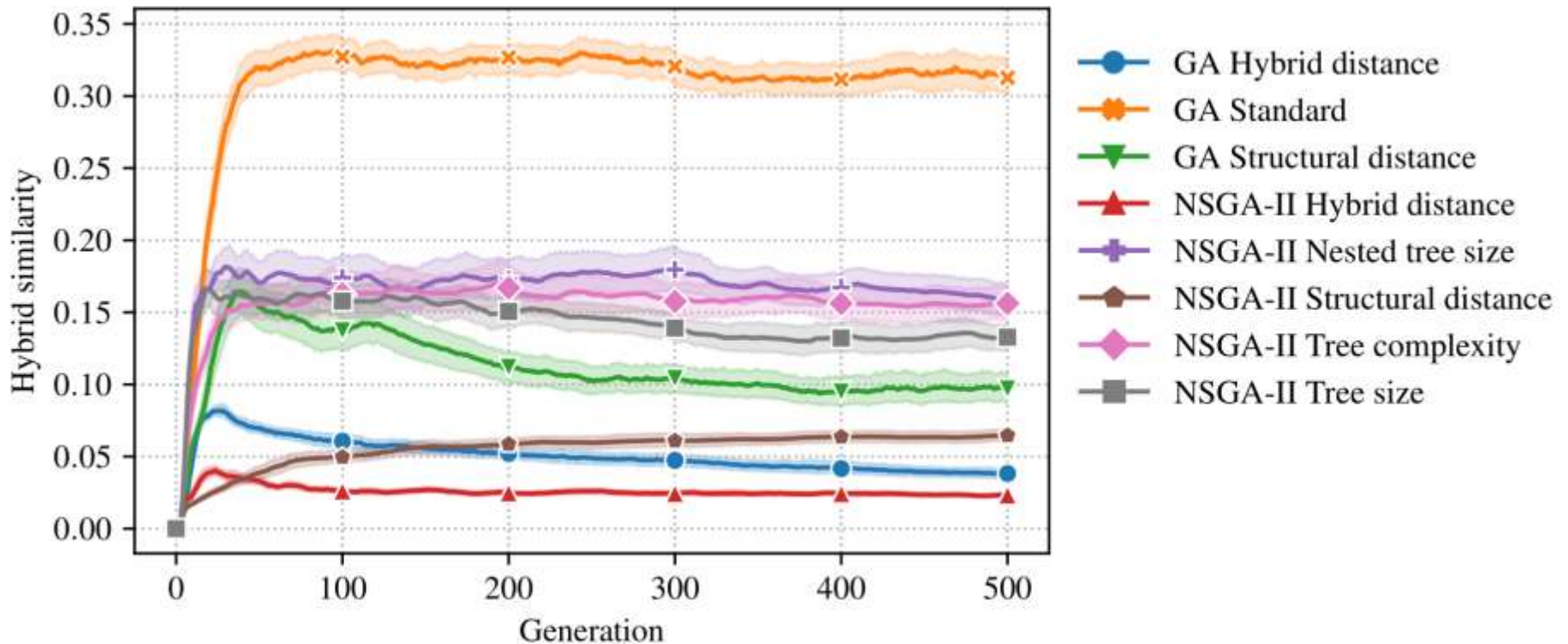
Evolution of Diversity

Structural similarity $S = 1 - D$



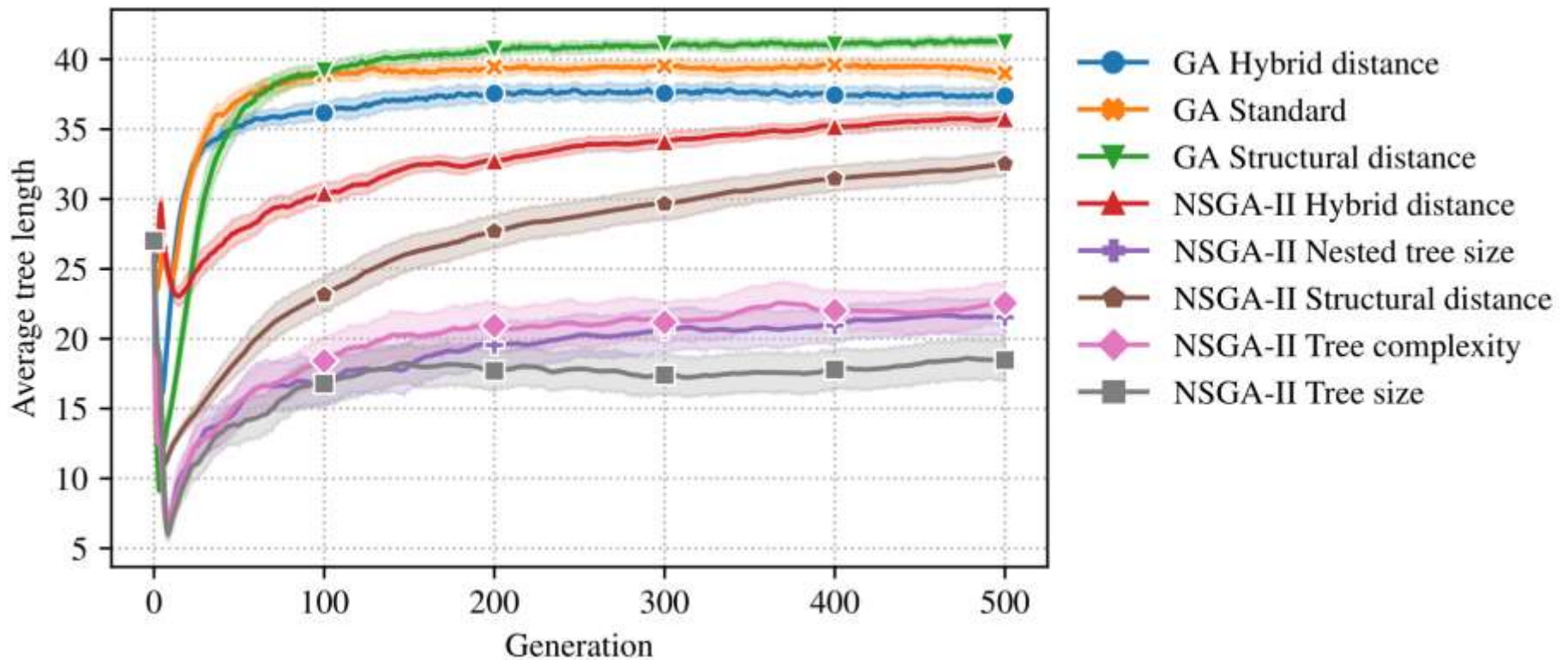
Evolution of Diversity

Hybrid similarity $S = 1 - D$



Evolution of Tree Length

Average tree length





Runtime Analysis

Distance computation overhead

Algorithm	Runtime (s)
Poly-10 (250 training rows)	
GA Standard (baseline)	160.1
GA Hybrid distance	285.1
GA Structural distance	301.9
NSGA-II Hybrid distance	317.5
NSGA-II Structural distance	293.1
NSGA-II Complexity	195.5
NSGA-II Nested Tree Length	196.5
NSGA-II Tree Length	189.8
Breiman-I (5000 training rows)	
GA Standard (baseline)	899.3
GA Hybrid distance	1034.9
GA Structural distance	1204.1
NSGA-II Hybrid distance	1092.0
NSGA-II Structural distance	1140.6
NSGA-II Complexity	830.4
NSGA-II Nested Tree Length	842.0
NSGA-II Tree Length	820.3



Summary

Hash-based tree distance for GP

- Computationally-efficient
- Semantically-aware
- Hash-value sequences → mining of common patterns and building blocks
- Easy to integrate with other GP flavors or other operators (e.g. Crossover)
- Open-source: <https://dev.heuristiclab.com/>

Online diversity control

- Feasible for large populations
- Proof of concept successfully compares to other GP variants

Future research

- More sophisticated diversity tuning
- Detailed analysis of diversity
- Mining of common subtrees

Discussion



Online Diversity Control in Symbolic Regression via a Fast Hash-based Tree Similarity Measure

Bogdan Burlacu, Michael Affenzeller, Gabriel Kronberger, Michael Kommenda

Josef Ressel Centre for Symbolic Regression (JRZ)
Heuristic and Evolutionary Algorithms Laboratory (HEAL)
University of Applied Sciences Upper Austria, Hagenberg

Contact:

Bogdan Burlacu
Heuristic and Evolutionary
Algorithms Lab (HEAL)
Softwarepark 11
4232 Hagenberg, Austria

E-mail:

Bogdan.burlacu@fh-hagenberg.at

Web:

<http://heal.heuristiclab.com>

<http://dev.heuristiclab.com>