SymReg

JOSEF RESSEL CENTER FOR
SYMBOLIC REGRESSION

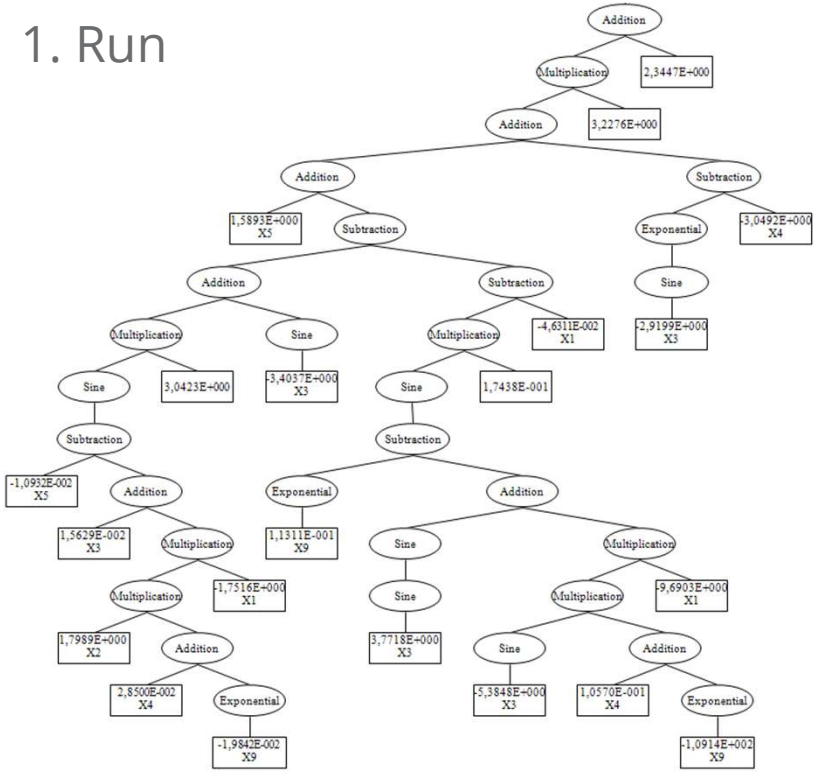
Empirical Analysis of Variance for GP based Symbolic Regression

Lukas Kammerer

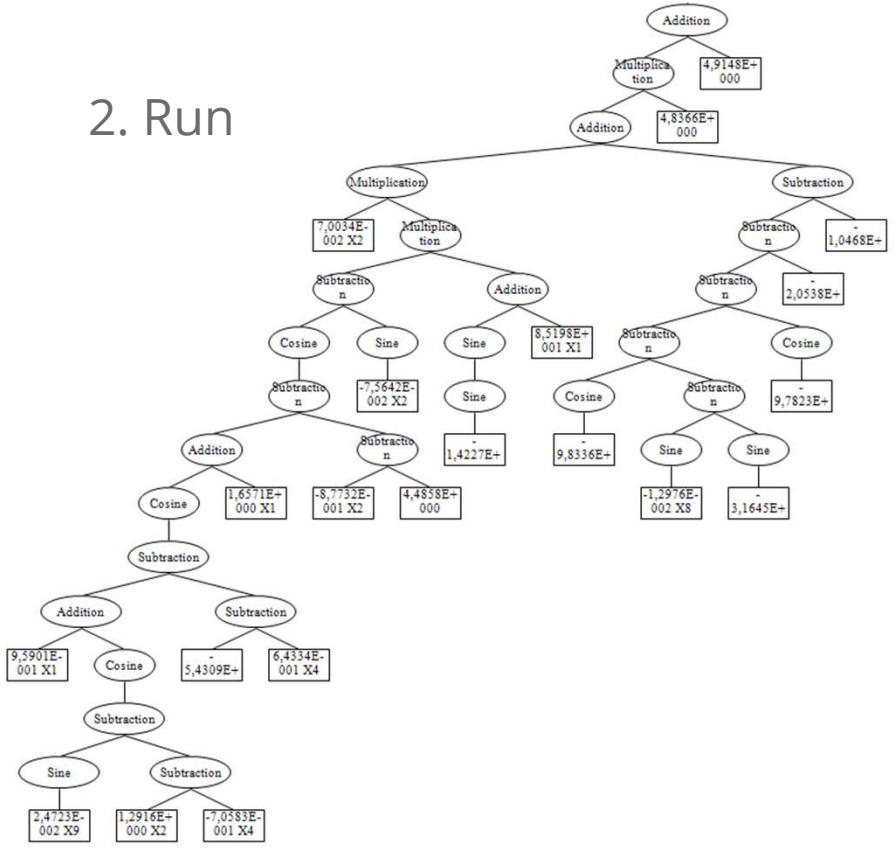
JRZ SymReg Workshop
24.02.2021

Motivation

1. Run

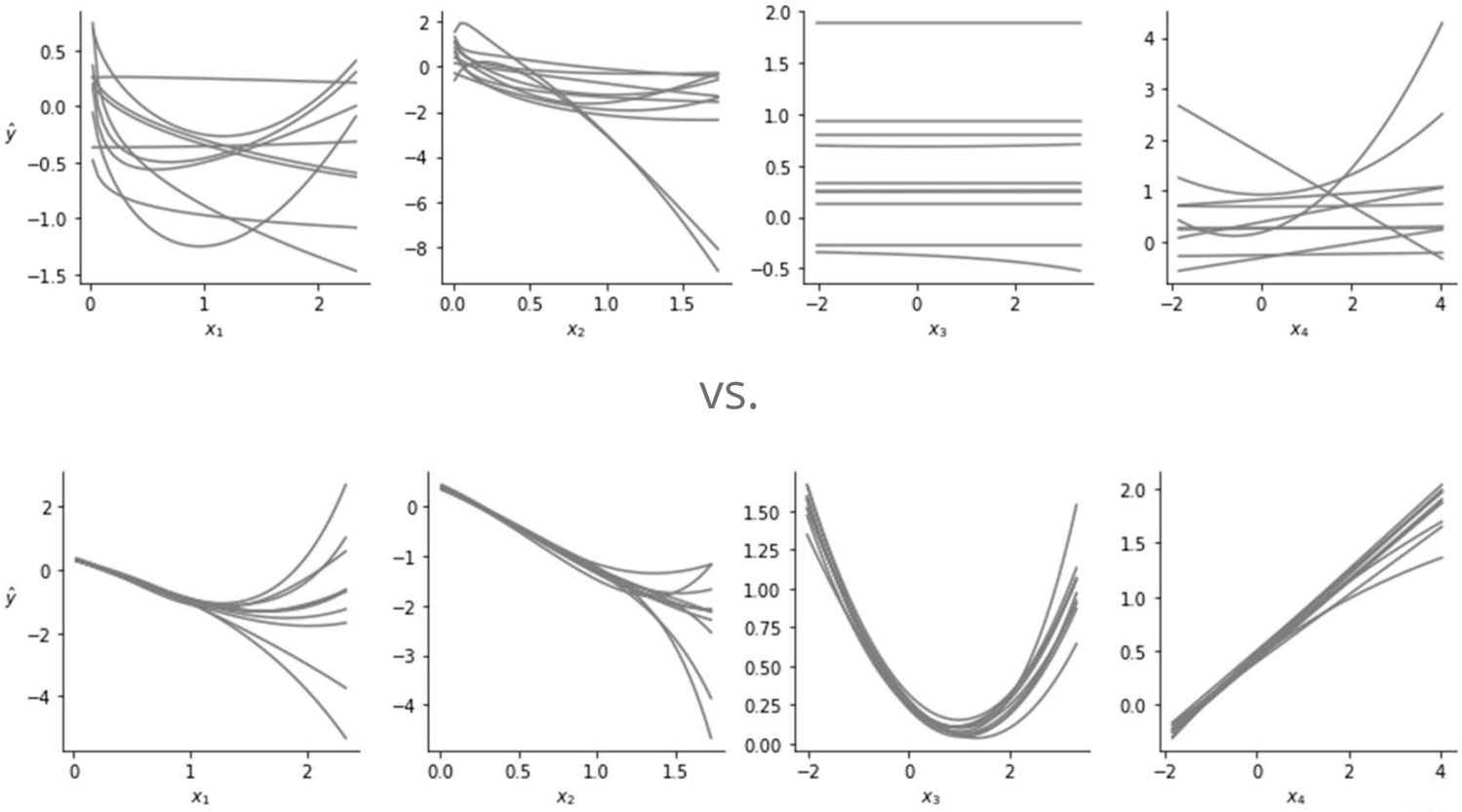


2. Run



SymReg

Motivation



vs.

SymReg

Phenotypical Spread

Given:

- $\mathbf{m}$ with k models
- Dataset D with n samples

$$s(\boldsymbol{m}, D) = \frac{1}{n}\sum_{i=0}^{n} \mathrm{IQR}(\{\mathrm{m}_1(D_i) \ \ldots \ m_k(D_i)\})$$

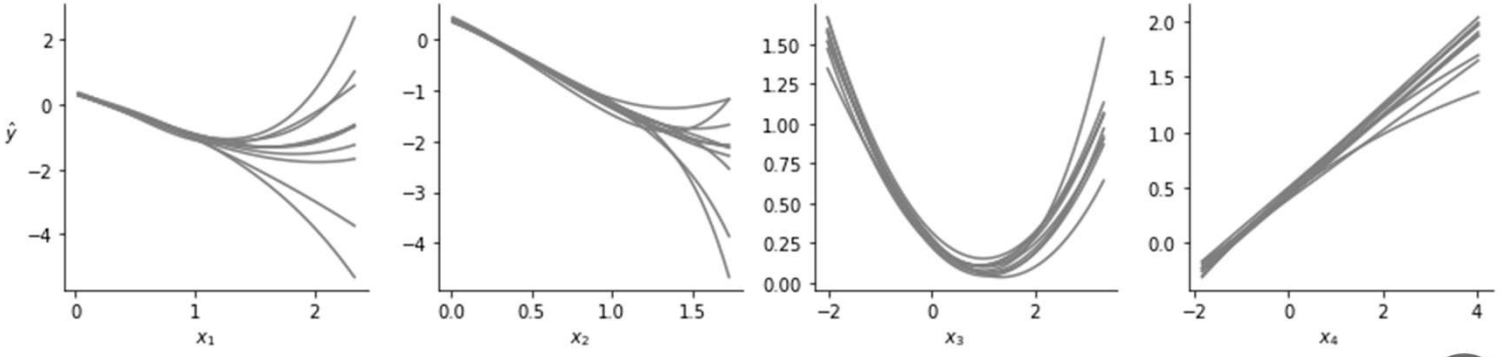
SymReg

Phenotypical Spread

Given:

- $\mathbf{m}$ with k models
- Dataset D with n samples

$$s(\boldsymbol{m}, D) = \frac{1}{n}\sum_{i=0}^{n} \mathrm{IQR}(\{\mathrm{m}_1(D_i) \ \ldots \ m_k(D_i)\})$$



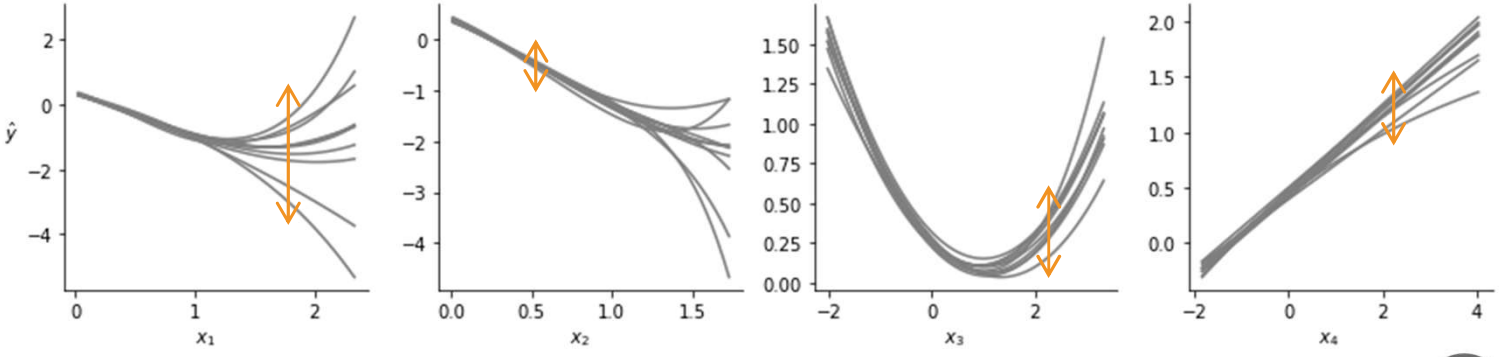
SymReg

Phenotypical Spread

Given:

- $\mathbf{m}$ with k models
- Dataset D with n samples

$$s(\boldsymbol{m}, D) = \frac{1}{n}\sum_{i=0}^{n} \mathrm{IQR}(\{\mathrm{m}_1(D_i) \ \ldots \ m_k(D_i)\})$$



SymReg

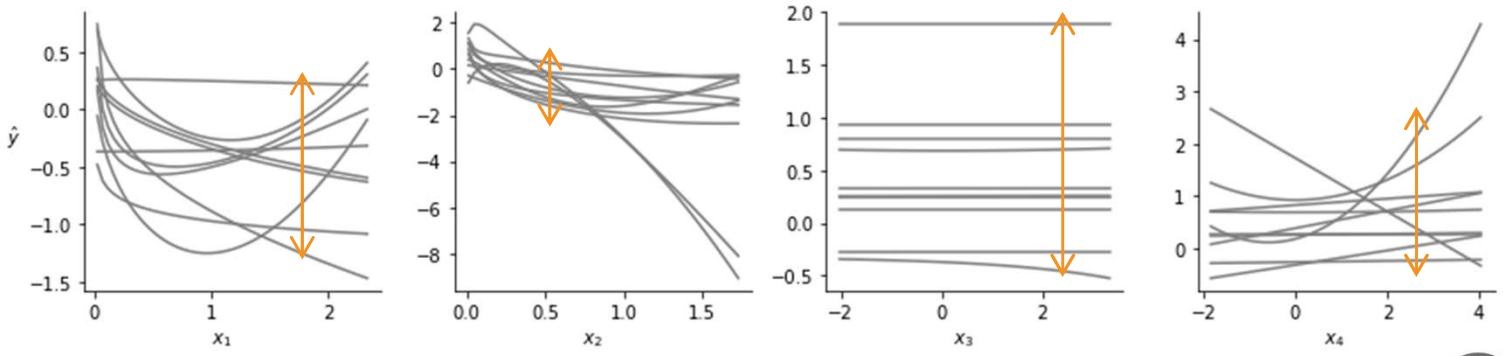
Phenotypical Spread

Given:

- $\mathbf{m}$ with k models
- Dataset D with n samples

$$s(\boldsymbol{m}, D) = \frac{1}{n}\sum_{i=0}^{n} \mathrm{IQR}(\{\mathrm{m}_1(D_i) \ \ldots \ m_k(D_i)\})$$

cf. Maarten Keijzer and Vladan Babovic. 2000. Genetic Programming, Ensemble Methods and the Bias/Variance Tradeoff – Introductory Investigations. In *European Conference on Genetic Programming*. Springer, 76–90.



SymReg

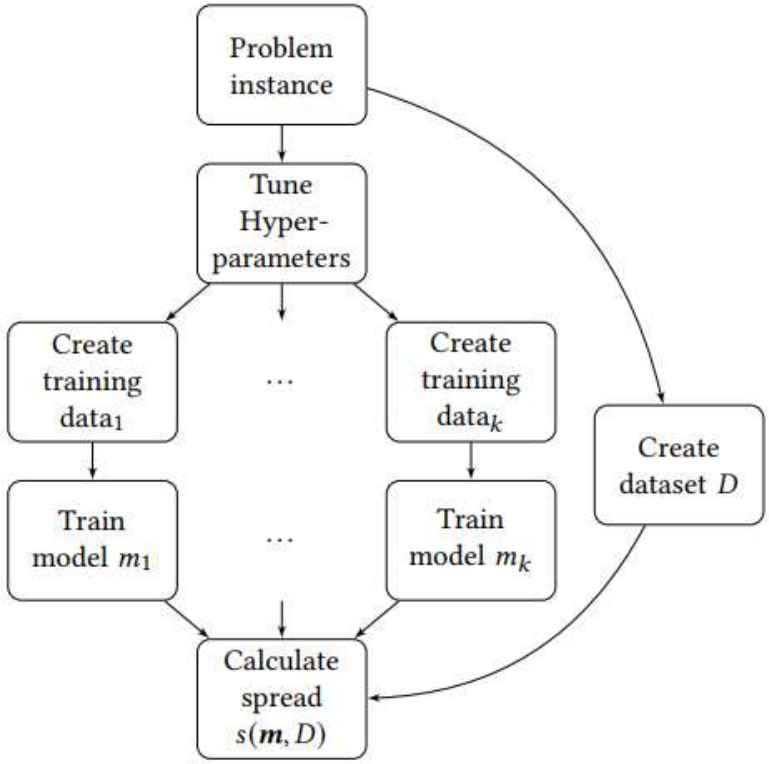
Phenotypical Spread

Given:

- $\mathbf{m}$ with k models
- Dataset D with n samples

$$s(\boldsymbol{m}, D) = \frac{1}{n}\sum_{i=0}^{n} \mathrm{IQR}(\{\mathrm{m}_1(D_i) \ \ldots \ m_k(D_i)\})$$

- 96 problem instances
- 5 (non-)evolutionary Algorithms

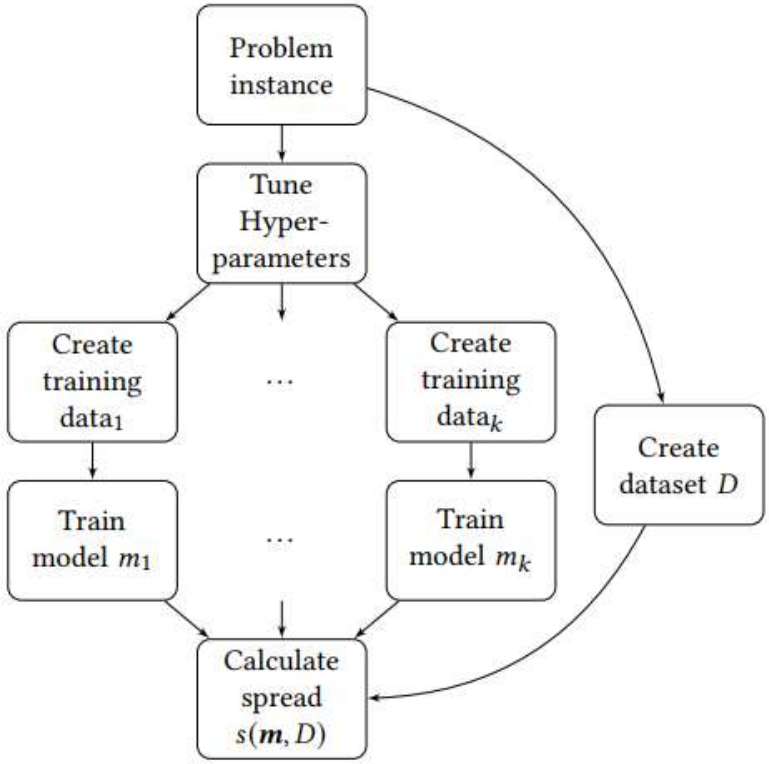


SymReg

Experimental Setup

- PennML Benchmark Library
  - 32 real world problems
  - 62 Friedman problems

- Algorithms
  - GP
  - GP with local optimization (GP NLS)
  - Linear regression (LR)
  - Polynomial regression (PR)
  - Random forest regression (RF)



SymReg

Experimental Setup

- GP and GP NLS
  - max. tree length ∈ {10, 25, 50, 75}
  - population size ∈ {100, 500, 1000}
  - max. generations ∈ {100, 200, 1000}

  cf. Michael Kommenda, Bogdan Burlacu, Gabriel Kronberger and Michael Affenzeller. 2019. Parameter identification for symbolic regression using nonlinear least squares. Genetic Programming and Evolvable Machines (2019), 1–31.
- RF
  - 200 trees
  - $m \in \{0.25, 0.5, 0.75\}$
  - $r \in \{0.1, 0.2, \ldots, 0.7\}$
- PR
  - $\alpha \in \{0, 0.5, 1\}$
  - $\lambda \in \{1 \cdot 10\text{-}7, 2.5 \cdot 10\text{-}7, 5 \cdot 10\text{-}7, 7.5 \cdot 10\text{-}7, 1 \cdot 10\text{-}6, 2.5 \cdot 10\text{-}6 \ldots 7.5 \cdot 10\text{-}2\}$
  - total polynomial degree ∈ {2, 3, 4, 5}. (if > 20 features: max. 3)

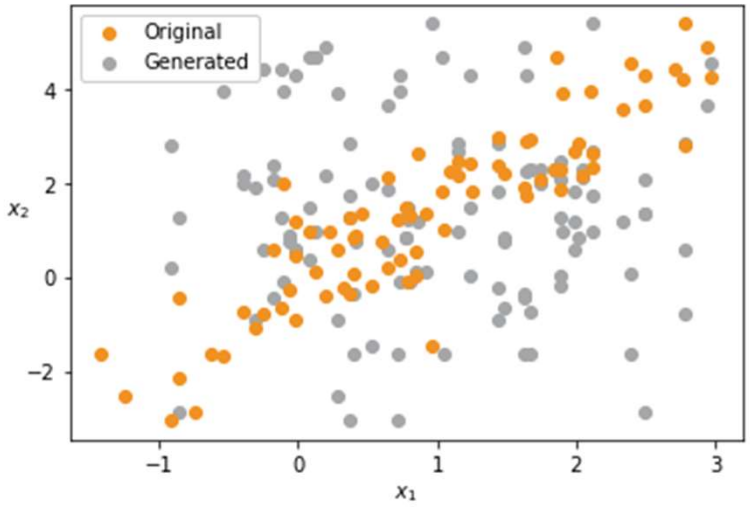
GP and GP NLS Settings

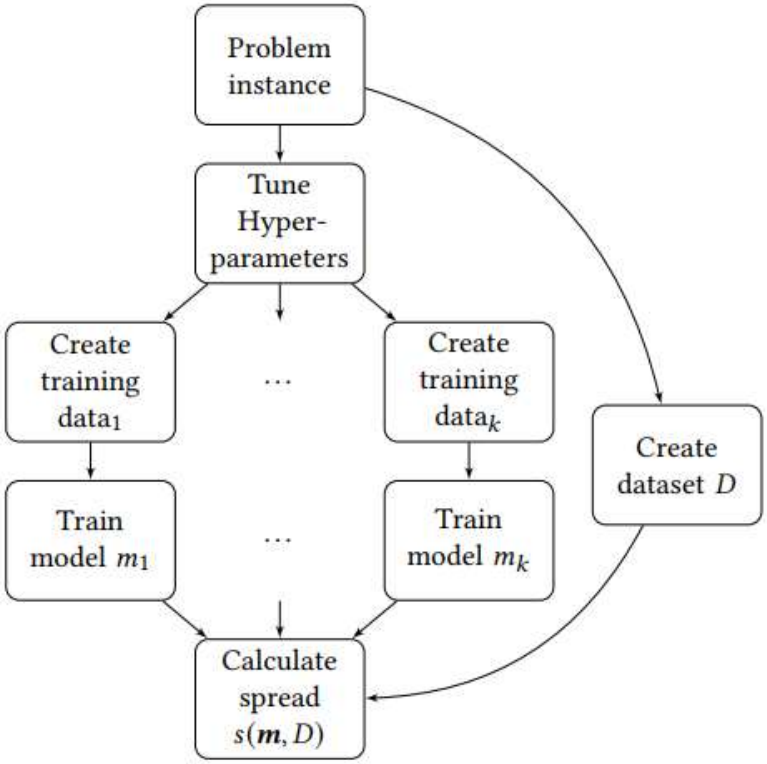
Function set	+, -, ×, ÷, x^2, $\sqrt{x}$, exp(x), log(x)
Max. evaluated models	100 000
Selection	Tournament group size 2
Crossover operator	Subtree crossover, 100% probability
Mutation probability	25%
Mutation operator	Change symbol, replace branch, remove branch, single-point
Fitness function	Mean squared error

SymReg

Experimental Setup

- Dataset D generated by sampling from origianl dataset

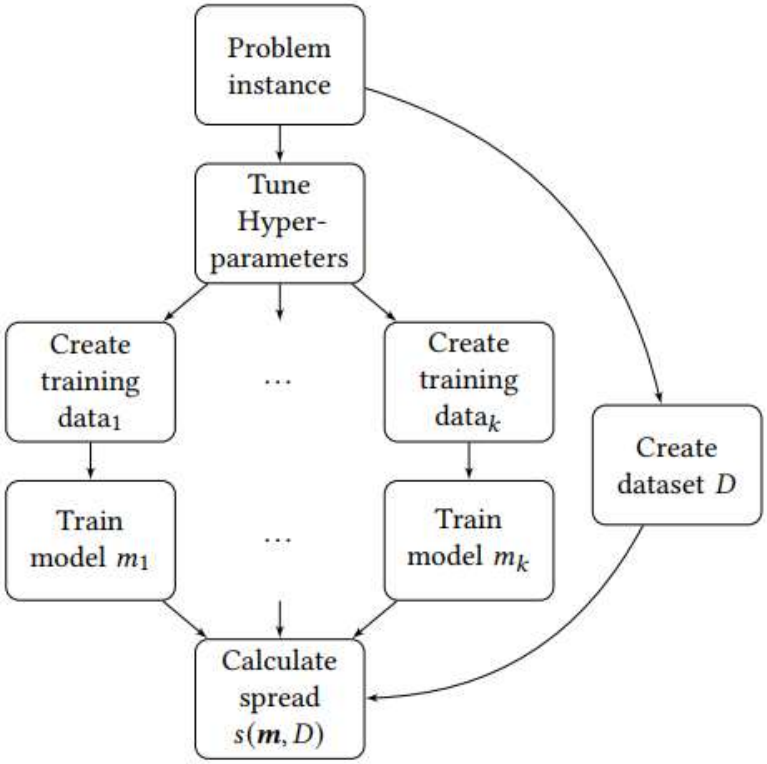




SymReg

Experimental Setup

Problem	LR	PR	RF	GP	GP NLS
192_vineyard	0,82	0,83	0,67	1.0	1.47
195_auto_price	2058,8	1445,8	751,49	1625,6	2256,4
210_cloud	0,14	0,13	0,15	0,24	0,35
228_elusage	2,62	2,87	3,5	2,15	2,58
...	...	...	...	...	...

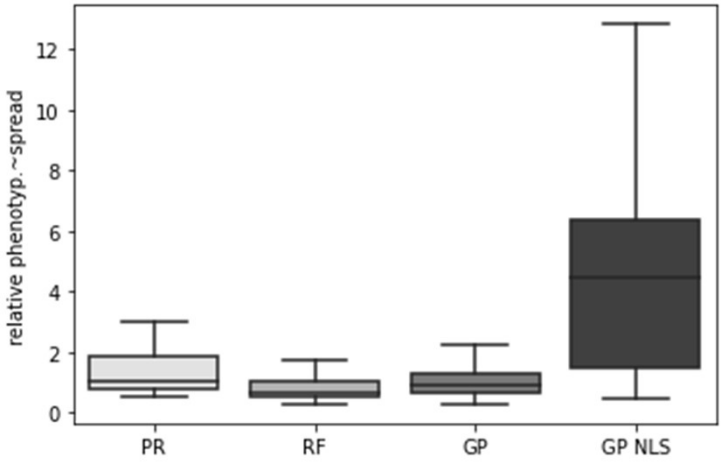


SymReg

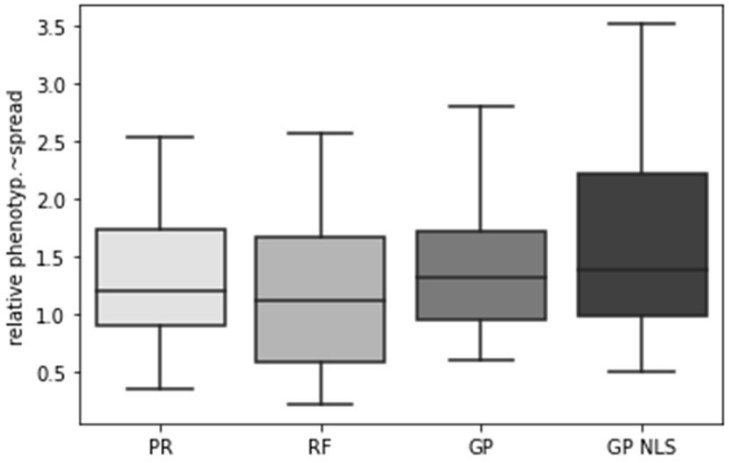
Distribution of Spread

Values per problem instance relative to value of linear regression

Friedman Problems



Real World Problems



SymReg

Statistical Significance

- Rank-wise analysis

Problem	LR	PR	RF	GP	GP NLS
192_vineyard	0,82	0,83	0,67	1.0	1.47
195_auto_price	2058,8	1445,8	751,49	1625,6	2256,4
210_cloud	0,14	0,13	0,15	0,24	0,35
228_elusage	2,62	2,87	3,5	2,15	2,58
...	...	...	...	...	...

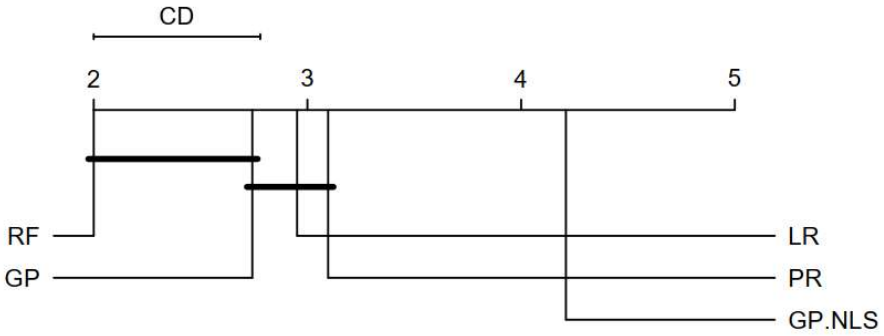
Problem	LR	PR	RF	GP	GP NLS
192_vineyard	2	3	1	4	5
195_auto_price	4	2	1	3	5
210_cloud	2	1	3	4	5
228_elusage	3	4	5	1	2
...	...	...	...	...	...

SymReg

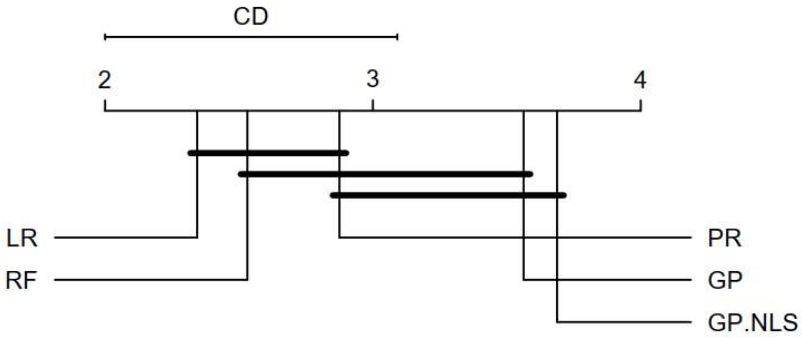
Statistical Significance

- Rank-wise analysis

Friedman Problems



Real World Problems

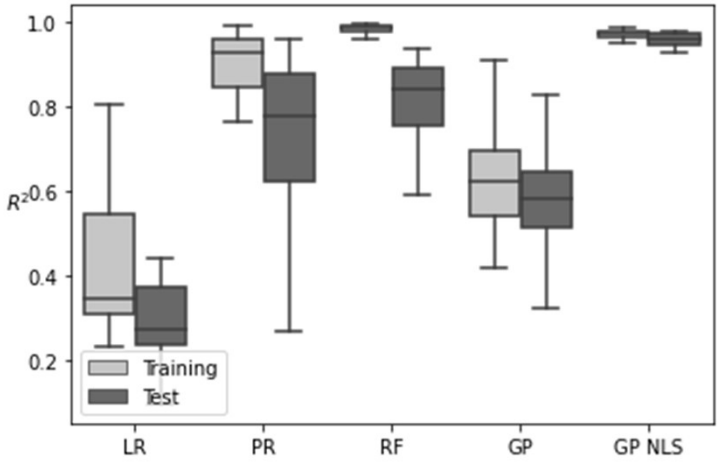


SymReg

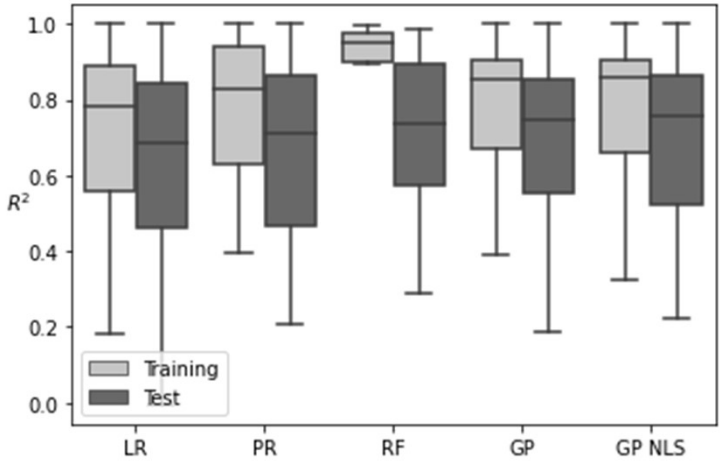
Prediction Error

Median R^2 of all 50 models per Problem

Friedman Problems



Real World Problems

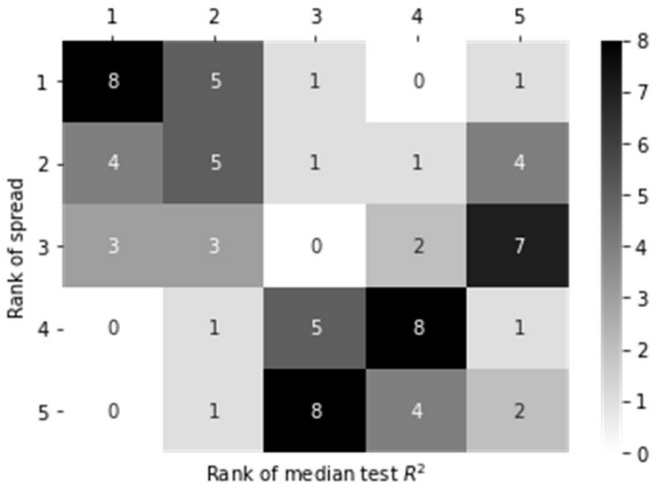


SymReg

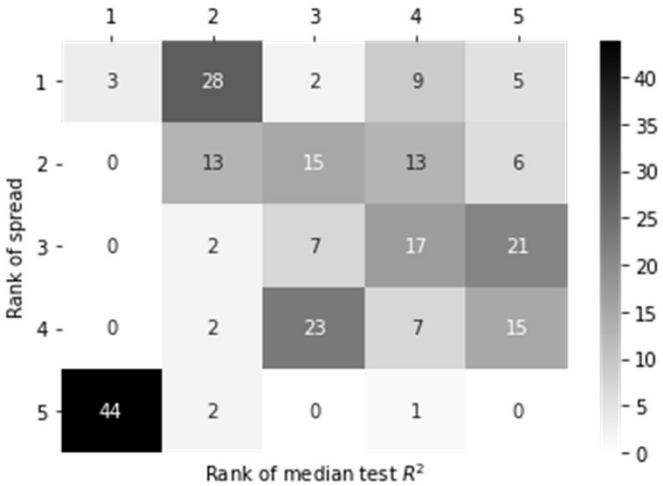
Prediction Error vs. Phenotypical Spread

Rank correlation between prediction error and phenotypical Spread

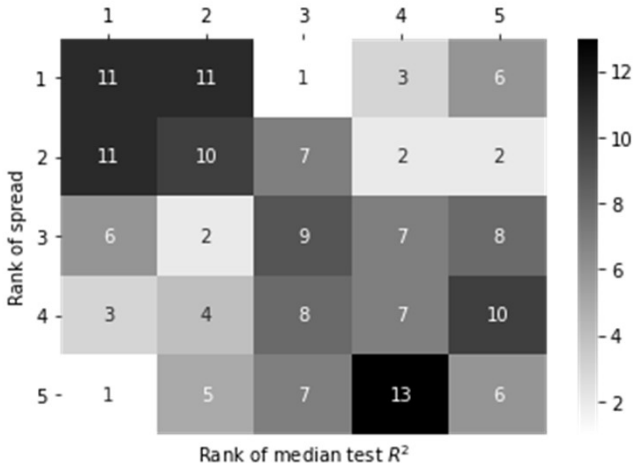
Friedman Problems
without collinearity



Friedman Problems
with collinearity



Real World Problems



SymReg

Spread in Training and Test

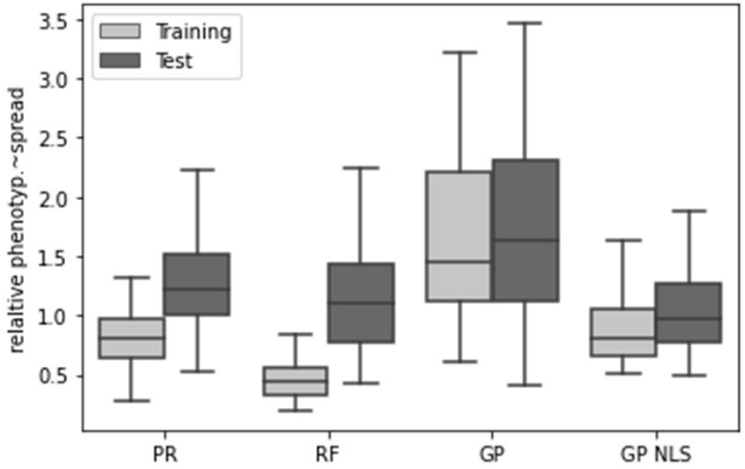
→ Different Calculation of Spread

Sample Index	m_1	m_2	...	m_{50}	IQR Predictions for **Test** instances	IQR Predictions for **Training** instances
1	Train	Train	Test	Train	...	...
2	Train	Test	Train	Train	...	...
3	Test	Test	Train	Test	...	...
4	Train	Train	Train	Train	...	...
...	...	...	...	...	...	...

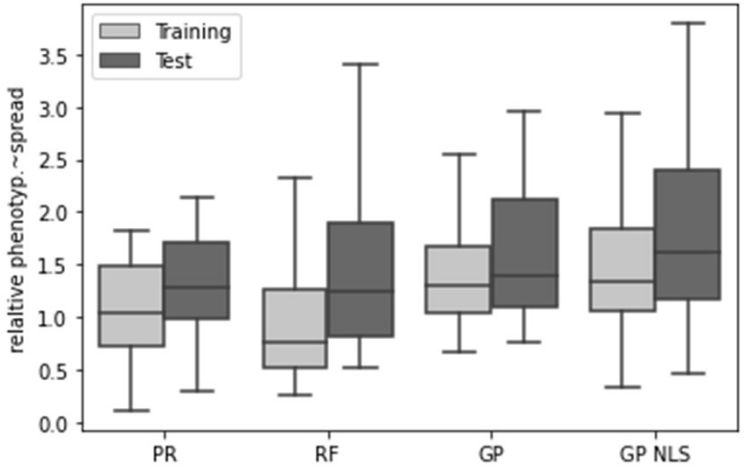
SymReg

Spread in Training and Test

Friedman Problems



Real World Problems



SymReg

Conclusion

- Clear differences in modelling accuracy between Friedman and real world Problems
  → High accuracy on Friedman Problems → good results on whole benchmark

- Real world problems and Friedman problems without collinearity:
  Higher accuracy ↔ Lower Spread and vice versa

- Friedman problems with collinearity:
  High accuracy ↔ High Spread
  – Only in extrapolation

SymReg

JOSEF RESSEL CENTER FOR
SYMBOLIC REGRESSION

Questions?