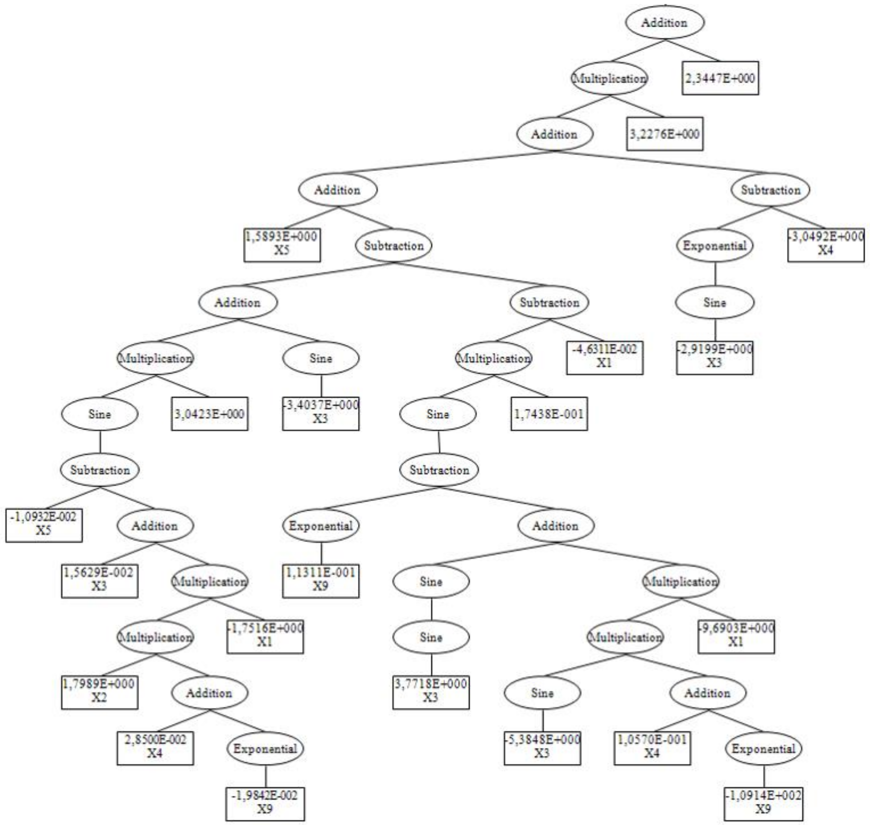
Symbolic Regression with Fast Function Extraction and Nonlinear Least Squares Optimization

Lukas Kammerer, Gabriel Kronberger, Michael Kommenda

SymReg
JOSEF RESSEL CENTER FOR SYMBOLIC REGRESSION

Motivation

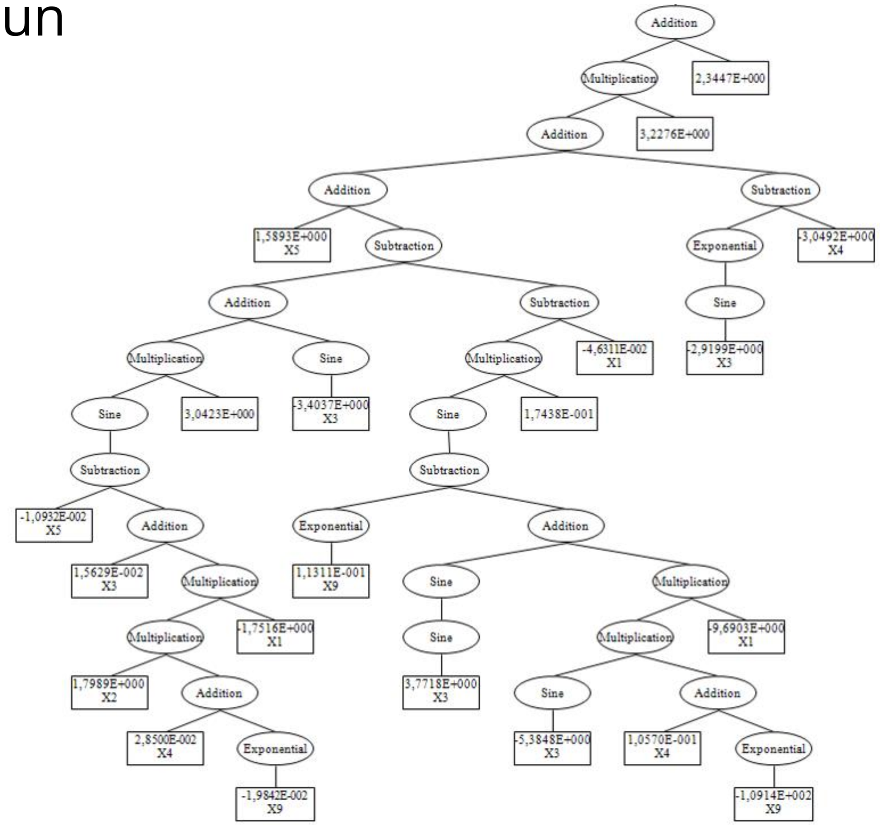
Symbolic Regression with GP



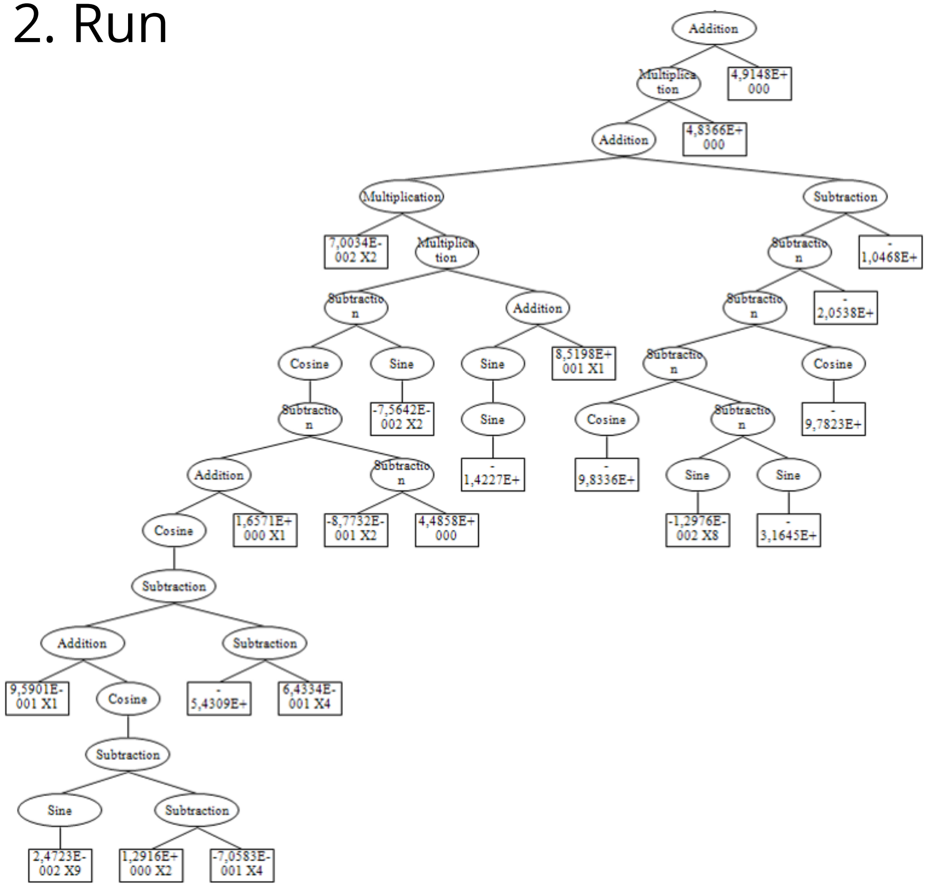
Motivation

Symbolic Regression with GP

1. Run



2. Run



Previous Work

Trent McConaghy. "FFX: Fast, Scalable, Deterministic Symbolic Regression Technology." *GPTP IX*. 2011

Tony Worm and Kenneth Chiu. "Prioritized Grammar Enumeration: Symbolic Regression by Dynamic Programming." *Proceedings of the 15th GECCO*. 2013.

Lukas Kammerer et al. "Symbolic Regression by Exhaustive Search" *GPTP XVII*. 2019

Michael Kommenda. "Parameter Identification for Symbolic Regression using Nonlinear Least Squares." *Genetic Programming and Evolvable Machines 21.3*. 2020

Fast Function Extraction (FFX)

Trent McConaghy. "FFX: Fast, Scalable, Deterministic Symbolic Regression Technology." *GPTP IX*. 2011

Example: Given features x_1, x_2, x_3

Step 1: Generate Base Functions

- Univariate: $x_1, x_1^2, \log(x_1), \log(x_1^2), \exp(\mathrm{x}_1), \exp(x_1^2), abs(x_1), \ldots \quad x_2, x_2^2, \log(x_2) \ldots$
- Bivariate: $x_1 x_2,\ x_1^2 x_2,\ \log(x_1)\, x_2, \ldots$

Step 2: Run ElasticNet Regression with Base Functions as Features:

$$c_1 x_1 + c_2 x_1^2 + c_3 \log(x_1) + \ \ldots + c_n$$

ElasticNet Regression

Hastie, Trevor, et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. New York: springer, 2009.

Given a linear Model: $\hat{f}(\boldsymbol{X}, \boldsymbol{c}) = \boldsymbol{Xc}$ with $\boldsymbol{X}$ as features.

Find c where $\underbrace{\|y - Xc\|^2}_{\text{Mean Squared Error}} + \underbrace{\lambda\left(\alpha\,\|c\|_1 + \frac{1}{2}(1-\alpha)\,\|c\|^2\right)}_{\text{Regularization Term}}$ is minimal.

Fast Function Extraction with Nonlinear Least Squares Optimization

Example: Given features x_1, x_2, x_3

Step 1: Generate Base Functions

$$x_1, x_1^2, \log(x_1 + c_1), \log(x_1^2 + c_2), \exp(c_3 * x_1), \ldots \quad x_2, \ldots$$

Step 2: Optimize linear and nonlinear parameters p with NLS Optimization

$$p_1 x_1 + p_2 x_1^2 + p_3 \log(x_1 + p_4) + p_5 \exp(\mathrm{p}_6 \mathrm{x}_1) + \ldots + \mathrm{p}_\mathrm{n}$$

Fast Function Extraction with Nonlinear Least Squares Optimization

Example: Given features x_1, x_2, x_3

Step 1: Generate Base Functions

$$x_1, x_1^2, \log(x_1 + c_1), \log(x_1^2 + c_2), \exp(c_3 * x_1), \ldots \quad x_2, \ldots$$

Step 2: Optimize linear and nonlinear parameters p with NLS Optimization

$$p_1 x_1 + p_2 x_1^2 + p_3 \log(x_1 + p_4) + p_5 \exp(\mathrm{p}_6 \mathrm{x}_1) + \ldots + \mathrm{p}_\mathrm{n}$$

NLS with Regularized Variable Projection

cf. Chen, Guang-Yong, et al. "A regularized variable projection algorithm for separable nonlinear least-squares problems." IEEE Transactions on Automatic Control 64.2 (2018)

Example: Given model

$$p_1 x_1 + p_2 x_1^2 + p_3 \log(x_1 + p_4) + p_5 \exp(\mathrm{p}_6 \mathrm{x}_1) + \ \ldots + \mathrm{p}_\mathrm{n}$$

with linear and nonlinear parameters.

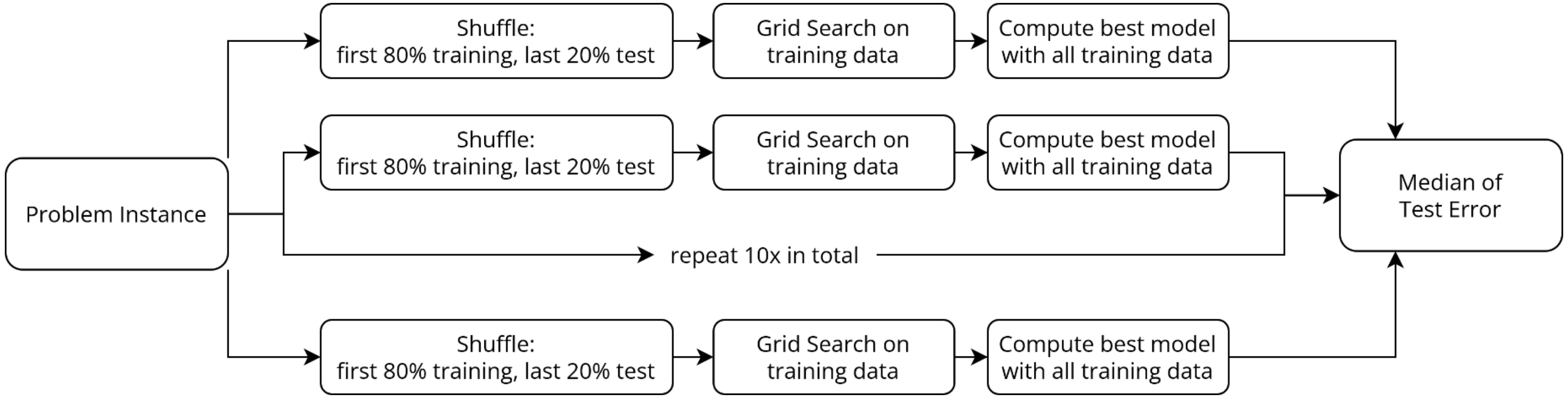
1. Optimize nonlinear parameters with gradient descent while linear parameters are given by regularized linear least squares optimization.

2. Sparsification of linear parameters by removing terms with least variance.

Experimental Setup

cf. Olson, Randal S., et al. "PMLB: a large benchmark suite for machine learning evaluation and comparison." BioData mining 10.1 (2017)

- Workflow for each problem instance:



- Hyperparameter for FFX and FFX VarPro:
  - Max. number of terms ∈ [3, 5, 10, 20, 30]
  - Allow bivariate base functions ∈ [true, false]
  - FFX only: Ratio of L1 Regularization ∈ [0.0, 0.5, 1.0]

Results with Artificial Data

cf. Chen, Chen, et al. "A multilevel block building algorithm for fast modeling generalized separable systems." Expert Systems with Applications 109 (2018)

- Aircraft Lift Coefficient:

$$C_L = C_{L\alpha}(\alpha - \alpha_0) + C_{L\delta_e}\delta_e \frac{S_{HT}}{S_{ref}},$$

- Aircraft Maximum Lift Coeff.

$$\begin{aligned} C_{L\max} =& C_{L\max,W} - C_{L\alpha,WF} \cdot \Delta\alpha_{W/c} + C_{L\alpha,H}\left(\frac{S_H}{S}\right)\left[\alpha_{CL\max}\left(1 - \frac{\partial\varepsilon}{\partial\alpha}\right) - \varepsilon_{0,H} + \psi_H\right] \\ &+ C_{L\alpha,c}\left(\frac{S_c}{S}\right)\left[\alpha_{CL\max}\left(1 + \frac{\partial\varepsilon_c}{\partial\alpha}\right) + \varepsilon_{0,c} + \psi_H\right] \end{aligned}$$

- Rocket Fuel Flow

$$\dot{m} = \frac{p_0 A^*}{\sqrt{T_0}}$$

- Spinning Cylinder Flow

$$\psi = (V_\infty r \sin\theta)\left(1 - \frac{R^2}{r^2}\right) + \frac{\Gamma}{2\pi}\ln\frac{r}{R},$$

200 instances per dataset; Normal distributed Noise

Results with Artificial Data

Dataset	Normalized Mean Squared Error (Test)			Number of Terms			Number of Symbols		
	FFX	FFX VarPro	Difference	FFX	FFX VarPro	Difference	FFX	FFX VarPro	Difference
Aircraft Lift Coefficient	0.015	0.0129	0.0021	28	19	9	133	70	63
Aircraft Max. Lift Coefficient	0.0854	0.067	0.0184	26	31	-5	120	131	-11
Rocket Fuel Flow	0.0102	0.0107	-0.0005	19	10	9	95	37	58
Spinning Cylinder Flow	0.0127	0.0131	-0.0004	21	11	10	102	41	61

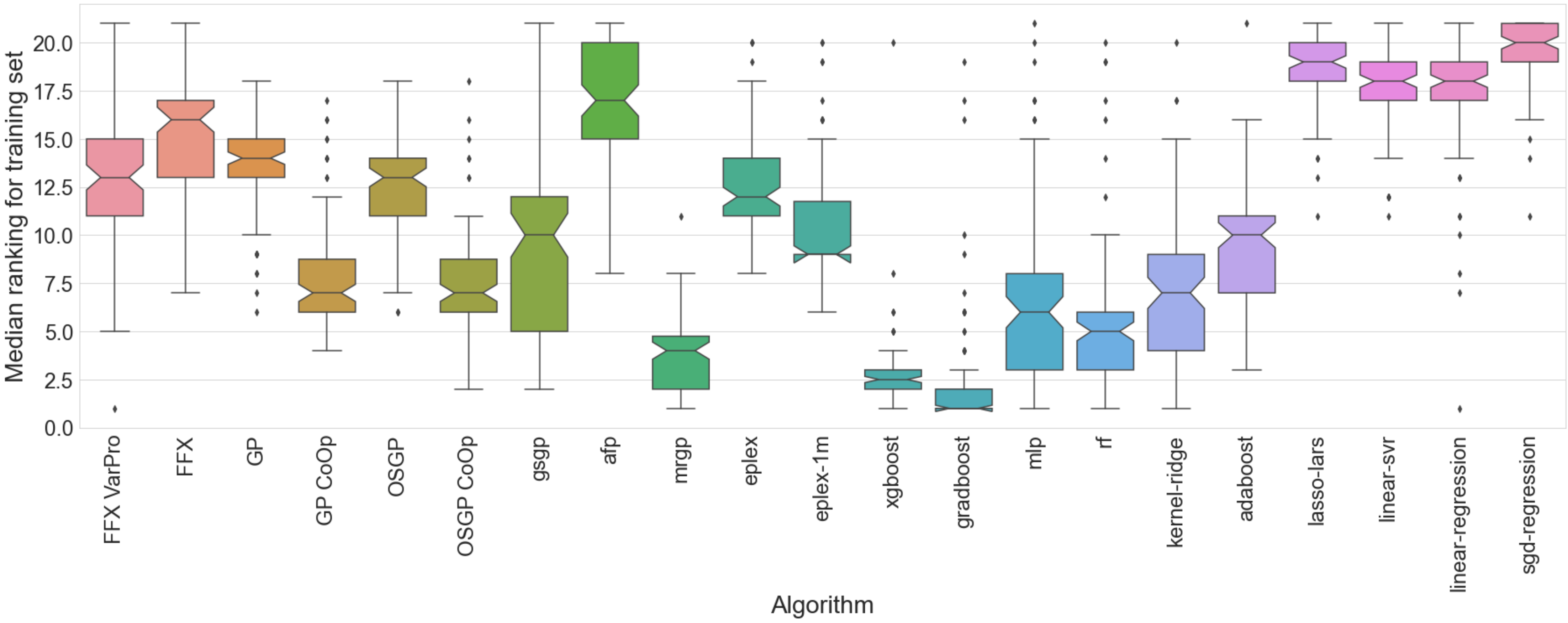
→ Higher Accuracy and/or less complex models with FFX VarPro

Results with Real-World Data

cf. Orzechowski, Patryk, William La Cava, and Jason H. Moore. "Where are we now? A large benchmark study of recent symbolic regression methods." Proceedings of the Genetic and Evolutionary Computation Conference. 2018.

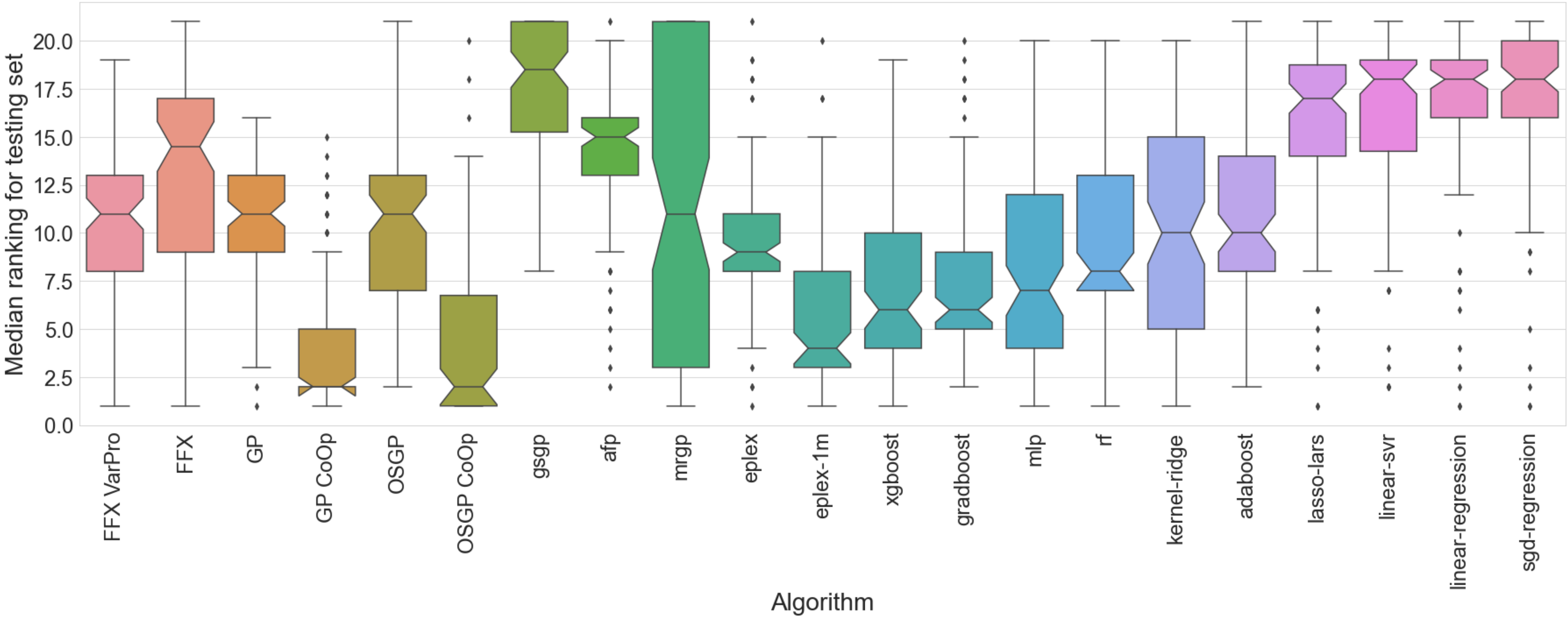
- PennML-Benchmark Suite
  - Comparison of many common regression algorithms
  - 94 regression benchmark problems
  - Accuracy ranking per Problem

Results – Training Accuracy



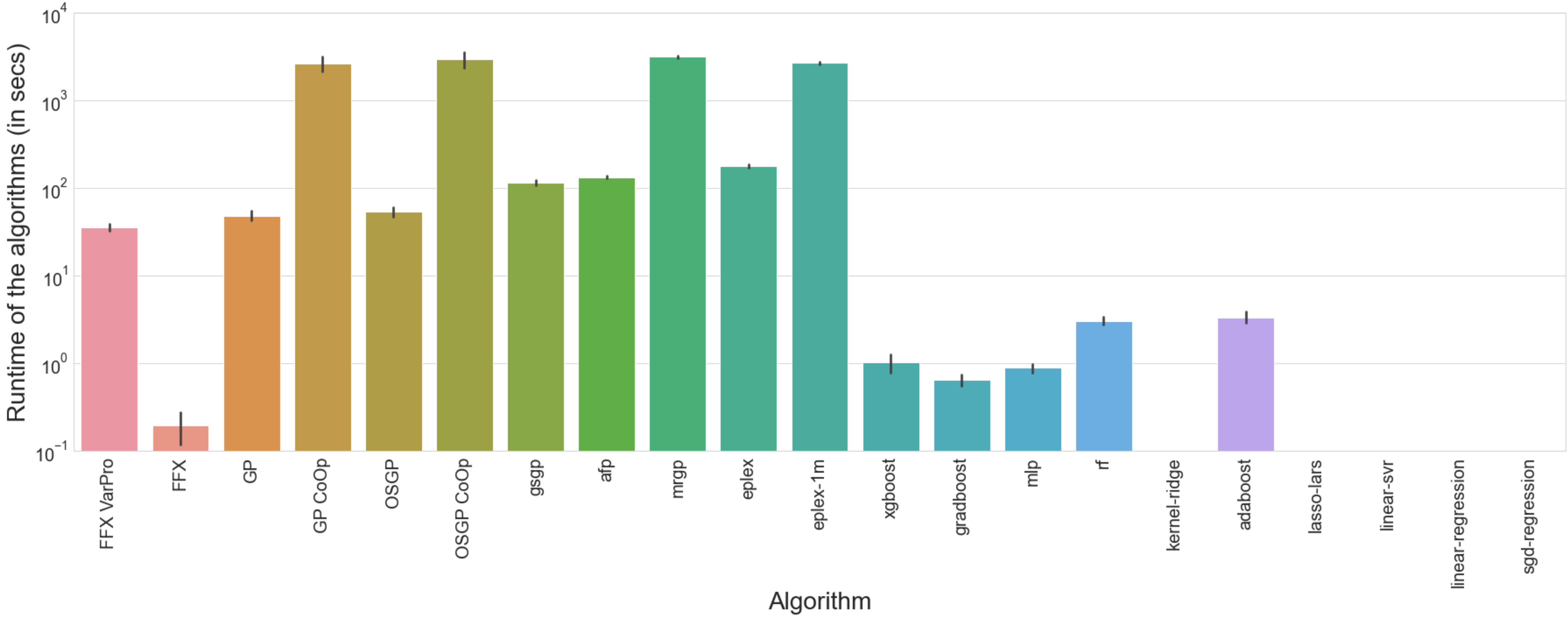
cf. Michael Kommenda. "Parameter Identification for Symbolic Regression using Nonlinear Least Squares." Genetic Programming and Evolvable Machines 21.3. 2020
cf. Orzechowski, Patryk, William La Cava, and Jason H. Moore. "Where are we now? A large benchmark study of recent symbolic regression methods." Proceedings of the Genetic and Evolutionary Computation Conference. 2018.

Results – Test Accuracy



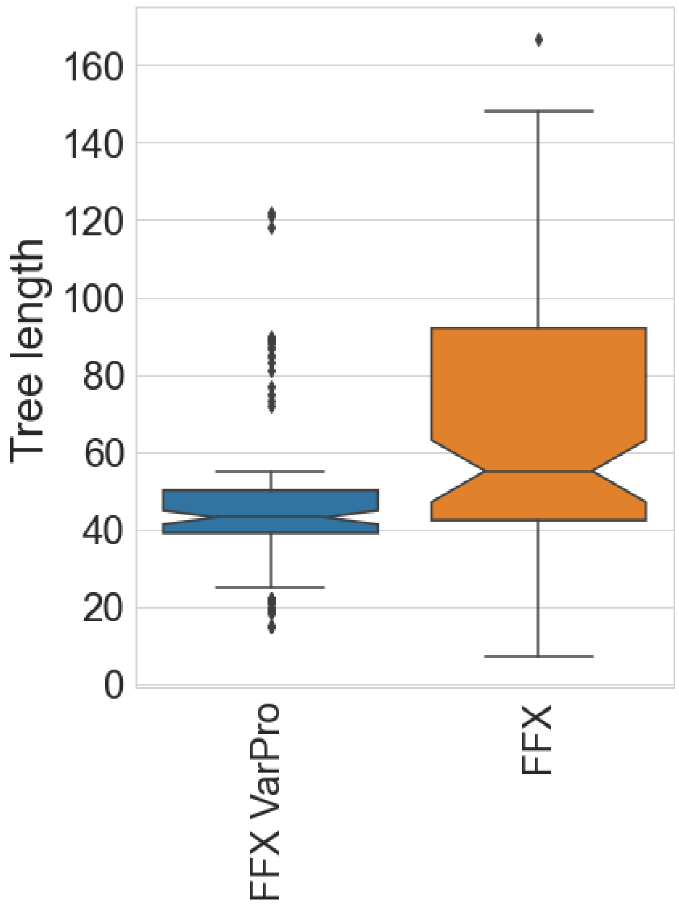
cf. Michael Kommenda. "Parameter Identification for Symbolic Regression using Nonlinear Least Squares." Genetic Programming and Evolvable Machines 21.3. 2020
cf. Orzechowski, Patryk, William La Cava, and Jason H. Moore. "Where are we now? A large benchmark study of recent symbolic regression methods." Proceedings of the Genetic and Evolutionary Computation Conference. 2018.

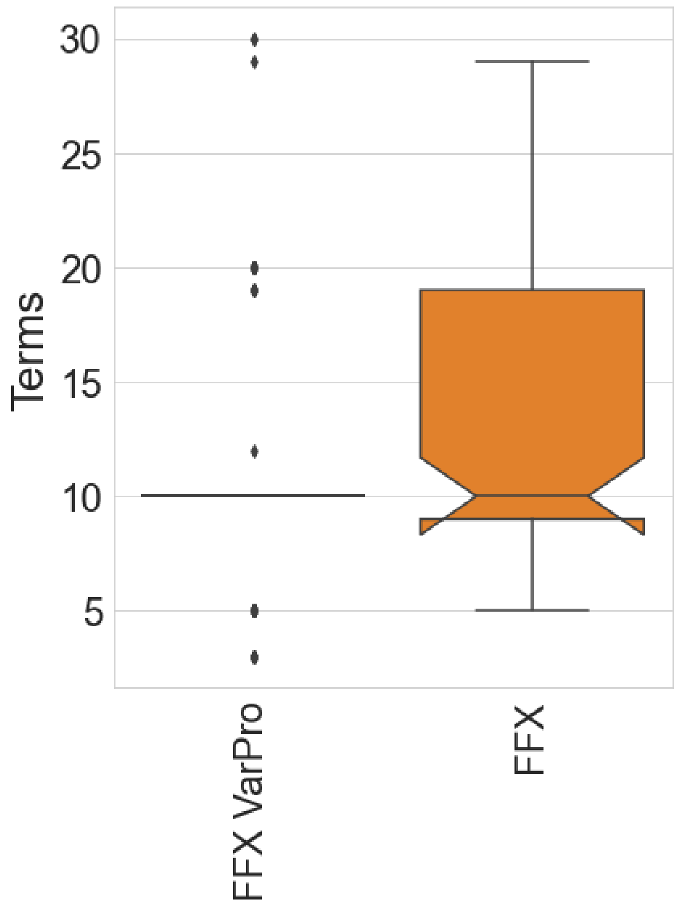
Results – Runtime



cf. Michael Kommenda. "Parameter Identification for Symbolic Regression using Nonlinear Least Squares." Genetic Programming and Evolvable Machines 21.3. 2020
cf. Orzechowski, Patryk, William La Cava, and Jason H. Moore. "Where are we now? A large benchmark study of recent symbolic regression methods." Proceedings of the Genetic and Evolutionary Computation Conference. 2018.

Results – Complexity





Conclusion and Outlook

- Higher Accuracy and/or less complex Models
- Much higher Runtime

→Implementation of more nonlinear functions
→More robustness against overfitting