A Unified Framework for Deep Symbolic Regression

Thirty-sixth Conference on Neural Information Processing Systems

Tue Nov 29th through Dec 1st, 2022. New Orleans, Louisiana, US

NEURAL INFORMATION PROCESSING SYSTEMS

Mikel Landajuela, Chak Lee, Jiachen Yang, Ruben Glatt, Claudio P. Santiago, Ignacio Aravena, Terrell N. Mundhenk, Garrett Mulcahy, Brenden K. Petersen

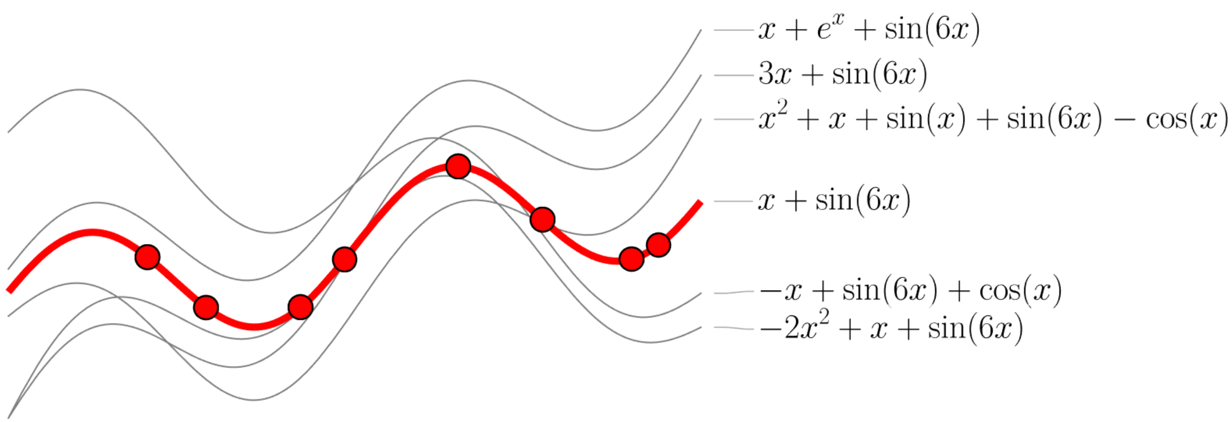
LLNL-PRES-841546-DRAFT

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Lawrence Livermore National Laboratory

Symbolic Regression: A classical Problem

Given a dataset (X, y), where each point $X_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$, find an analytic expression $f : \mathbb{R}^n \to \mathbb{R}$ such that $f(X_i) \approx y_i$



- Symbolic Regression (SR) leads to interpretable models with high performance and generalizability, even in the small dataset regime Broløs et al., 2021; Wilstrup et al., 2021
- SR has received lot of attention in recent years Cranmer et al., 2020; Udrescu et al., 2020; Petersen et al., 2021; Landajuela et al., 2021; Biggio et al., 2021 ; Kamienny et al., 2022 ; …

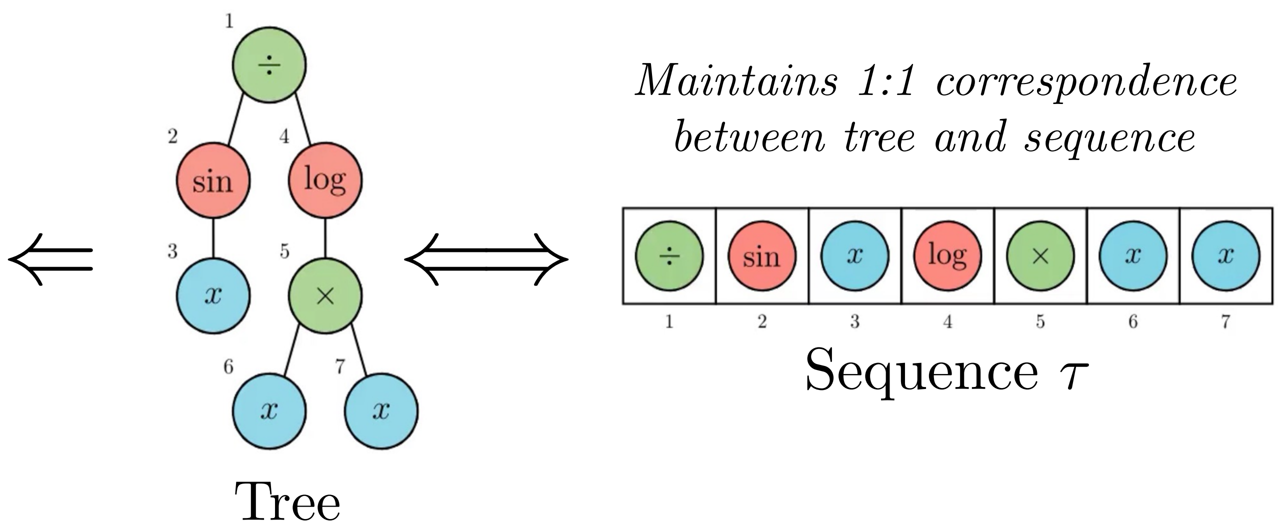
Symbolic Regression as Discrete Optimization

- Using expression trees, the problem becomes a discrete optimization one:

$$\underset{n \leq N, \tau_1, \ldots, \tau_n}{\arg\max} \left[R(\mathtt{ET}(\tau_1, \ldots, \tau_n)) \right] \text{ with } \tau_i \in \mathcal{L} = \{+, \ldots, \sin, \ldots, x_1 \ldots\}$$

$$\frac{\sin x}{\log x^2}$$

Expression



- Exponentially large search space $|\mathcal{L}|^N$. SR is NP-hard (Virgolin et al., 2022), i.e., the search for the best solution can be intractable.

Solution Strategies for Symbolic Regression

- Over the last few years, there are now several quite different approaches to SR:

  - **Problem Simplification**
    Udrescu et al., 2019 and 2020
    Exploits (X, y) data to simplify a SR problem into lower-dimensional sub-problems.

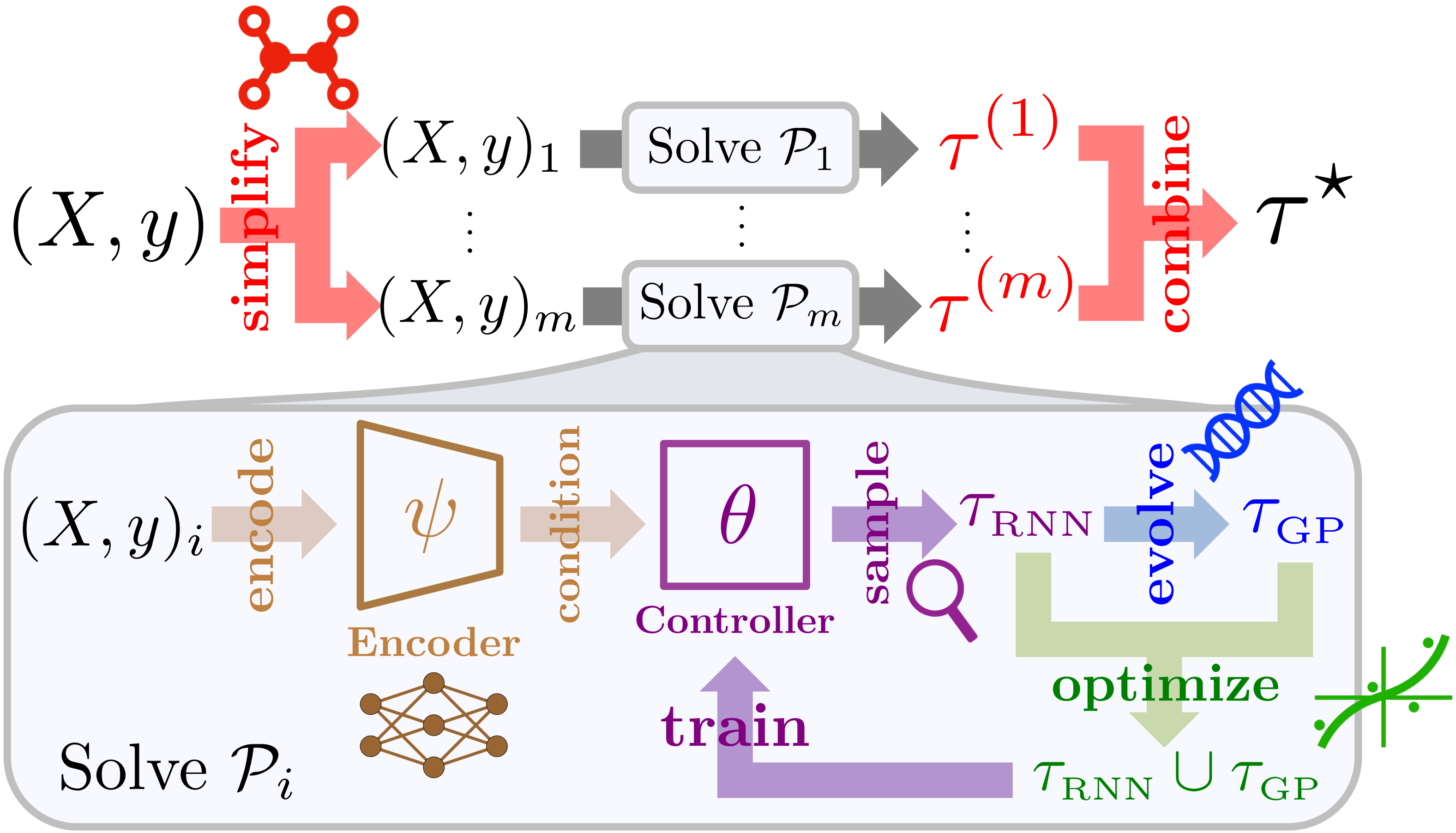
  - **Neural-guided Search**
    Bello et al., 2017; Petersen et al., 2021
    Neural network learns to search over time, with the ability to incorporate in situ constraints.

  - **Genetic Programming**
    Koza, 1994; Mundhenk et al., 2021; …
    Rapidly explores the search space via genetic operators.

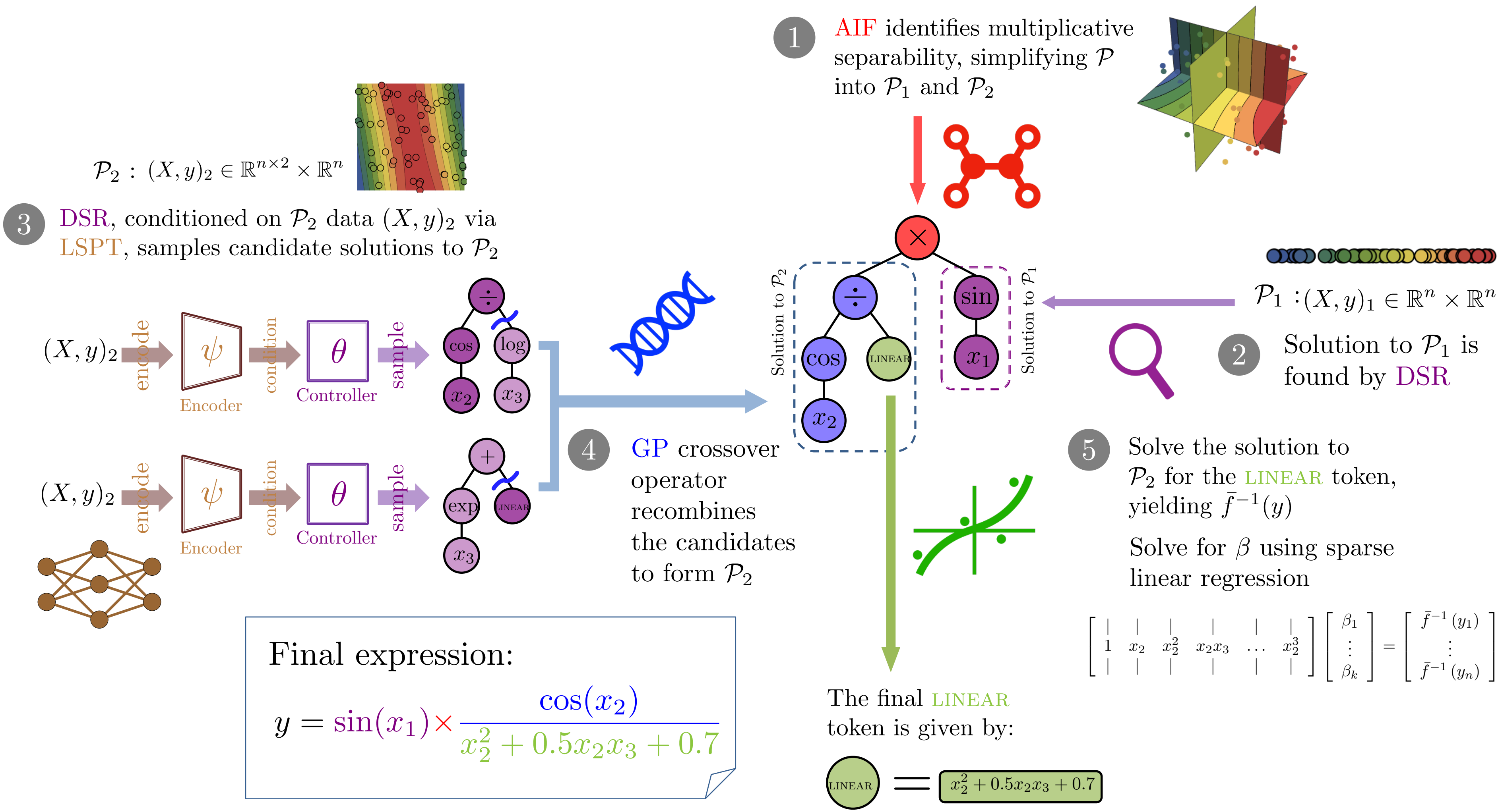
  - **Large Scale Pre-training**
    Biggio et al., 2021; Kamienny et al., 2022; …
    Leverages big data, learning from many other problems by conditioning on the (X, y) data.

  - **Linear Regression**
    Legendre,1805; Brunton et al., 2016; …
    Quickly learn sparse coefficients of a linear combination of basis functions.

uDSR: A Unified framework for Deep Symbolic Regression

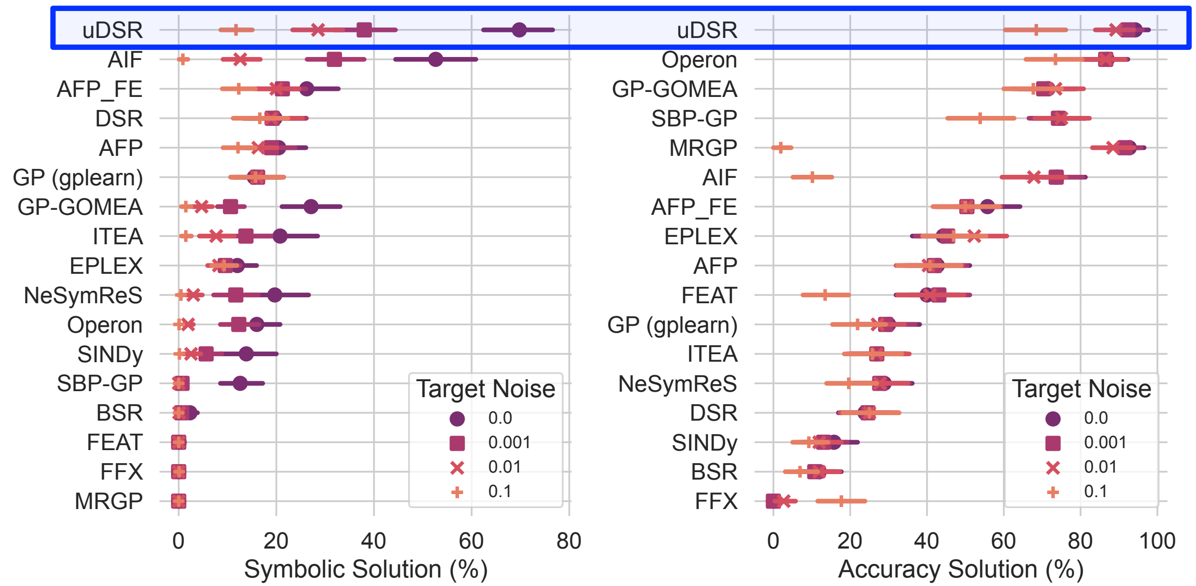


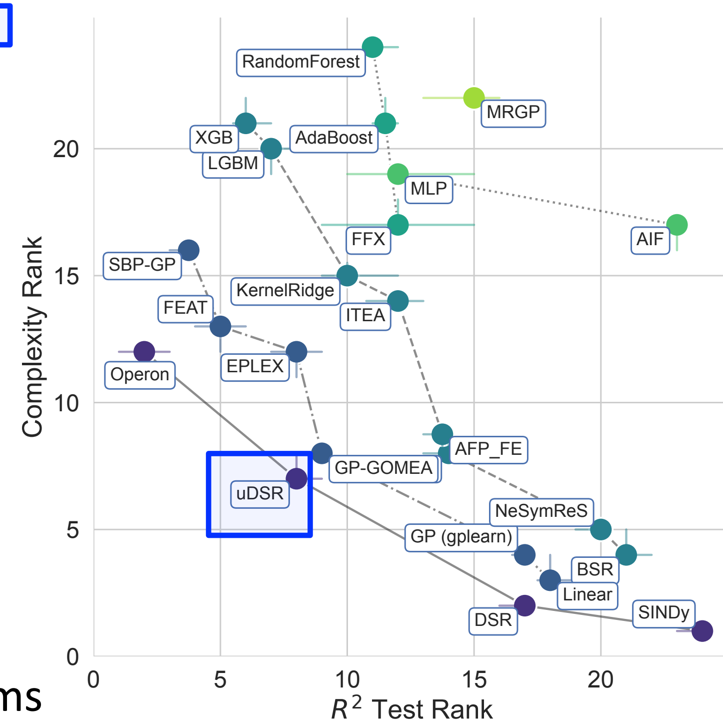
uDSR: A Unified framework for Deep Symbolic Regression



Results on SRBench

- Benchmarking using the open-source pipeline SRBench (La Cava et al.,2021) (252 datasets from PMLB):





- uDSR outperforms all other 14 benchmarked methods in symbolic and accuracy recovery for ground-truth problems

- uDSR falls on the Pareto frontier (accuracy-complexity) on black-box SR problems