

# 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

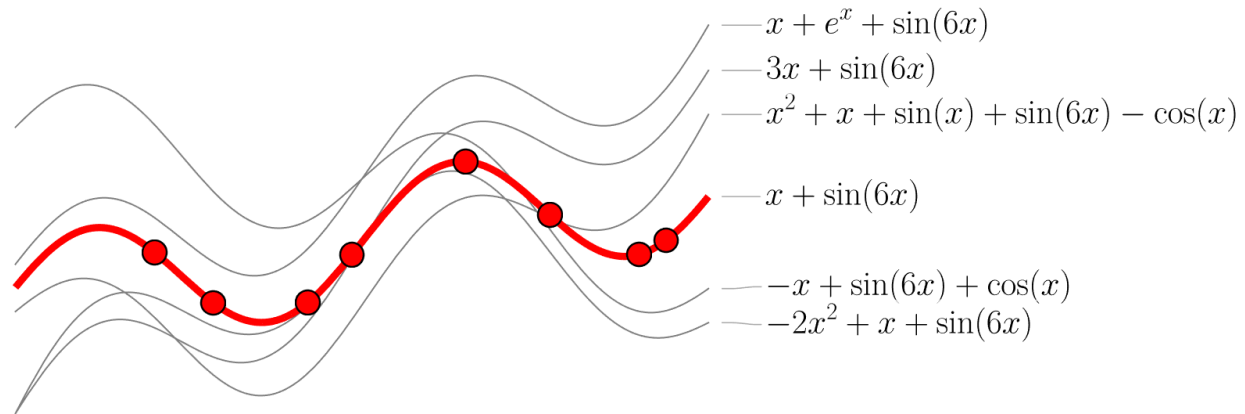


**Mikel Landajuela**, Chak Lee, Jiachen Yang, Ruben Glatt, Claudio P. Santiago, Ignacio Aravena, Terrell N. Mundhenk, Garrett Mulcahy, Brenden K. Petersen



# 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  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  such that  $f(X_i) \approx y_i$

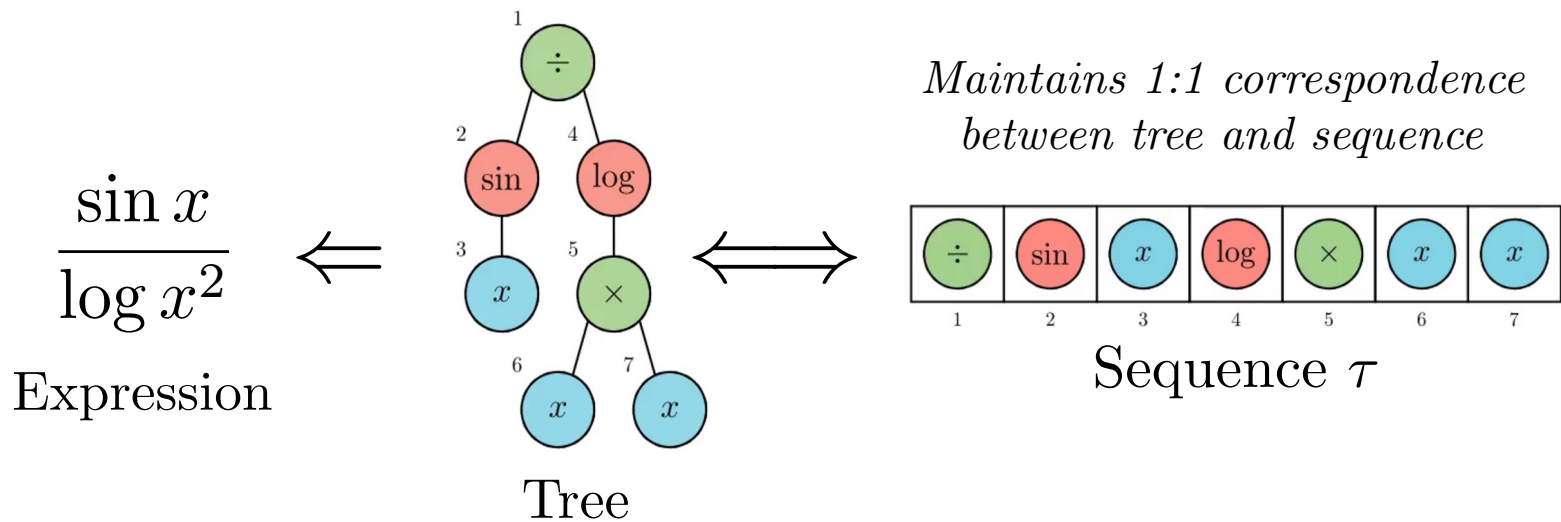


- Symbolic Regression (SR) leads to **interpretable** models with **high performance** and **generalizability**, even in the **small dataset** regime Broløfs et al., 2021; Wilstrup et al., 2021
- SR has received **lot of attention** in recent years Cranmer et al., 2020; Udrescu et al., 2020; Petersen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

$$\arg \max_{n \leq N, \tau_1, \dots, \tau_n} [R(\text{ET}(\tau_1, \dots, \tau_n))] \text{ with } \tau_i \in \mathcal{L} = \{+, \dots, \sin, \dots, x_1 \dots\}$$



- 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; ...

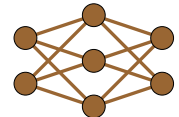
Rapidly **explores** the search space via genetic operators.



- 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.



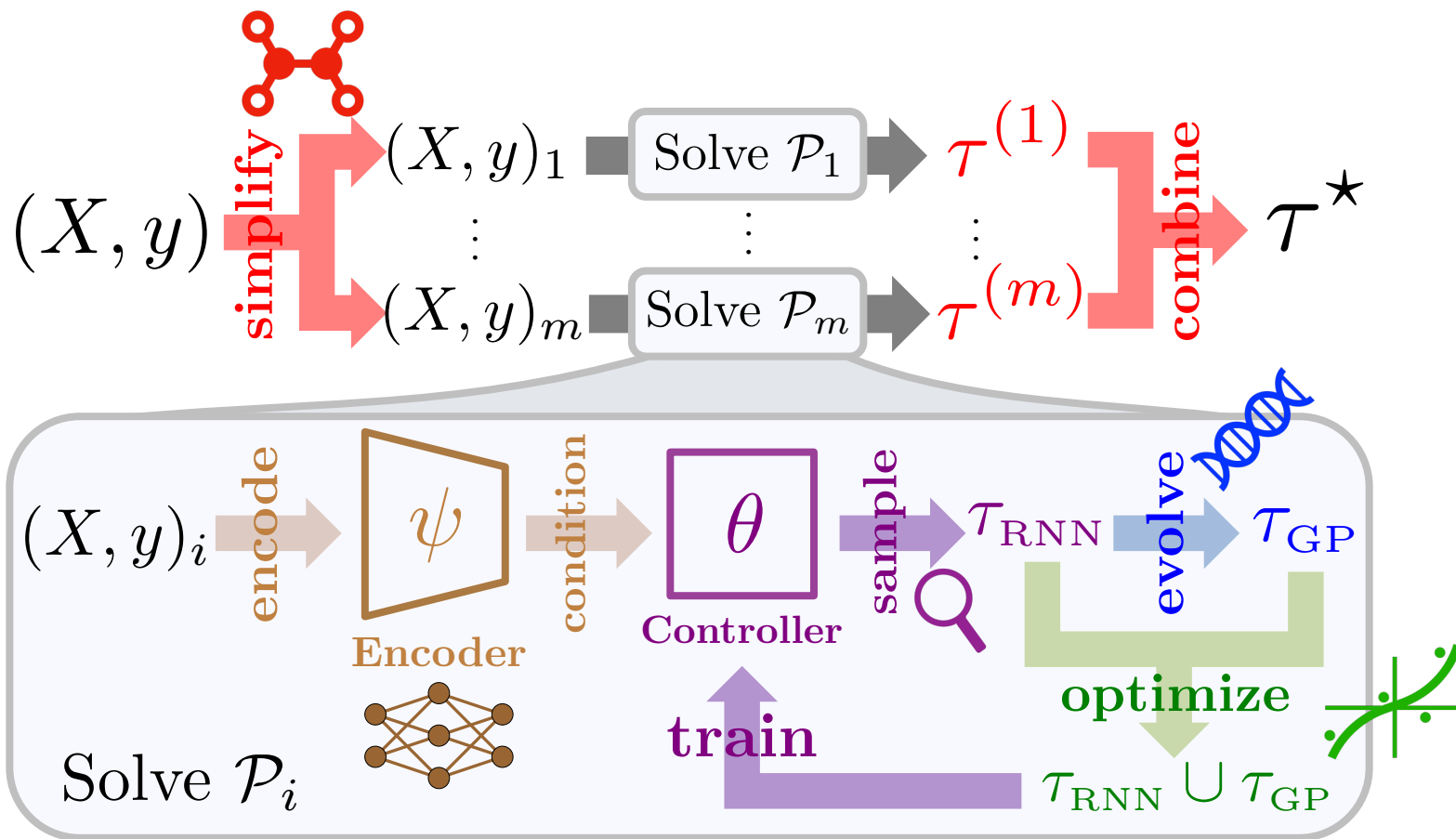
- Linear Regression

Legendre, 1805; Brunton et al., 2016; ...

Quickly learn sparse coefficients of a linear **combination of basis** functions.

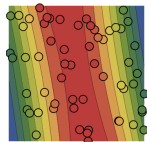


# uDSR: A Unified framework for Deep Symbolic Regression

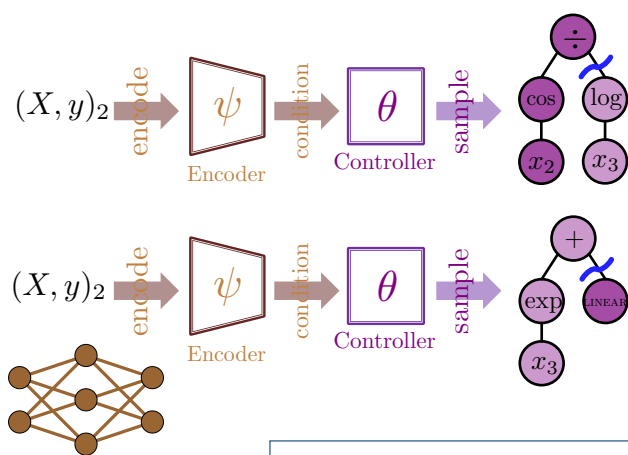


# uDSR: A Unified framework for Deep Symbolic Regression

$$\mathcal{P}_2 : (X, y)_2 \in \mathbb{R}^{n \times 2} \times \mathbb{R}^n$$



- 3 DSR, conditioned on  $\mathcal{P}_2$  data  $(X, y)_2$  via LSPT, samples candidate solutions to  $\mathcal{P}_2$

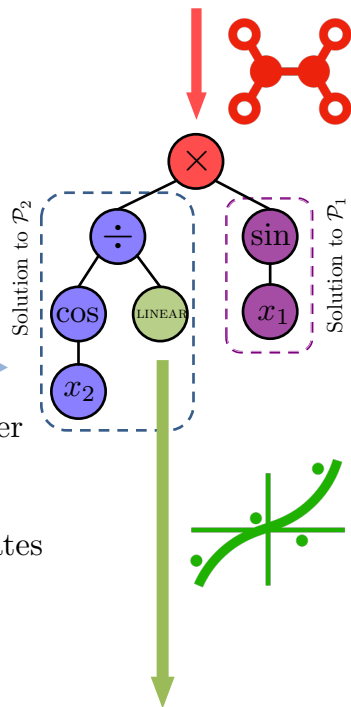
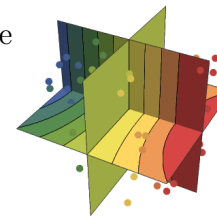


- 4 GP crossover operator recombines the candidates to form  $\mathcal{P}_2$

Final expression:

$$y = \sin(x_1) \times \frac{\cos(x_2)}{x_2^2 + 0.5x_2x_3 + 0.7}$$

- 1 AIF identifies multiplicative separability, simplifying  $\mathcal{P}$  into  $\mathcal{P}_1$  and  $\mathcal{P}_2$



$$\mathcal{P}_1 : (X, y)_1 \in \mathbb{R}^n \times \mathbb{R}^n$$



- 2 Solution to  $\mathcal{P}_1$  is found by DSR

- 5 Solve the solution to  $\mathcal{P}_2$  for the LINEAR token, yielding  $\bar{f}^{-1}(y)$   
Solve for  $\beta$  using sparse linear regression

$$\begin{bmatrix} | & | & | & | & | & | \\ 1 & x_2 & x_2^2 & x_2x_3 & \dots & x_2^3 \\ | & | & | & | & | & | \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} \bar{f}^{-1}(y_1) \\ \vdots \\ \bar{f}^{-1}(y_n) \end{bmatrix}$$

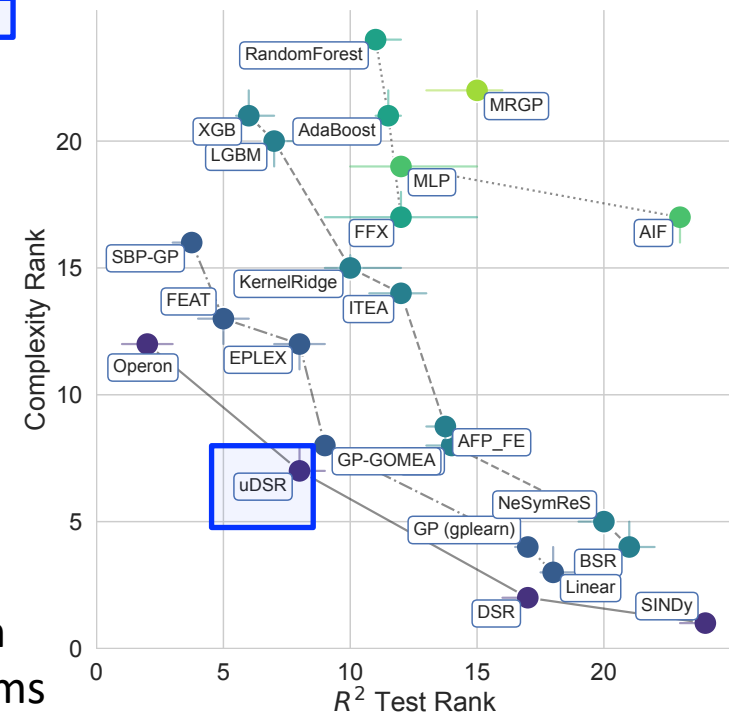
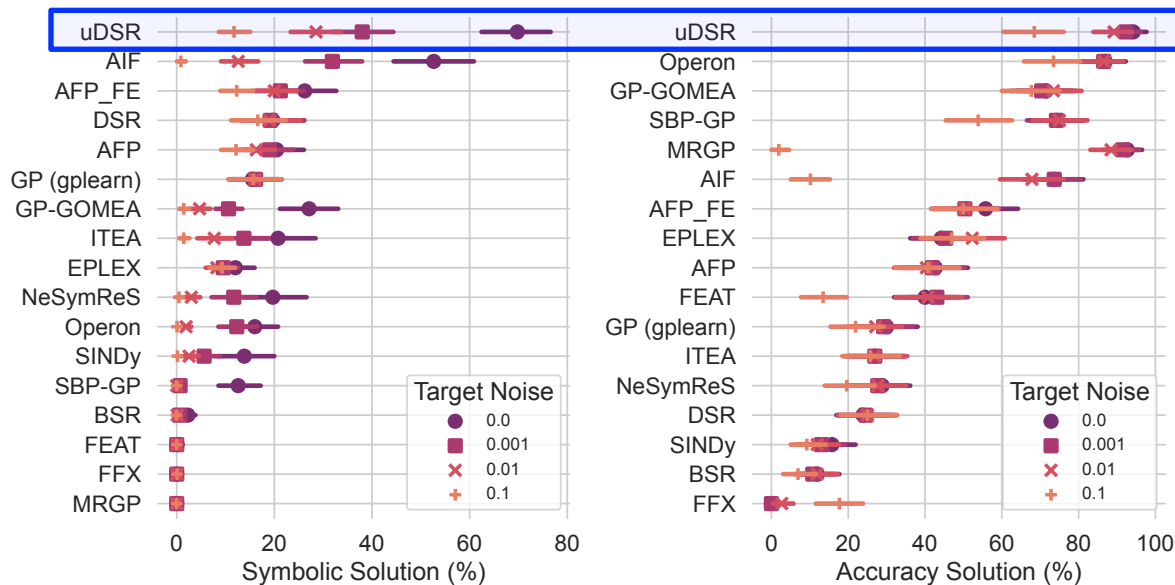
The final LINEAR token is given by:

$$\text{LINEAR} = x_2^2 + 0.5x_2x_3 + 0.7$$



# 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



This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.