Unraveling the Complexity of Multivariate Systems with Symbolic Regression

Giorgio Morales

Gianforte School of Computing
Montana State University

MONTANA STATE UNIVERSITY

NISL
Numerical Intelligent Systems Laboratory

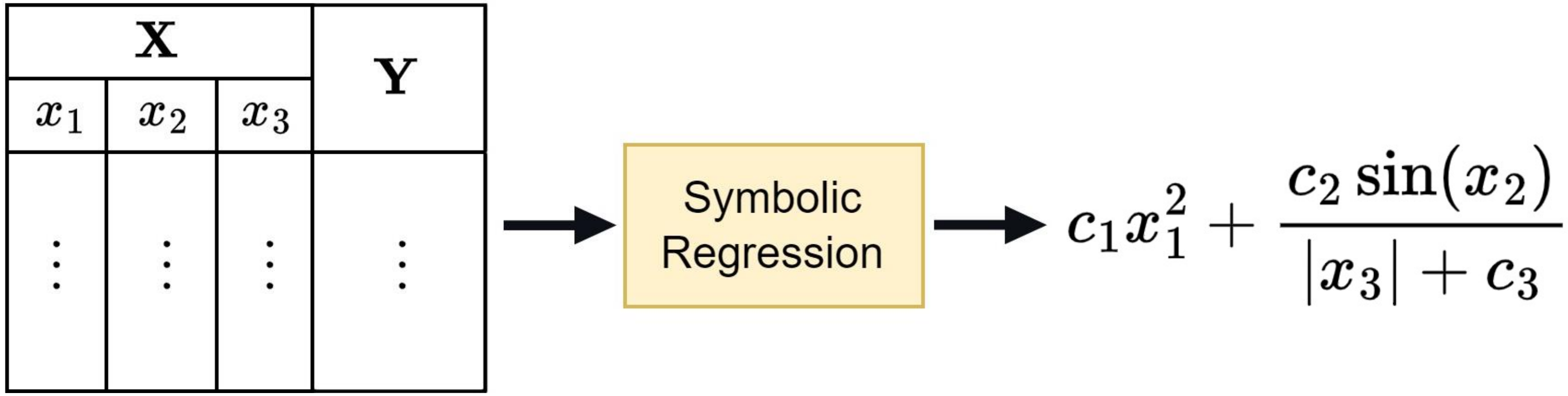
Equations are Interpretable

$$F = G\frac{m_1 m_2}{r^2}$$

- F is proportional to m_1
- F is proportional to m_2
- F is inversely proportional to the square of distance r

Introduction

- “Black-box” models are not suitable for understanding the underlying processes
- A goal of science is to discover causal explanations for the observable world
- Symbolic regression (SR) aims to identify underlying relationships in the studied phenomena



Multivariate Symbolic Regression

- Consider a multivariate system $y = f(\mathbf{x}) = f(x_1, \dots, x_t)$ ($\mathbf{x} \in R^t$)
- $f(\cdot)$ is the underlying function of the system
- We assume f can be expressed as a mathematical expression that uses:
  - Unary operators: $\sin, \cos, \log$, etc.
  - Binary operators: $+, -, *, /$
- Goal: Given a dataset of observations, generate a mathematical expression $\tilde{f}$ that approximates f
- The skeleton of $3x^2 + e^{2x} - 4$ is $c_1x^2 + e^{c_2x} + c_3$

SR vs. Function Fitting

- Step 1: Assume the function has a prior form:

$$y = w_2 \, \sigma(w_1 x_1 + b_1) + b_2$$

- Step 2: Find coefficients that best fit the data (e.g., gradient descent)

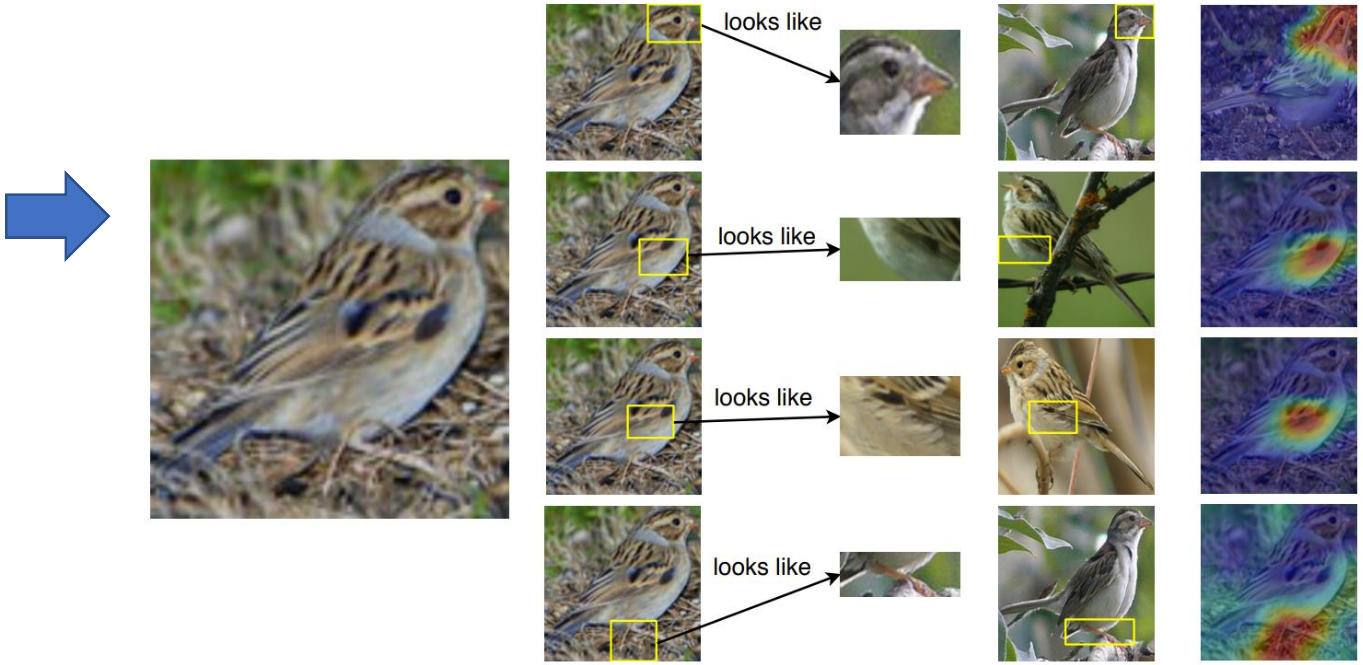
- **Q: Why SR is harder than function fitting?**

Problem Definition

- The primary focus of SR methods is to minimize the prediction error
- SR methods that rely on genetic programming (GP) suffer from slow computation
- GP-based SR methods do not consider past experiences
- SR methods often fail to identify the functional form that explains the relationship between each variable and the system's response

Interpretability vs. Explainability

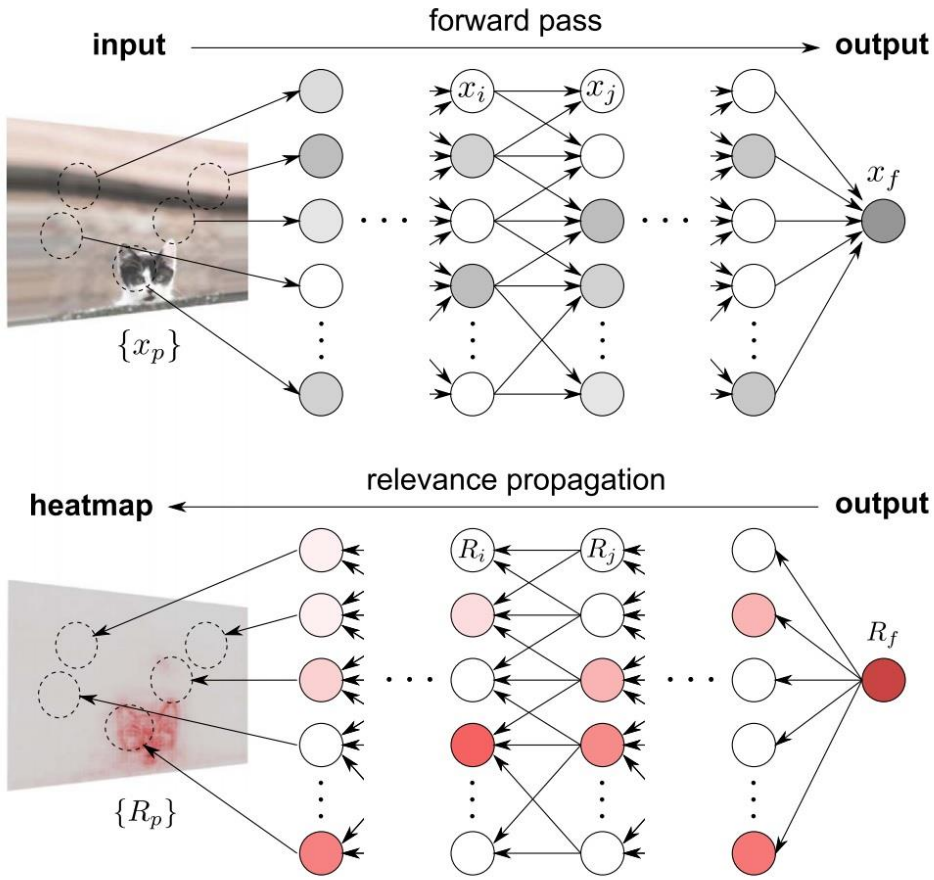
- Interpretability: Allows to identify cause-effect relationships within the system's inputs and outputs.
- Explainability: Associated with the internal logic and mechanisms inside a machine learning system.



This Looks Like That (Chen et al., 2019)

Interpretability vs. Explainability

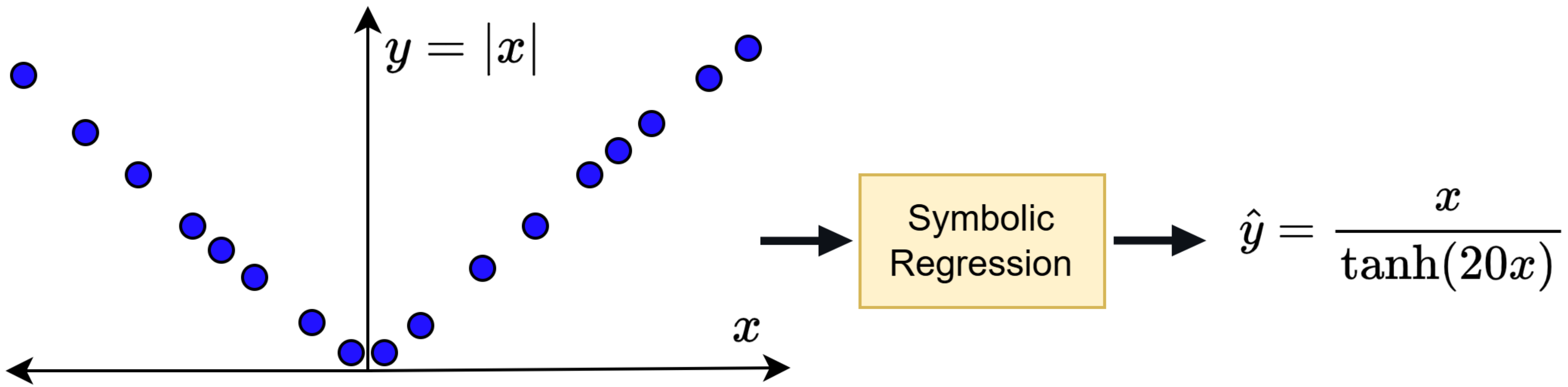
- Interpretability: Allows to identify cause-effect relationships within the system's inputs and outputs.
- Explainability: Associated with the internal logic and mechanisms inside a machine learning system.



Explaining nonlinear classification decisions with deep Taylor decomposition (Montavon et al., 2016)

Problem Definition

- SR models (i.e., equations) are white-box models
- Explainable models: Allow to understand the models' inner mechanisms
- Interpretations: Allow humans to identify cause-effects

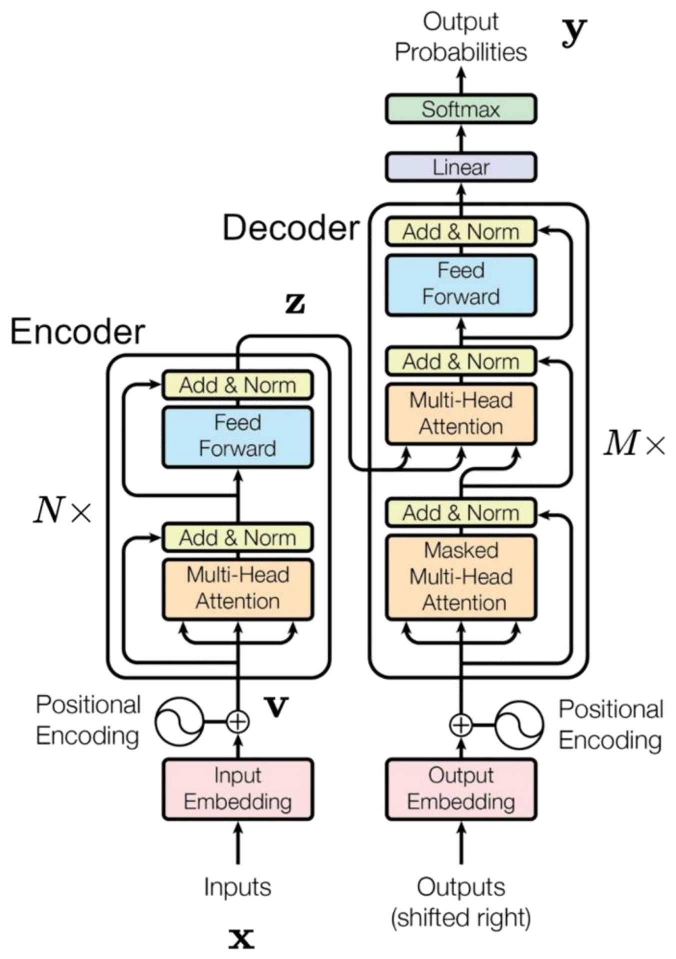


Contributions

- A method that learns univariate symbolic skeletons to explain the functional form between an independent variable and the system's response
- We introduce an SR problem called multi-set symbolic skeleton prediction (MSSP)
- We present a novel transformer network model called "Multi-Set Transformer" to solve the MSSP problem

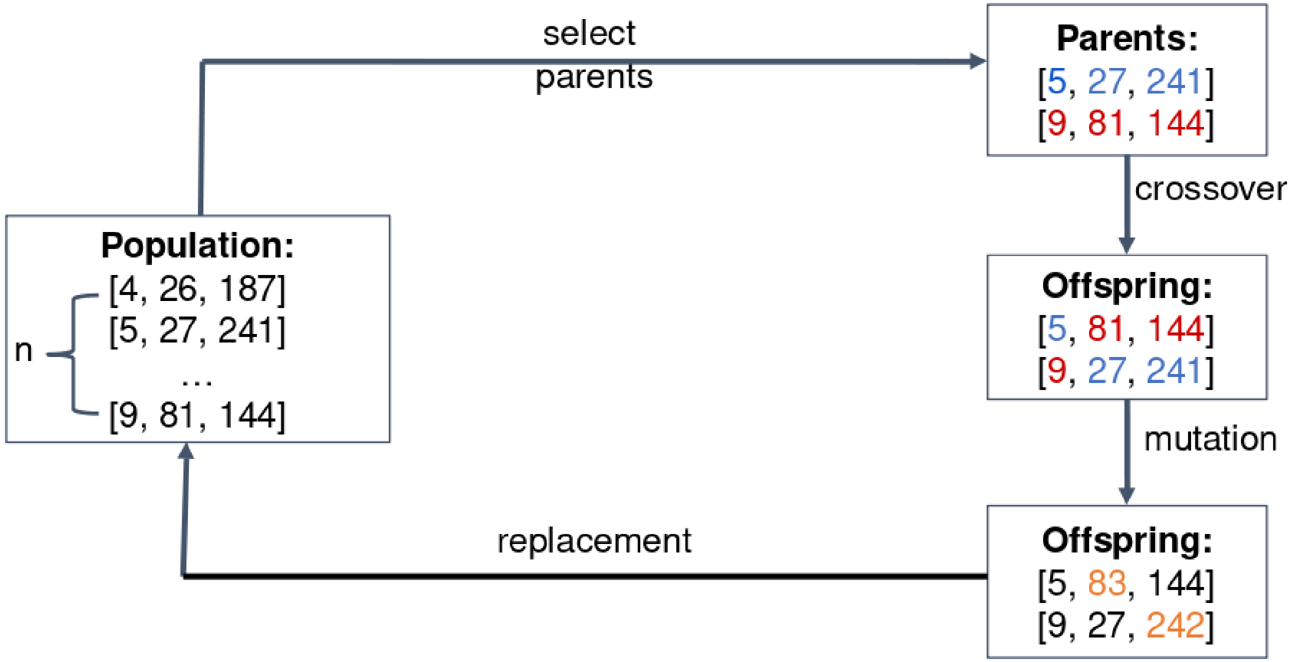
Transformers

- The transformer architecture is designed to process sequential data
- It uses an encoder-decoder architecture
- It leverages the concept of self-attention mechanism



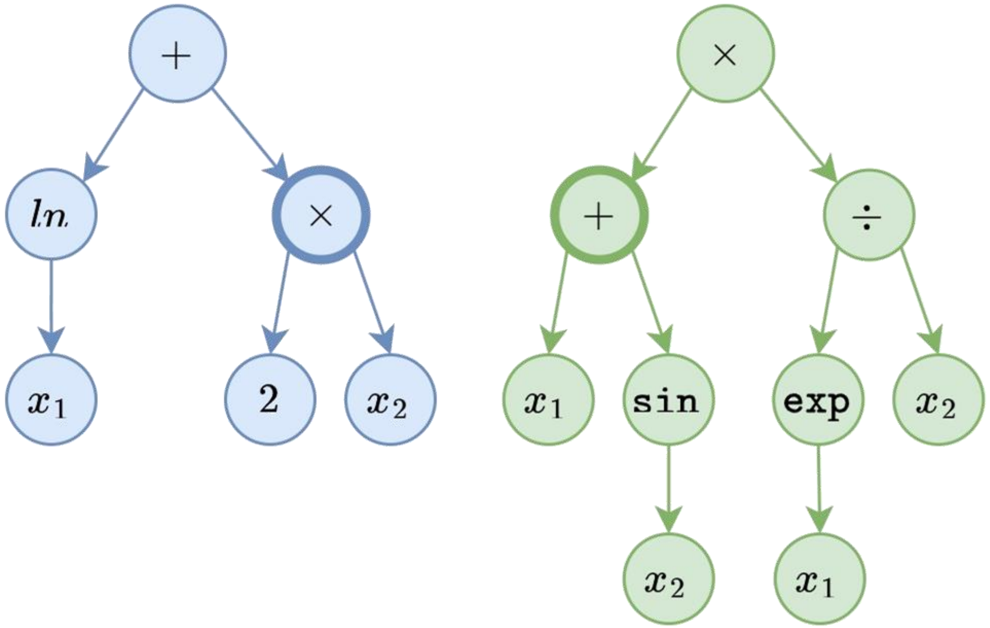
Attention is all you need (Vaswani et al., 2017)

Genetic Algorithms



Multi- and many-objective factored
evolutionary algorithms (Peerlinck, 2023)

Genetic Programming

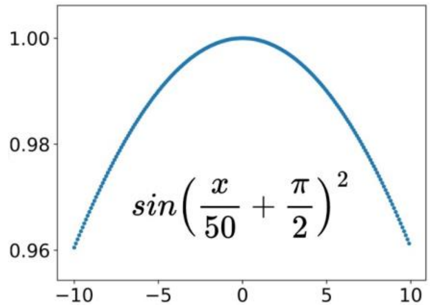


Related Work

- Cramer (2020) presented **PySR**, a GP-based library for SR
- Biggio *et al.* (2021) designed a method called **NeSymReS** using a pre-trained transformer model
- He *et al.* (2022) proposed **TaylorGP** which uses a Taylor polynomial to extract important features
- Kamienny *et al.* (2022) proposed the transformer model **E2E** that estimates the full mathematical expression
- Bendinelli *et al.* (2023) modified the method proposed by Biggio *et al.* to include *a priori* information

Multi-set Symbolic Skeleton Prediction

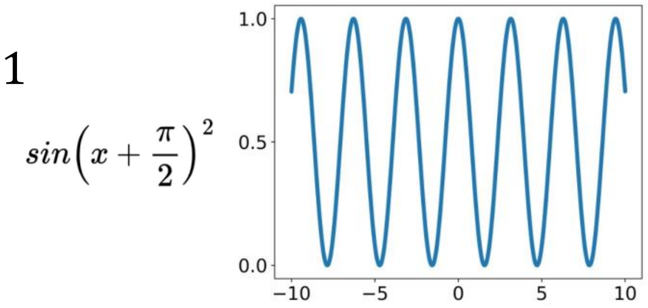
- Consider a system $f(\mathbf{x}) = \sin\left(\frac{x_1}{10\, x_2} + \frac{\pi}{2}\right)^2$
- From a dataset $(\mathbf{X}, \mathbf{Y})$ construct a subset where $x_2 = 5$



Symbolic Skeleton Prediction (SSP)
NeSymReS

$c_1 x_1^2 + c_2$

$x_2 = 0.1$



Symbolic Skeleton Prediction (SSP)
NeSymReS

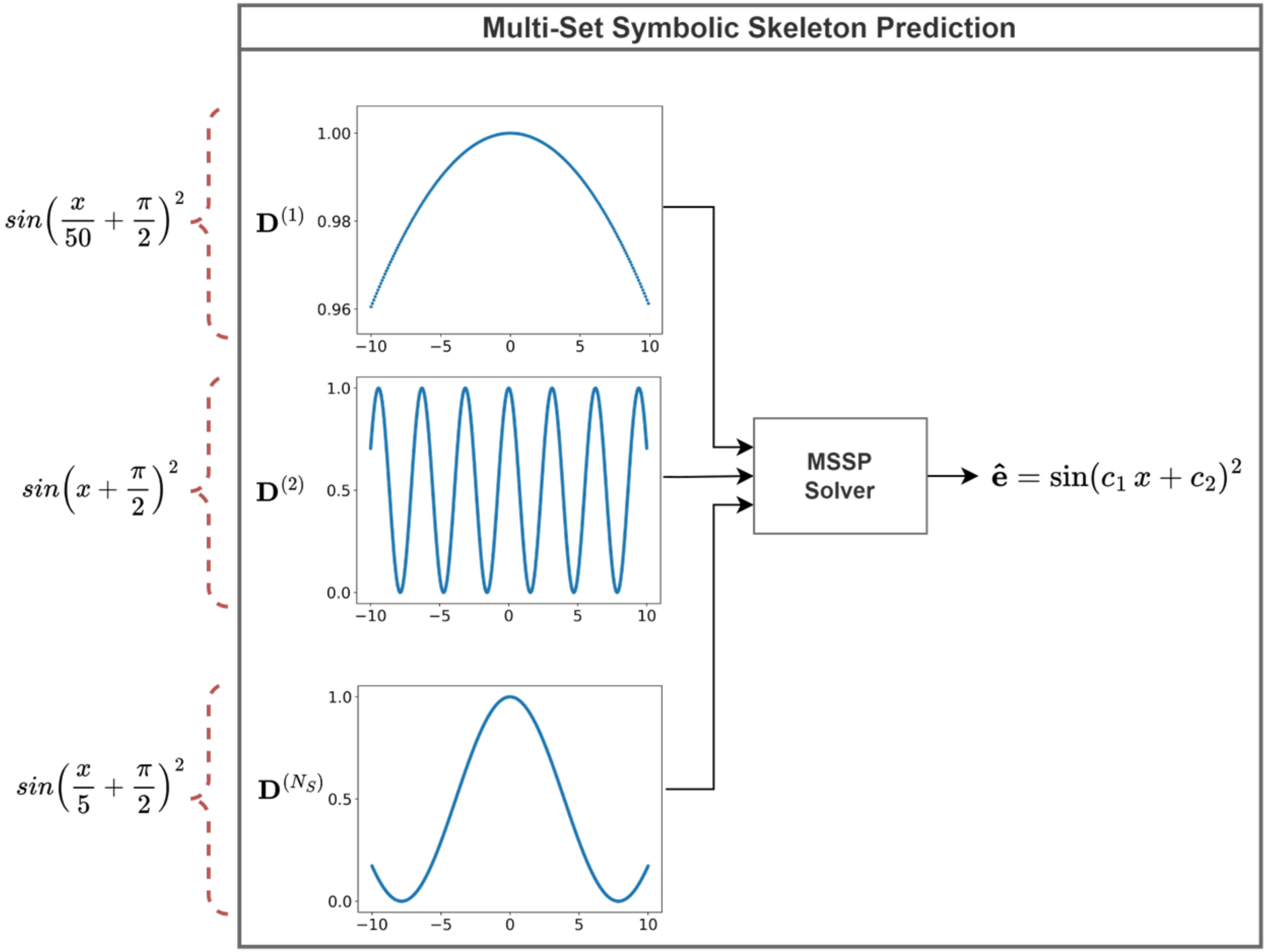
$\sin(c_1 x_1 + c_2)^2$

Multi-set Symbolic Skeleton Prediction

- The SSP problem would benefit from the injection of additional context data
- We propose to process the information from multiple sets simultaneously to produce a symbolic skeleton that is common to all input sets
- We refer to this new problem as multi-set symbolic skeleton prediction (MSSP)

Multi-set Symbolic Skeleton Prediction

$$f(\mathbf{x}) = \sin\left(\frac{x_1}{10\,x_2} + \frac{\pi}{2}\right)^2$$

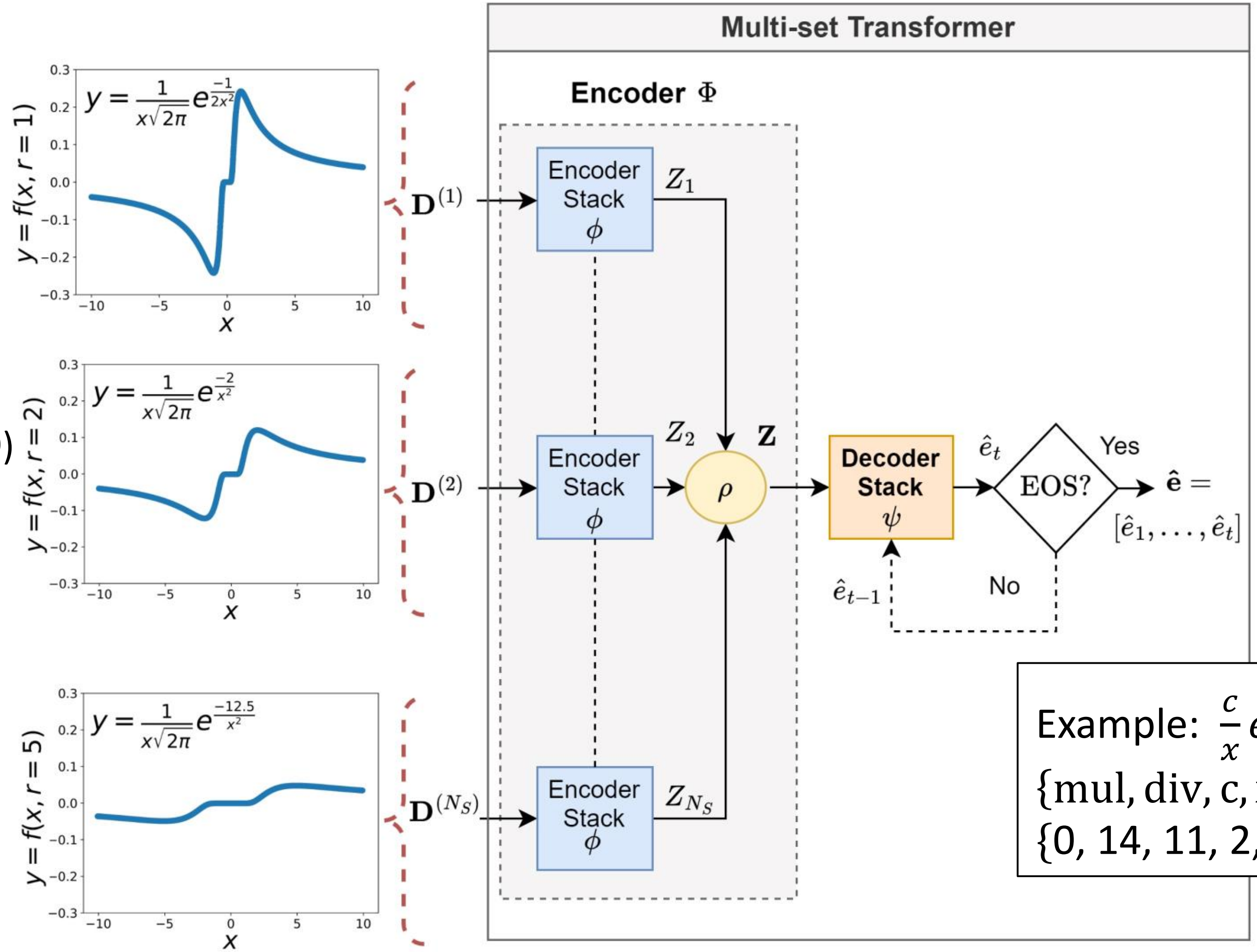


Multi-set Symbolic Skeleton Prediction

- The Transformer processes sequences with fixed orderings (one at a time)
- Sets don't have inherent order. Permutations don't alter their semantics
- We propose a Multi-Set Transformer architecture to solve the MSSP problem
- The Multi-Set Transformer is pre-trained on a large dataset of synthetic expressions

Multi-set Symbolic Skeleton Prediction

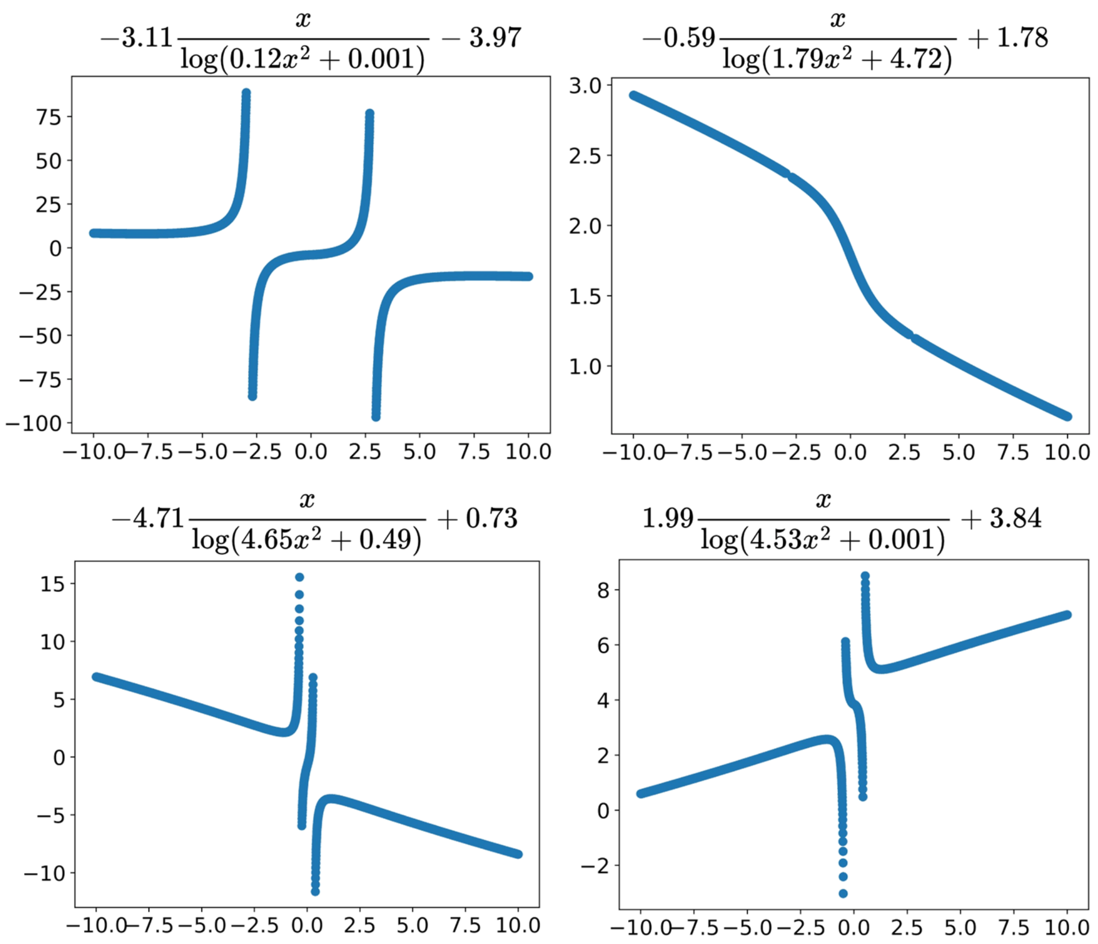
Induced set attention block (ISAB)
"Set Transformer" (Lee et al. 2019)



Example: $\frac{c}{x} e^{\frac{c}{\sqrt{x}}}$ is expressed as
$\{\text{mul}, \text{div}, \text{c}, \text{x}, \text{exp}, \text{div}, \text{c}, \text{square}, \text{x}\} \equiv$
{0, 14, 11, 2, 3, 12, 11, 2, 18, 3, 1}

Multi-set Symbolic Skeleton Prediction

Four sets of input–response pairs generated from the symbolic skeleton $\frac{c_1 x}{\log(c_2 x_2^2 + c_3)} + c_4$



Model Training

- We created a dataset $\mathbf{Q}_1$ of 10^6 expressions for training and 10^5 expressions for validation. They allow up to 1 nested operation

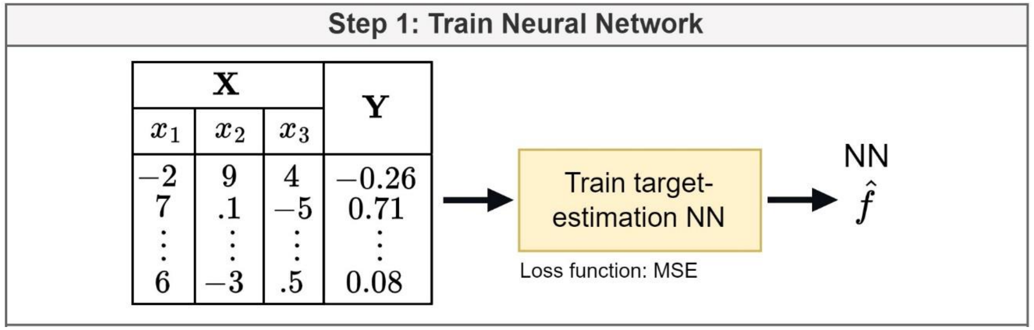
- Architecture details:
  - # of input sets: $N_S = 10$
  - # of input–response pairs in each input set: $n = 3000$
  - # of encoder blocks: $N = 5$
  - # of decoder blocks: $M = 5$

Dataset available at https://huggingface.co/datasets/AnonymousGM/MultiSetTransformerData 🤗

MONTANA STATE UNIVERSITY

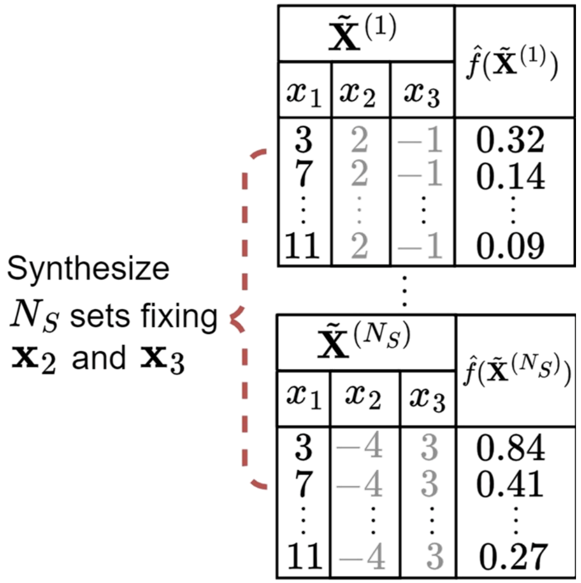
Univariate Symbolic Skeleton Prediction

- f can be approximated based on observed data using a neural network (NN)
- Let $(\mathbf{X}, \mathbf{Y})$ be a dataset, so that $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_{N_R}\}$ and $\mathbf{y} = \{y_1, \dots, y_{N_R}\}$
- A NN $\hat{f}(\cdot)$ is trained to capture the association between $\mathbf{X}$ and $\mathbf{Y}$
- Given an input $\mathbf{x}_\mathrm{j}$, $\hat{y}_j = \hat{f}(\mathbf{x}_\mathrm{j})$ and $\hat{y}_j \approx y_j$



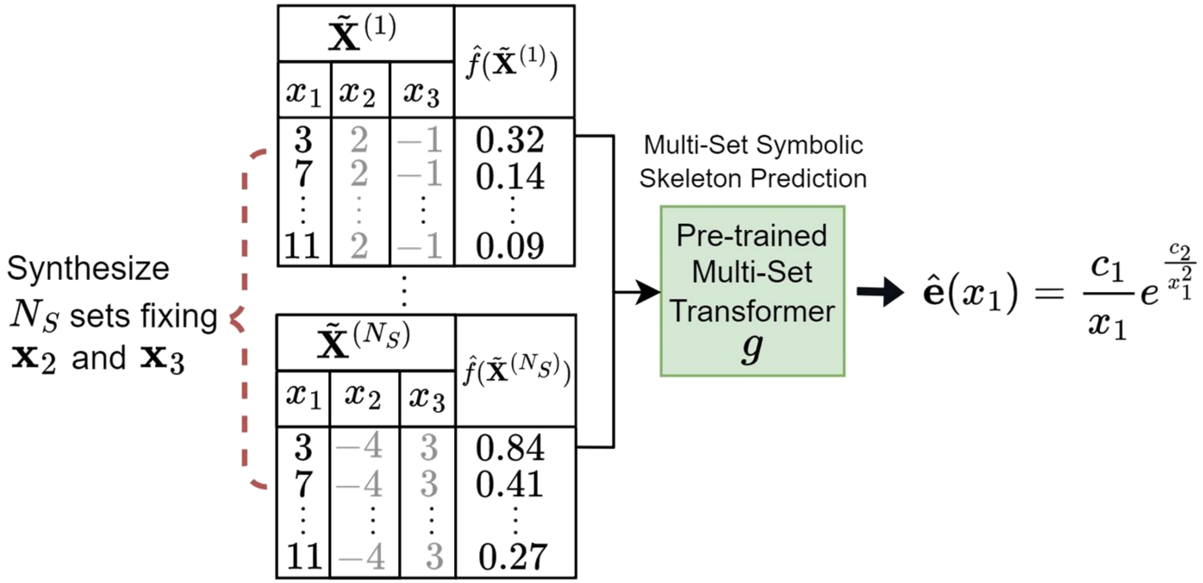
Univariate Symbolic Skeleton Prediction

- Suppose we are currently analyzing the v-th variable, x_v
- We generate N_S synthetic sets and estimate their response with $\hat{f}$



Univariate Symbolic Skeleton Prediction

- Suppose we are currently analyzing the v-th variable, x_v
- We generate N_S synthetic sets and estimate their response with $\hat{f}$
- The collection of N_S sets and their estimated response is fed into the Multi-Set Transformer



Skeleton Performance Evaluation

$$f(x_1, x_2, x_3) = 5.5 + (1 - x_1/4)^2 + \sqrt{x_2 + 10}\,\sin(x_3/5)$$

$$\mathbf{e}(x_1) = (c_1\, x_1 + c_2)^2 + c_3$$

$$\mathbf{e}(x_2) = c_4\sqrt{x_2 \ + c_5} + c_6$$

$$\mathbf{e}(x_3) = c_7 \sin(c_8 x_3) + c_9$$

MONTANA STATE UNIVERSITY

Skeleton Performance Evaluation

$\mathbf{e}(x_1) = \sin(c_1\, x_1 + c_2) + c_3$

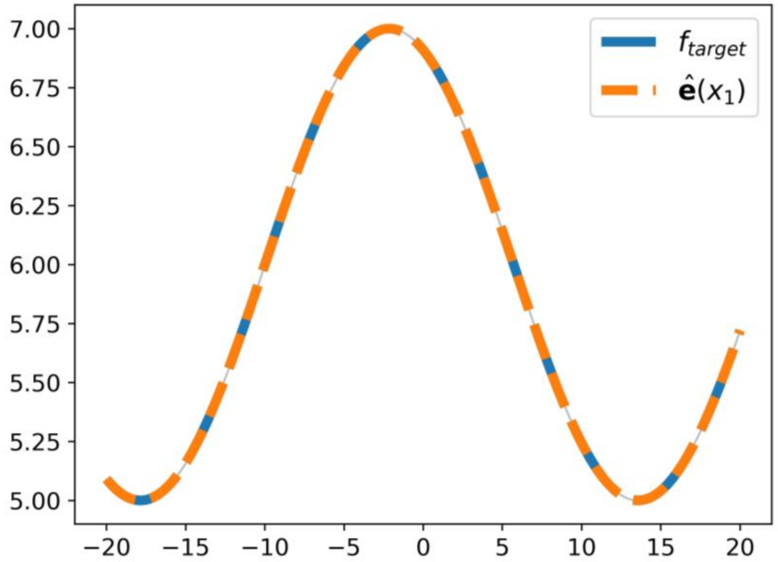
Method 1's result: $\hat{\mathbf{e}}_1(x_1) = c_9 \cos(c_{10}\, x_1 + c_{11}) + c_{12}$

Method 2's result: $\hat{\mathbf{e}}_2(x_1) = c_{13} \sin(c_{14}\, x_1 + c_{15})^4 + c_{16}$

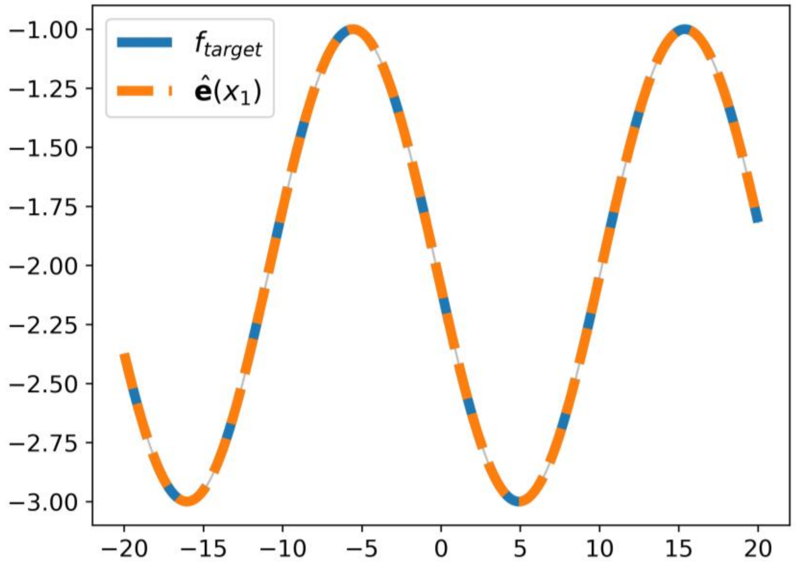
Comparison Is $\hat{\mathbf{e}}_1(x_1)$ equivalent to $\mathbf{e}(x_1)$?

Sample coefficients of $\mathbf{e}(x_1)$:

$f_{target}(x_1) = \sin(0.2\, x_1 + 2) + 6$



$f_{target}(x_1) = \sin(-0.3\, x_1 - 0.1) - 2$



Fit coefficients of $\hat{\mathbf{e}}_1(x_1)$

$1 \cos(-0.2\, x_1 - 0.43) + 6$

$1 \cos(0.3\, x_1 + 1.67) - 2$

MONTANA STATE UNIVERSITY

Skeleton Performance Evaluation

$$\mathbf{e}(x_1) = \sin(c_1\, x_1 + c_2) + c_3$$

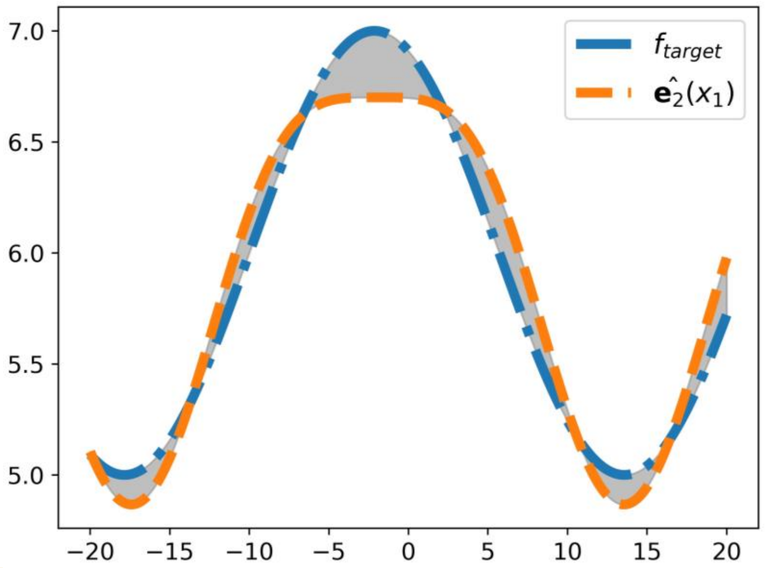
Method 1's result: $\hat{\mathbf{e}}_1(x_1) = c_9 \cos(c_{10}\, x_1 + c_{11}) + c_{12}$

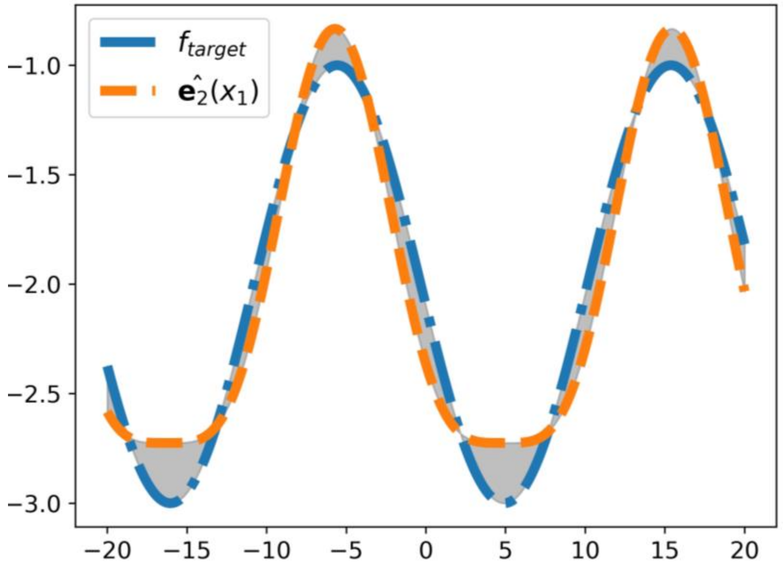
Method 2's result: $\hat{\mathbf{e}}_2(x_1) = c_{13} \sin(c_{14}\, x_1 + c_{15})^4 + c_{16}$

Comparison Is $\hat{\mathbf{e}}_2(x_1)$ equivalent to $\mathbf{e}(x_1)$?

Sample coefficients of $\mathbf{e}(x_1)$: $f_{target}(x_1) = \sin(0.2\, x_1 + 2) + 6$ $f_{target}(x_1) = \sin(-0.3\, x_1 - 0.1) - 2$

Fit coefficients of $\hat{\mathbf{e}}_2(x_1)$





$-1.83 \sin(0.101\, x_1 - 9.23)^4 + 6.7$

$1.89 \sin(0.149\, x_1 - 3.86)^4 - 2.72$

MONTANA STATE UNIVERSITY

Eq.	Underlying Equation	Domain range
E1	$(3.0375x_1x_2 + 5.5\sin(9/4(x_1 - 2/3)(x_2 - 2/3)))/5$	$[-5, 5]^2$
E2	$5.5 + (1 - x_1/4)^2 + \sqrt{x_2 + 10}\sin(x_3/5)$	$[-10, 10]^2$
E3	$(1.5e^{1.5x_1} + 5\cos(3x_2))/10$	$[-5, 5]^2$
E4	$((1 - x_1)^2 + (1 - x_3)^2 + 100(x_2 - x_1^2)^2 + 100(x_4 - x_3^2)^2)/10000$	$[-5, 5]^4$
E5	$\sin(x_1 + x_2x_3) + e^{1.2x_4}$	$x_1 \in [-10, 10], x_2 \in [-5, 5],$ $x_3 \in [-5, 5], x_4 \in [-3, 3]$
E6	$\tanh(x_1/2) + \lvert x_2\rvert\cos(x_3^2/5)$	$[-10, 10]^3$
E7	$(1 - x_2^2)/\sin(2\pi x_1 + 1.5)$	$[-5, 5]^2$
E8	$x_1^4/(x_1^4 + 1) + x_2^4/(x_2^4 + 1)$	$[-5, 5]^2$
E9	$\log(2x_2 + 1) - \log(4x_1^2 + 1)$	$[0, 5]^2$
E10	$\sin(x_1e^{x_2})$	$x_1 \in [-2, 2], x_2 \in [-4, 4]$
E11	$x_1\log(x_2^4)$	$[-5, 5]^2$
E12	$1 + x_1\sin(1/x_2)$	$[-10, 10]^2$
E13	$\sqrt{x_1}\log(x_2^2)$	$x_1 \in [0, 20], x_2 \in [-5, 5]$

Experimental Results

Method	Learned Expressions
PySR	$0.4094\, x_3 + \big\lvert \lvert x_1 - 3.5101 \rvert - 1.952 \big\rvert + 4.1115$
TaylorGP	$-0.497\, x_1 \; + \; 0.001498\, x_2 \; + \; 0.3868\, x_3 \; + \; 8.619$
NeSymRes	$-x_1 \; + \; 0.4029\, x_3 \; + \; \exp(\exp(-0.000755\, x_2)) \; + \; 5.8753$

Equations obtained by other methods for problem E2

E2: $5.5 + (1 - x_1/4)^2 + \sqrt{x_2 + 10}\,\sin(x_3/5)$

Experimental Results

Method	x_1	x_2	x_3
PySR	$c_1 + \|c_2 + \|c_3 + x_1\|\|$	c_1	$c_1 + c_2 x_3$
TaylorGP	$c_1 + c_2 x_1$	$c_1 + c_2 x_2$	$c_1 + c_2 x_3$
NeSymRes	$c_1 + c_2 x_1$	$c_1 + \exp(\exp(x_2))$	$c_1 + c_2 x_3$
E2E	$c_1 + c_2 x_1 + c_3 (c_4 + c_5 x_1)^2$	$c_1 + c_2 (c_3 + c_4 x_2)$	$c_1 + c_2 x_3 + c_3 (c_4 + c_5 \cos(c_6 + c_7 x_3))$
MST	$c_1 + c_2 (c_3 + c_4 x_1)^2$	$c_1 \sqrt{c_2 x_2 + c_3} + c_4$	$c_1 + c_2 \sin(c_3 x_3 + c_4)$
Target $\mathbf{e}(x)$	$c_1 + (c_2 + c_3 x_1)^2$	$c_1 \sqrt{x_2 + c_2} + c_3$	$c_1 + c_2 \sin(c_3 x_3)$

Comparison of skeleton prediction results for problem E2

E2: $5.5 + (1 - x_1/4)^2 + \sqrt{x_2 + 10} \sin(x_3/5)$

MONTANA STATE UNIVERSITY

Experimental Results

Eq.	Var.	PySR	TaylorGP	NeSymReS	E2E	MST
E1	x_1	1.4 ± 0.8	1.4 ± 0.8	0.9 ± 0.7	$\mathbf{0.2 \pm 0.4}$	$\mathbf{0.01 \pm 0.02}$
	x_2	1.5 ± 0.9	1.5 ± 0.9	1.3 ± 0.8	1.5 ± 0.9	$\mathbf{0 \pm 0}$
E2	x_1	303.5 ± 167.3	310.0 ± 170.1	310.0 ± 170.1	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$
	x_2	5.4 ± 5.0	4.2 ± 5.3	4.6 ± 5.0	4.2 ± 5.3	$\mathbf{0.02 \pm 0.03}$
	x_3	1.7 ± 1.0	1.7 ± 1.0	1.7 ± 1.0	$\mathbf{0.01 \pm 0.02}$	$\mathbf{0 \pm 0}$
E3	x_1	$2 \times 10^{12} \pm 5 \times 10^{12}$	939.4 ± 1419.9	1.9 ± 1.2	$\mathbf{0.8 \pm 1.8}$	$\mathbf{0.8 \pm 1.8}$
	x_2	1.3 ± 1.0	1.3 ± 1.0	0.8 ± 0.8	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$
E4	x_1	4576.2 ± 2695.7	4581.5 ± 2697.4	—	$\mathbf{2.3 \pm 3.6}$	$\mathbf{1.1 \pm 0.7}$
	x_2	79.6 ± 41.3	80.2 ± 40.8	—	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$
	x_3	3995.5 ± 2815.6	4304.6 ± 2843.7	—	$\mathbf{2.0 \pm 4.0}$	$\mathbf{1.0 \pm 0.9}$
	x_4	74.5 ± 48.0	75.5 ± 47.0	—	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$
E5	x_1	0.6 ± 0.05	0.6 ± 0.05	—	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$
	x_2	1.5 ± 1.0	1.5 ± 1.0	—	1.5 ± 1.0	$\mathbf{0 \pm 0}$
	x_3	0.6 ± 0.05	0.6 ± 0.05	—	0.6 ± 0.05	$\mathbf{0 \pm 0}$
	x_4	2.7 ± 1.2	487.4 ± 461.9	—	$\mathbf{0.6 \pm 0.8}$	$\mathbf{0.6 \pm 0.8}$
E6	x_1	0.8 ± 0.08	0.8 ± 0.08	0.3 ± 0.01	0.04 ± 0	$\mathbf{0 \pm 0}$
	x_2	16.8 ± 12.2	16.8 ± 12.2	13.8 ± 11.1	1.3 ± 0.9	$\mathbf{0 \pm 0}$
	x_3	2.9 ± 1.3	1.9 ± 0.7	1.6 ± 0.9	1.6 ± 0.9	$\mathbf{0 \pm 0}$

Skeleton evaluation performance comparison

Experimental Results

https://github.com/NISL-MSU/MultiSetSR/

Open in Colab

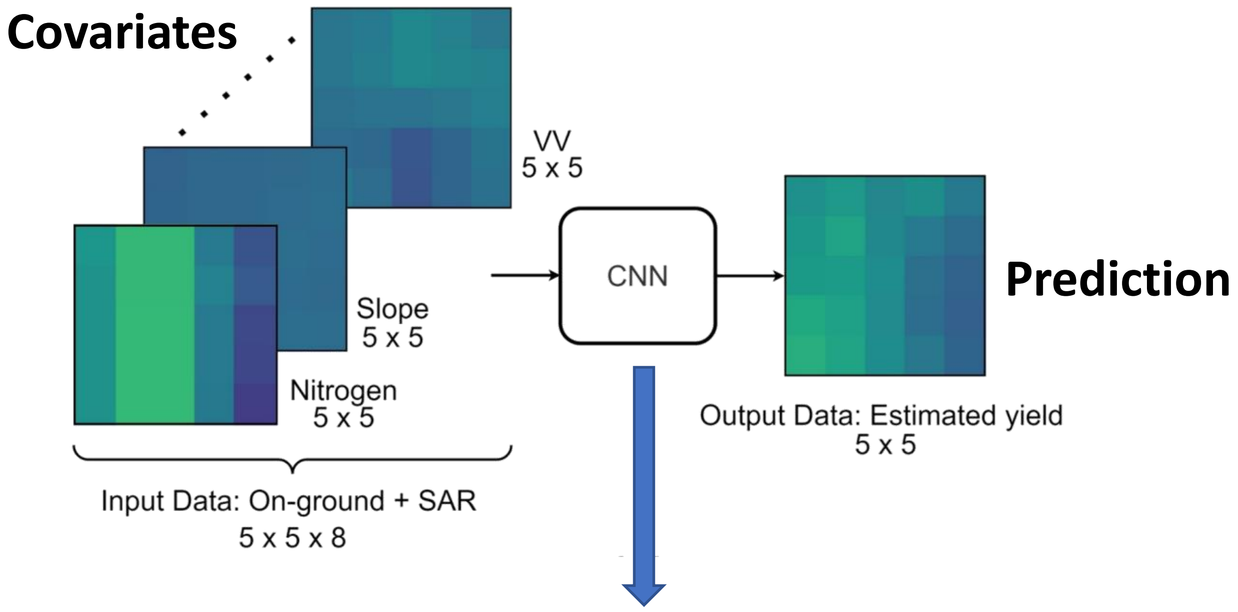
E7	x_1	29.5 ± 1.0	1.8 ± 2.0	1.8 ± 2.0	1.1 ± 1.3	$\mathbf{0 \pm 0}$
	x_2	63.6 ± 43.8	1.6 ± 1.0	42.4 ± 24.7	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$
E8	x_1	0.02 ± 0.01	0.04 ± 0.01	0.8 ± 1.2	0.02 ± 0.02	$\mathbf{0 \pm 0}$
	x_2	0.02 ± 0.01	0.04 ± 0.01	0.9 ± 1.3	0.02 ± 0.01	$\mathbf{0 \pm 0}$
E9	x_1	271.3 ± 446.8	239.8 ± 428.2	375.1 ± 485.5	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$
	x_2	$\mathbf{0 \pm 0}$	0.2 ± 0.09	2.7 ± 1.7	0.05 ± 0.01	$\mathbf{0 \pm 0}$
E10	x_1	$\mathbf{0 \pm 0}$	0.6 ± 0.2	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$
	x_2	$\mathbf{0 \pm 0}$	0.4 ± 0.06	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$
E11	x_1	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$
	x_2	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$
E12	x_1	21.8 ± 13.1	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$
	x_2	$\mathbf{0 \pm 0}$	2.4 ± 1.5	2.5 ± 1.6	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$
E13	x_1	$\mathbf{0 \pm 0}$	0.8 ± 0.5	0.7 ± 0.6	0.7 ± 0.8	$\mathbf{0 \pm 0}$
	x_2	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	3.8 ± 3.5	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$

Skeleton evaluation performance comparison

Crop Production

- In precision agriculture, black-box models are trained to relate input factors and outcome variables (e.g., crop yield)
- Those models can be used to understand the relationship between input variables and outputs

Yield Prediction Model

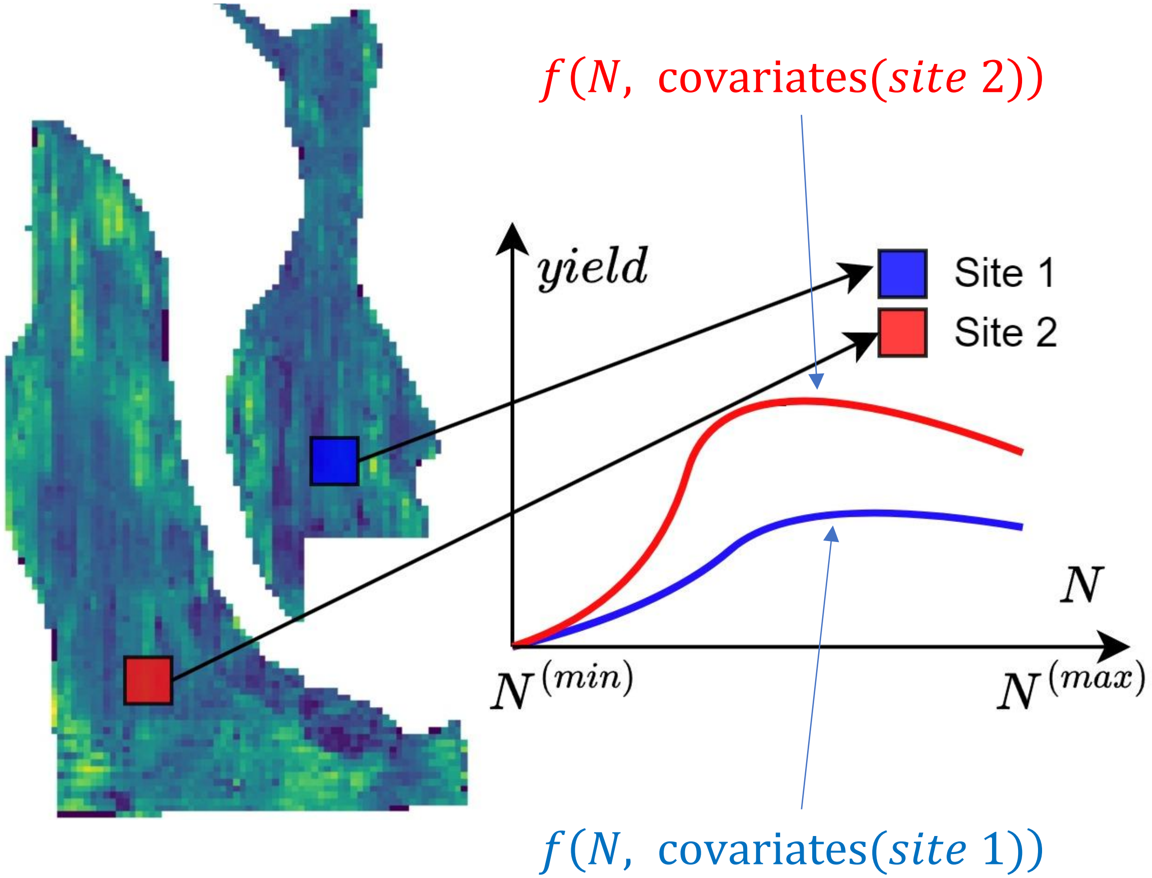


A Convolutional Neural Network designed for yield prediction called **"Hyper3DNetReg"**

MONTANA STATE UNIVERSITY

Crop Production

- Models can estimate N-response curves
- N-response curves are used for optimization and to understand fertilizer responsivity of the field
- SR methods could learn the mathematical expressions that describe site-specific N-response curves from data
- The use of mathematical expressions can accelerate optimization techniques



MONTANA STATE UNIVERSITY

Conclusions

- We aim to establish the basis for a decomposable SR method
- First "neural" SR method that prioritizes the identification of the functional relationships between each system variable and the response
- Experimental results demonstrate that our Multi-Set Transformer model learns to identify skeletons correctly and consistently for all system variables
- Our method can be viewed as a post-hoc interpretability method that generates mathematical expressions as interpretations of a black-box model's function

Thanks to:

- NISL
  - Dr. John Sheppard
  - Will Jardee
- DIFM Project
  - Dr. David Bullock
  - Dr. Brittani Edge
  - Dr. Paul Hegedus
  - Dr. Bruce Maxwell
  - Dr. Taro Mieno
  - Dr. Laila Puntel

MONTANA STATE UNIVERSITY

THANK YOU!

Contact

giorgio-morales

giorgiomorales.github.io/