SNIP: Bridging Mathematical Symbolic and Numeric Realms with Unified Pre-training

*Kazem Meidani *Parshin Shojaee Chandan K. Reddy Amir Barati Farimani

*Equal contribution

ICLR 2024 **Spotlight** Presentation

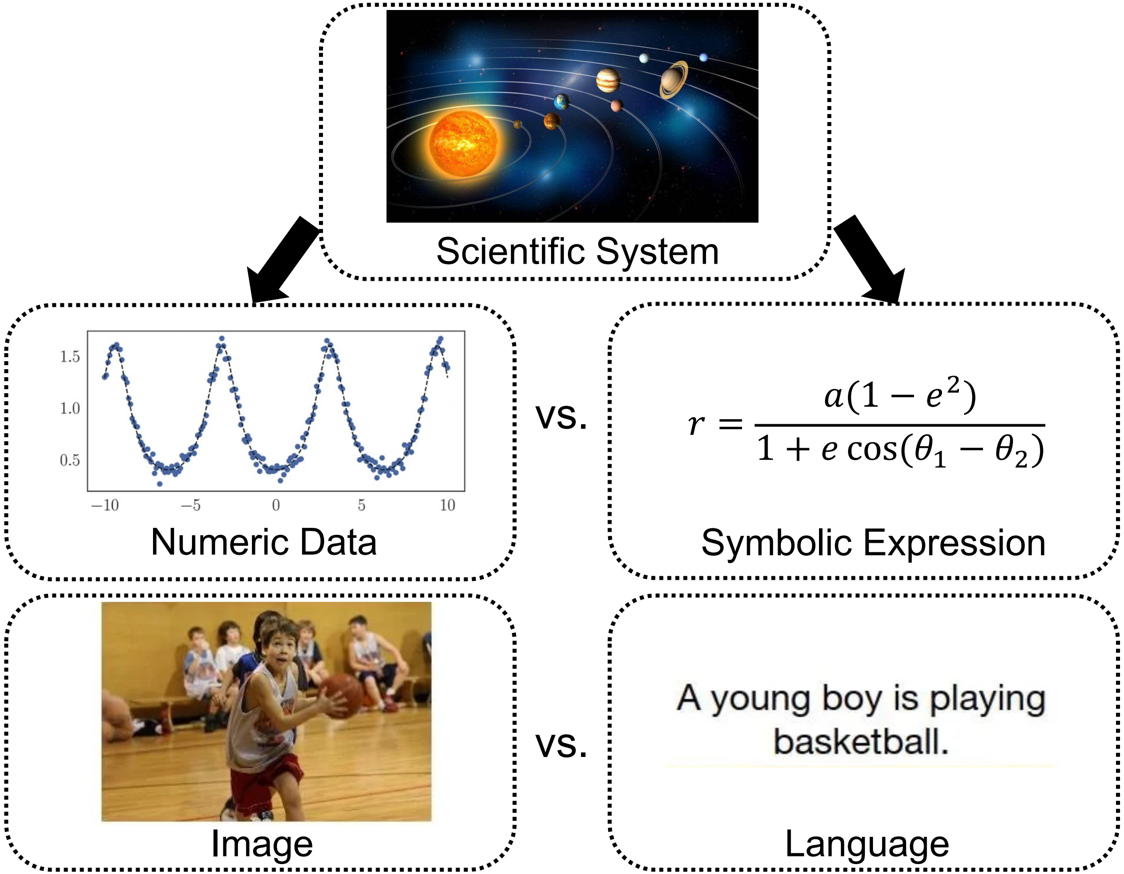
Carnegie Mellon University

ICLR

VIRGINIA TECH

Math: Language of Science

- Symbolic Mathematics has been ***unreasonably effective*** for understanding, predicting, and controlling various scientific systems.
- Obtaining mathematical equations from data is an essential part of ***scientific discovery***.
- Each system can be represented by two modalities of ***Numeric Data Observations*** and ***Symbolic Mathematical Expressions***
- **Multi-modal representation learning** has shown success in many domains including vision-language models.



Scientific System

Numeric Data

vs.

$$r = \frac{a(1-e^2)}{1 + e\cos(\theta_1 - \theta_2)}$$

Symbolic Expression

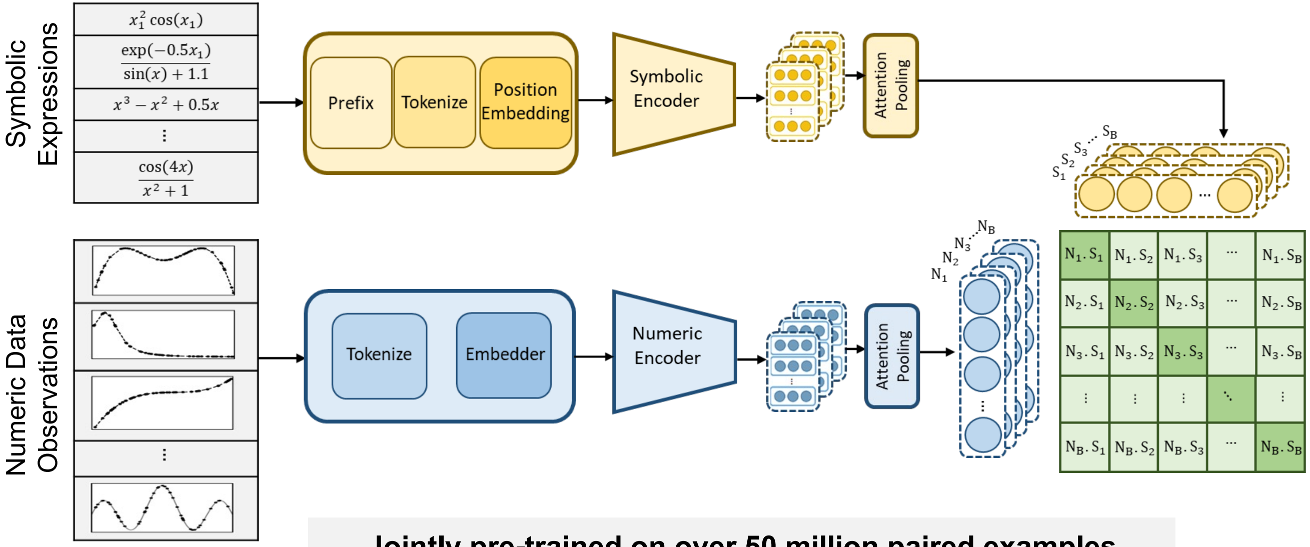
Image

vs.

A young boy is playing basketball.

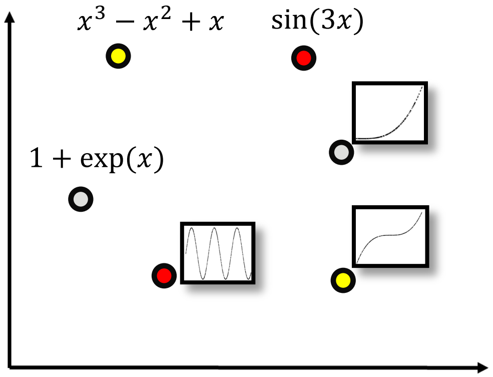
Language

SNIP: Symbolic Numeric Integrated Pre-training



$N_1 \cdot S_1$	$N_1 \cdot S_2$	$N_1 \cdot S_3$	⋯	$N_1 \cdot S_B$
$N_2 \cdot S_1$	$N_2 \cdot S_2$	$N_2 \cdot S_3$	⋯	$N_2 \cdot S_B$
$N_3 \cdot S_1$	$N_3 \cdot S_2$	$N_3 \cdot S_3$	⋯	$N_3 \cdot S_B$
⋮	⋮	⋮	⋱	⋮
$N_B \cdot S_1$	$N_B \cdot S_2$	$N_B \cdot S_3$	⋯	$N_B \cdot S_B$

Jointly pre-trained on over 50 million paired examples



CLIP-like Model for Math

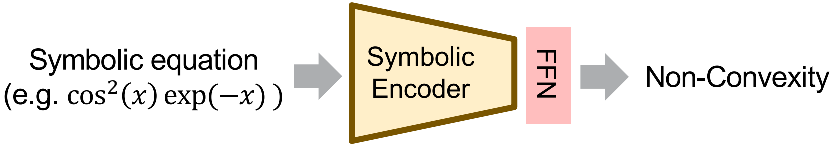
- Data samples are synthetically generated.
- Transformer Encoders are used for both symbolic and numeric encoders.
- Contrastive loss is used for pre-training joint embeddings.

$$\mathcal{L} = -\sum_{(v,s)\in B} \left(\log \mathrm{NCE}(\boldsymbol{Z}_S, \boldsymbol{Z}_V) + \log \mathrm{NCE}(\boldsymbol{Z}_V, \boldsymbol{Z}_S)\right)$$

$$\mathrm{NCE}(\boldsymbol{Z}_S, \boldsymbol{Z}_V) = \frac{\exp\left(\boldsymbol{Z}_S \cdot \boldsymbol{Z}_V^+\right)}{\sum_{\boldsymbol{Z}\in\{\boldsymbol{Z}_V^+,\boldsymbol{Z}_V^-\}} \exp\left(\frac{\boldsymbol{Z}_S\cdot\boldsymbol{Z}}{\tau}\right)}$$

Task 1: Cross-Modal Property Prediction

- Predicting numeric properties from symbolic input and vice vera.
- Example: Predicting convexity ratio (numeric property) of a function based on its symbolic expression

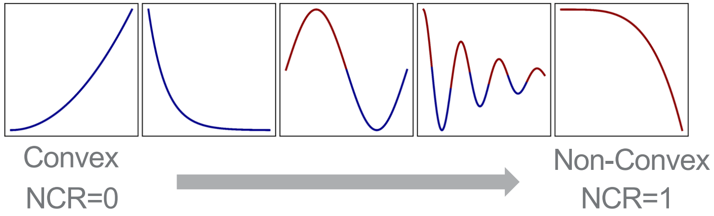


Symbolic equation (e.g. $\cos^2(x)\exp(-x)$)

Symbolic Encoder

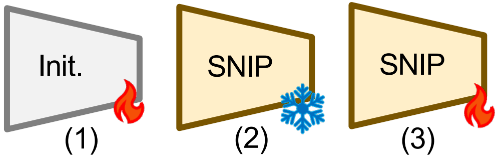
FFN

Non-Convexity

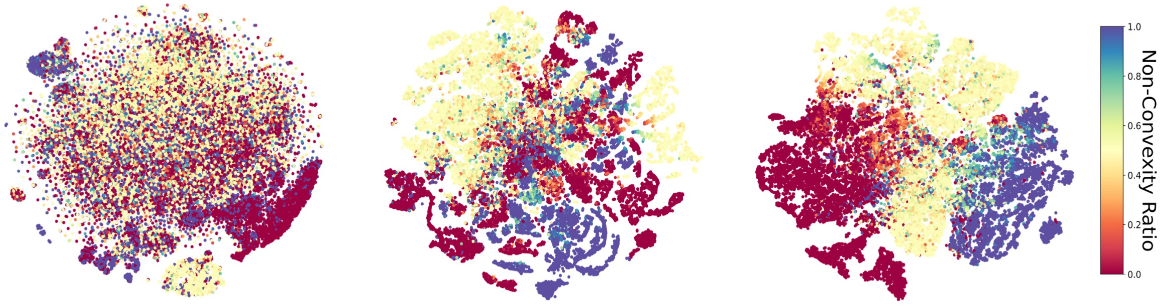


Using a predictor head, we compare:

(1) Supervised Model
(2) SNIP (frozen)
(3) SNIP (finetuned)



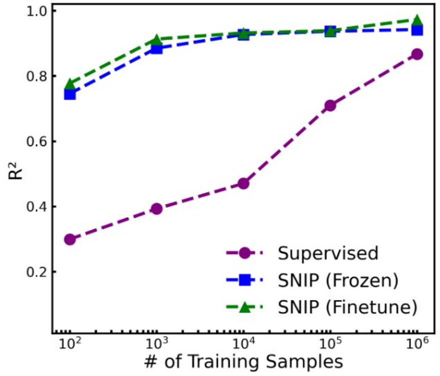
Model	Non-Convexity Ratio		Upwardness	
	↑ R^2	↓ NMSE	↑ R^2	↓ NMSE
Supervised	0.4701	0.5299	0.4644	0.5356
SNIP (frozen)	0.9269	0.0731	0.9460	0.0540
SNIP (finetuned)	**0.9317**	**0.0683**	**0.9600**	**0.0400**



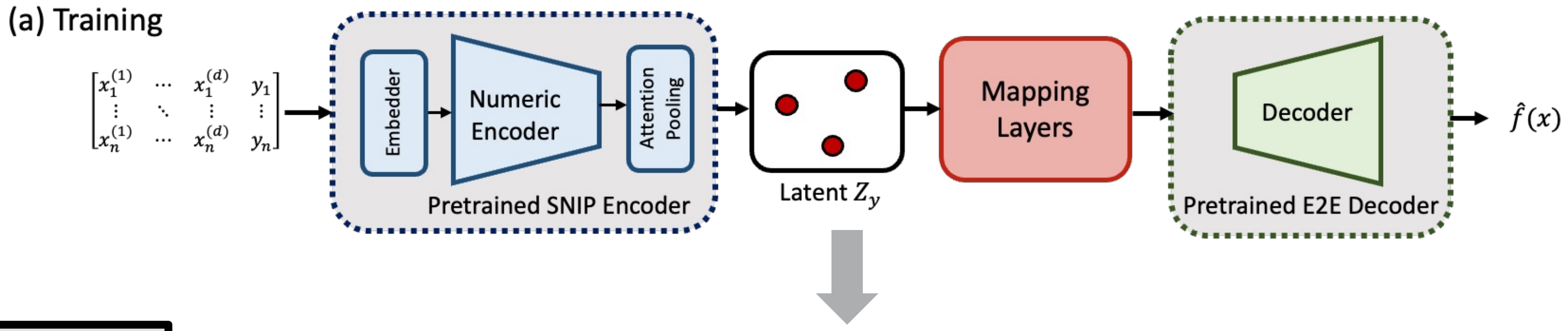
(a) Without Pretraining

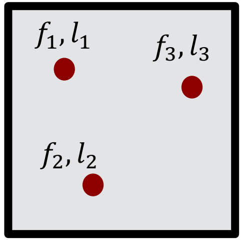
(b) Pretrained, Before Finetuning

(c) Pretrained, After Finetuning



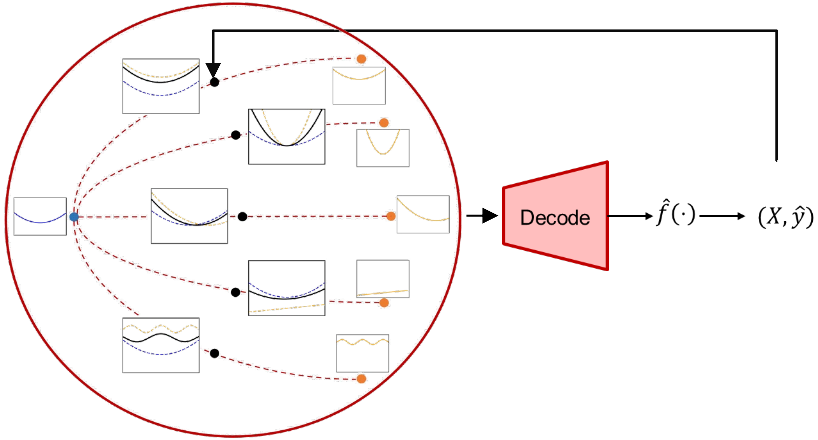
Task 2: SNIP for Symbolic Regression


(a) Training
$\begin{bmatrix} x_1^{(1)} & \cdots & x_1^{(d)} & y_1 \\ \vdots & \ddots & \vdots & \vdots \\ x_n^{(1)} & \cdots & x_n^{(d)} & y_n \end{bmatrix}$
Embedder
Numeric Encoder
Attention Pooling
Pretrained SNIP Encoder
Latent Z_y
Mapping Layers
Decoder
Pretrained E2E Decoder
$\hat{f}(x)$



Symbolic Expression Space

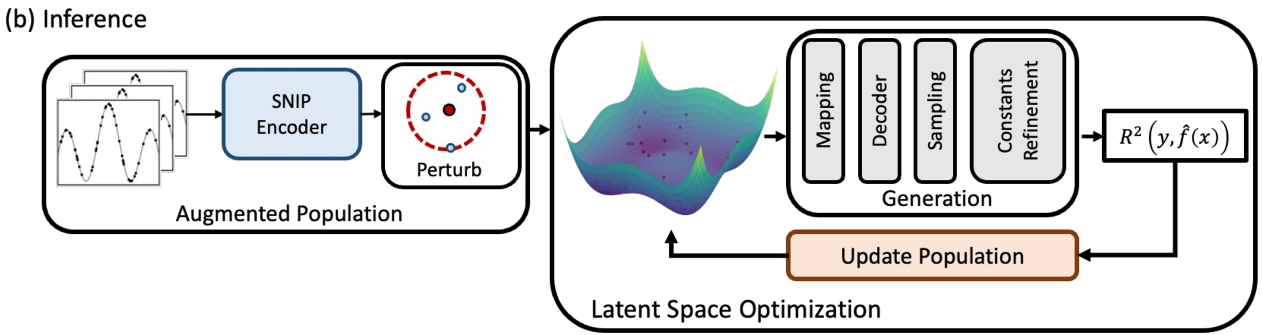
× Discrete / Combinatorial
× Very Large
× Numeric-Ignorant

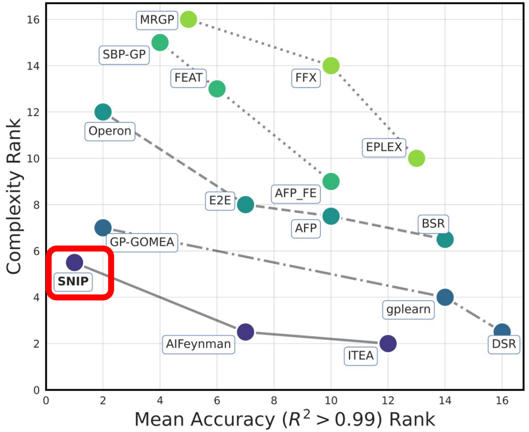


Latent Space

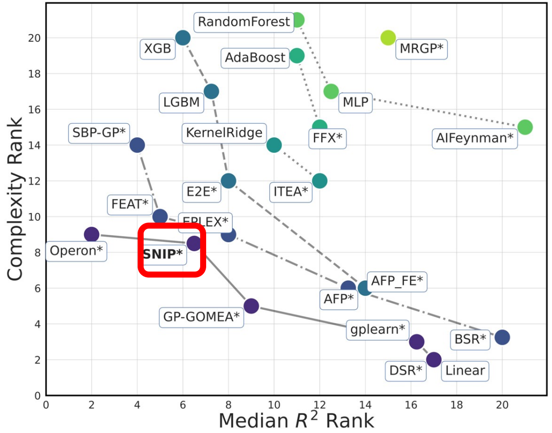
✓ Continuous
✓ Low-dimensional
✓ Interpolatable
✓ Numeric-informed

Task 2: SNIP for Symbolic Regression

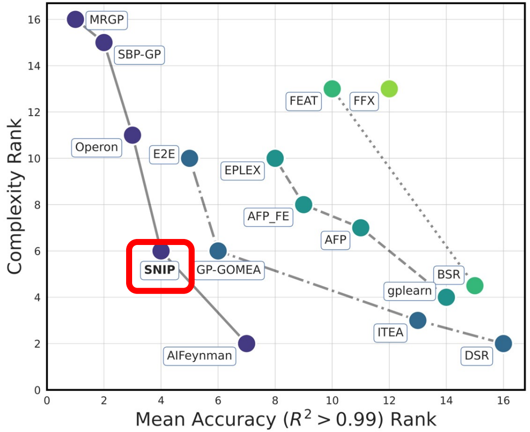




(a) Strogatz datasets

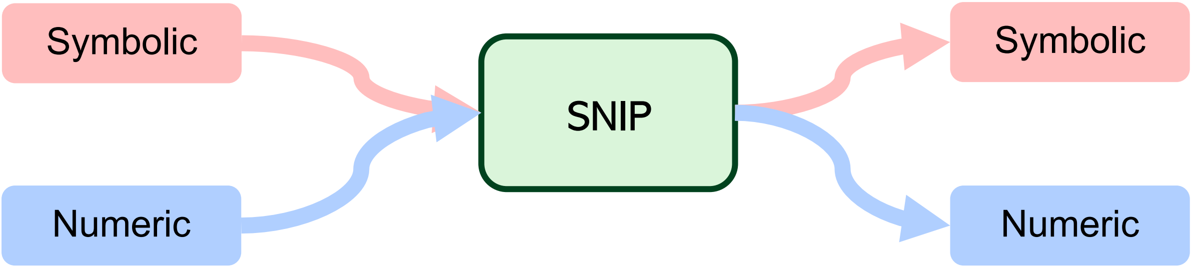


(b) Black-box datasets



(c) Feynman datasets

Future Work



Multi-Modal Pre-training

Multi-Modal Symbolic Regression

Contact: mmeidani@andrew.cmu.edu