# Vertical Symbolic Regression via Deep Policy Gradient

Nan Jiang, Md Nasim, Yexiang Xue. Department of Computer Science, Purdue University, USA

IJCAI JEJU 2024

## Symbolic Regression for Scientific Discovery

Learning an explicit symbolic expression (rather than black-box neural net) from data.
Given a dataset $\mathcal{D} = (x_1, y_1), \dots, (x_n, y_n)$ $(x_i \in \mathbb{R}^m$ and $y_i \in \mathbb{R})$ and a loss function $\mathcal{L}$. Symbolic regression searches for the optimal symbolic expression $\phi^*$ in the space of all candidate expressions (noted as $\Pi$) that minimizes the loss on the dataset $\mathcal{D}$:

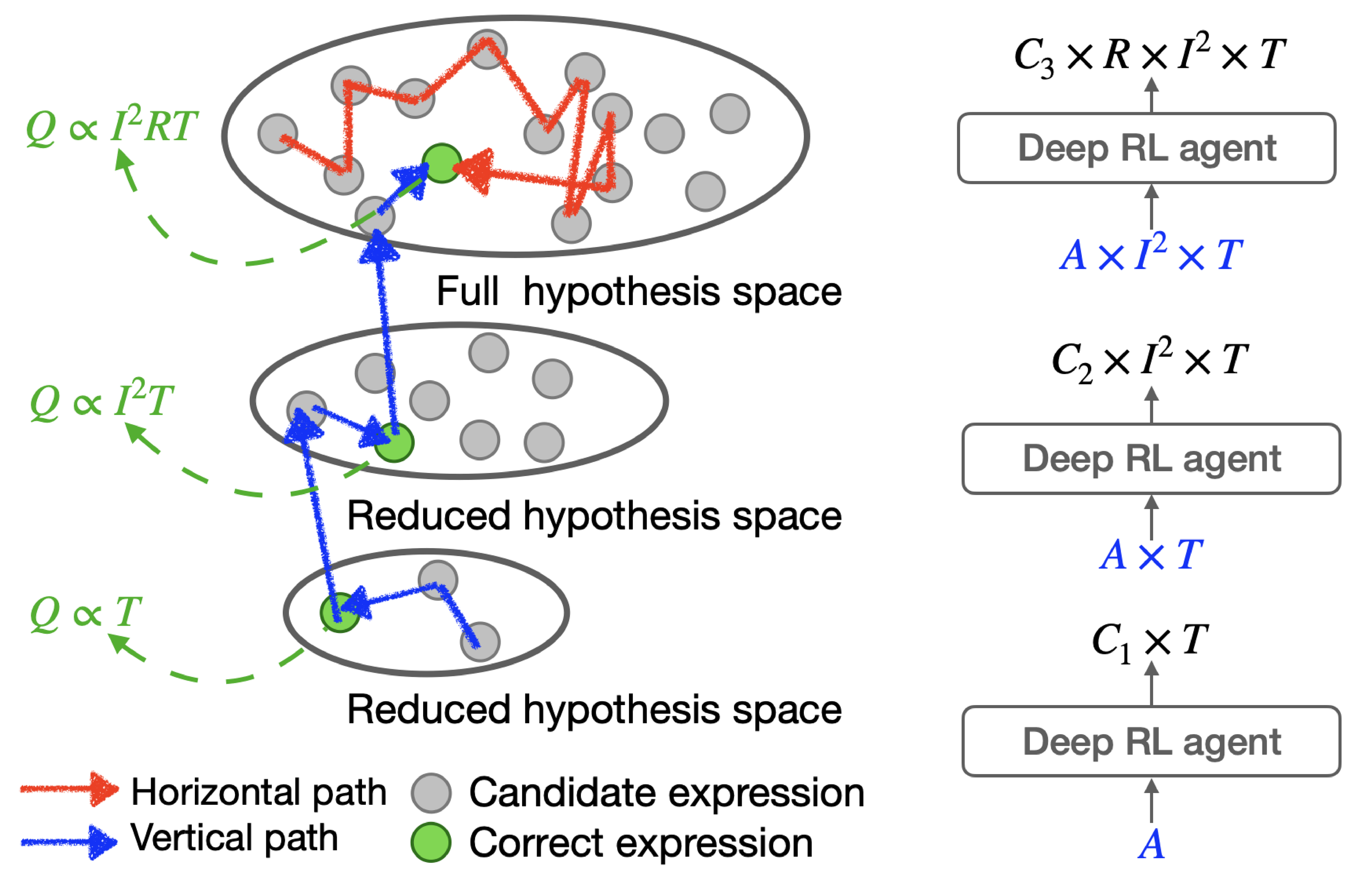
$$\phi^* \leftarrow \operatorname*{argmin}_{\phi \in \Pi} \frac{1}{n} \sum_{\{i=1\}}^{n} \mathcal{L}(x_i, y_i)$$

Current Challenges:

- Current methods, are too slow to find expressions with multiple variables.
- Most recent method (i.e., CVGP) that searches in vertical path are tightly integrated with genetic programming and integrate with other deep symbolic regressor, like deep symbolic regression, will cause (1) difficulty passing gradients to the parameters in the deep neural nets, (2)

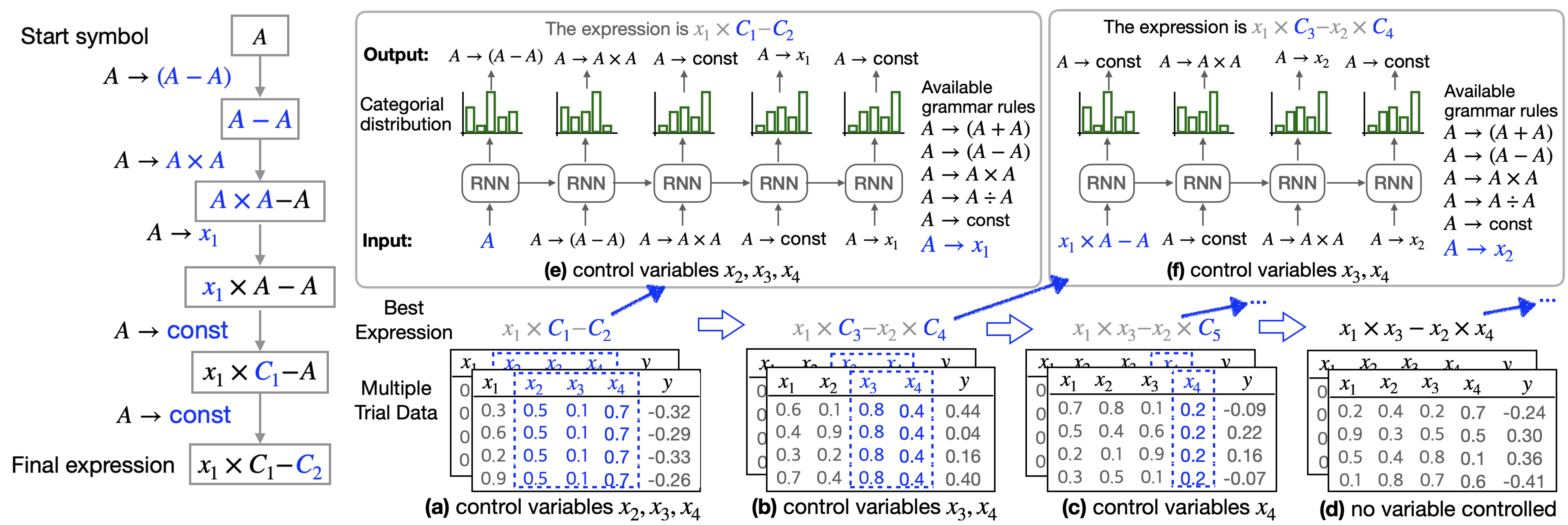
We propose an extended context-free grammar to represent expressions solve it!

## Vertical Path Scale up AI-driven scientific discovery




Horizontal path: current popular methods directly search for the expression in the full hypothesis space; Vertical path: search for the expression from reduced hypothesis space with more and more variables.

## Method: Vertical Symbolic Regression using Deep Policy Gradient (VSR-DPG)




## Experiment 1: Regression on Algebraic Equations

Table 1:median (50%-quartile) of NMSE values of the best-predicted expressions found by all the algorithms. The set of mathematical operator is {+, −, ×, sin, cos, const}. The 3-tuples at the top (·, ·, ·) indicate the number of free variables, singular terms, and cross terms in the ground-truth expressions generating the dataset. "T.O." implies the algorithm is timed out for 48 hours.

| Methods | (2, 1, 1) | (3, 2, 2) | (4, 4, 6) | (5, 5, 5) | (5, 5, 8) | (6, 6, 8) | (6, 6, 10) | (8, 8, 12) |
|---|---|---|---|---|---|---|---|---|
| VSR-GP | 0.005 | 0.028 | 0.086 | 0.014 | 0.066 | 0.066 | **0.104** | T.O. |
| GP | $7E-4$ | 0.023 | 0.044 | 0.063 | 0.102 | 0.127 | 0.159 | 0.872 |
| Eureqa | **<1E-6** | **<1E-6** | 0.024 | 0.158 | 0.284 | 0.433 | 0.910 | 0.162 |
| SPL | 0.006 | 0.033 | 0.144 | 0.147 | 0.307 | 0.391 | 0.472 | 0.599 |
| E2ETransformer | 0.018 | 0.0015 | 0.030 | 0.121 | 0.072 | 0.194 | 0.142 | 0.112 |
| DSR | < **1E-6** | 0.008 | 2.815 | 2.558 | 2.535 | 0.936 | 6.121 | 0.335 |
| PQT | 0.020 | 0.161 | 2.381 | 2.168 | 2.482 | 0.983 | 5.750 | 0.232 |
| VPG | 0.030 | 0.277 | 2.990 | 1.903 | 2.440 | 0.900 | 3.857 | 0.451 |
| GPMeld | $< 1E-6$ | 0.112 | 1.670 | 1.501 | 2.422 | 0.964 | 7.393 | T.O. |
| VSR-DPG (ours) | < **1E-6** | < **1E-6** | < **1E-6** | < **1E-6** | **0.026** | **0.063** | 0.114 | **0.101** |

**Our DSR-VPG finds symbolic expressions with the smallest Normalized-MSEs.**

## Experiment 2: Regression on Differential Equations

| | Lorenz Attractor (3 variables) | MHD Turbulence (5 variables) | Glycolysis Oscillations (7 variables) |
|---|---|---|---|
| SPL | **100%** | 50% | 14.2% |
| SINDy | **100%** | 0% | 0% |
| ProGED | 0% | 0% | 0% |
| ODEFormer | 0% | 0% | NA |
| VSR-DPG (ours) | **100%** | **100%** | **87%** |

Table 4: On the differential equation dataset, $(R^2 \geq 0.9999)$-based accuracy is reported over the best-predicted expression found by all the algorithms. Our VSR-DPG method can discover the governing expressions with a much higher accuracy rate than baselines.

## Acknowledgement

This research was supported by NSF grant CCF-1918327 and DOE – Fusion Energy Science grant: DE-SC0024583.