Vertical AI-driven Materials Scientific Discovery

Nan Jiang, Md Nasim, Yexiang Xue

Department of Computer Science, Purdue University

NEURAL INFORMATION PROCESSING SYSTEMS

Symbolic Regression for Scientific Discovery

Learning an explicit symbolic expression (rather than black-box neural net) from data.

Given a dataset $\mathcal{D} = (x_1, y_1), \ldots, (x_n, y_n)$ ($x_i \in \mathbb{R}^m$ and $y_i \in \mathbb{R}$) and a loss function $\mathcal{L}$. The objective of symbolic regression is to search for the optimal symbolic expression ϕ^* within the space of all candidate expressions Π that minimizes the average loss:

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

Current Challenges:

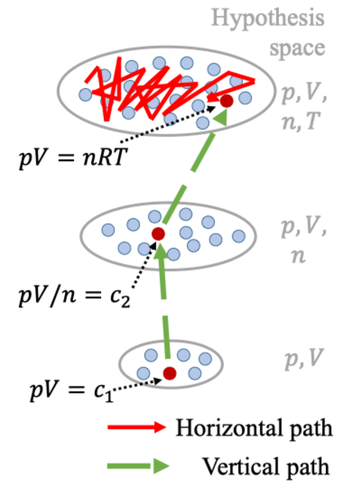
- Incredibly difficult because of the large search space of all possible expressions Π.
- Current methods are too slow to find expressions with ***Multiple variables.***
- We propose to use "***Control Variable Experiments***" (a classical scientific approach) to ***expedites*** scientific machine learning.

Vertical Path Scale up AI-driven scientific discovery

Horizontal way: directly search for the expression in the full hypothesis space

Vertical way: search for the expression incrementally.

1. Search the expression only have one input and one output variables (P, V)
2. Search the expression only have two inputs and one output variables (P,V, n)
3. Search for the whole expression.



Experiments

Table 1: Median (50%) and 75%-quantile NMSE values of the symbolic expressions found by all the algorithms on several noisy benchmark datasets. NMSE is normalized mean square errors.

Dataset	CVGP (ours)		GP		DSR		PQT		VPG		GPMeld		Eureqa	
configs	50%	75%	50%	75%	50%	75%	50%	75%	50%	75%	50%	75%	50%	75%
(3,2,2)	0.001	0.004	0.015	0.135	1.53	43.09	0.58	1.13	0.83	1.32	1.06	2.18	**<1e-6**	**<1e-6**
(4,4,6)	**0.008**	0.059	0.012	**0.054**	1.006	1.249	1.006	2.459	1.221	2.322	1.127	2.286	1.191	6.001
(5,5,5)	**0.011**	**0.019**	0.025	0.177	1.038	8.805	1.048	4.736	1.401	38.26	1.008	1.969	0.996	6.340
(5,5,8)	**0.007**	**0.013**	0.010	0.017	1.403	5.161	1.530	41.27	4.133	27.42	1.386	8.092	1.002	1.495
(6,6,8)	**0.044**	**0.074**	0.058	0.200	1.963	90.53	4.212	8.194	4.425	22.91	15.58	269.6	1.005	1.150
(6,6,10	**0.012**	**0.027**	0.381	0.820	1.021	1.036	1.006	1.048	1.003	1.020	1.022	1.689	1.764	49.041
(a) Datasets containing operators $\{\mathtt{inv}, +, -, \times\}$														
(3,2,2)	0.005	0.123	0.023	0.374	0.087	0.392	0.161	0.469	0.277	0.493	0.112	0.183	**<1e-6**	**<1e-6**
(4,4,6)	**0.028**	0.132	0.044	**0.106**	2.815	9.958	2.381	13.844	2.990	11.316	1.670	2.697	0.024	0.122
(5,5,5)	0.086	0.402	**0.063**	**0.232**	2.558	3.313	2.168	2.679	1.903	2.780	1.501	2.295	0.158	0.377
(5,5,8)	**0.014**	**0.066**	0.102	0.683	2.535	2.933	2.482	2.773	2.440	3.062	2.422	3.853	0.284	0.514
(6,6,8)	**0.066**	**0.166**	0.127	0.591	0.936	1.079	0.983	1.053	0.900	1.018	0.964	1.428	0.433	1.564
(6,6,10	**0.104**	**0.177**	0.159	0.230	6.121	16.32	5.750	16.29	3.857	19.82	7.393	21.709	0.910	1.927
(b) Datasets containing operators $\{\sin, \cos, +, -, \times\}$.														
(3,2,2)	0.039	0.083	0.043	0.551	0.227	7.856	0.855	2.885	0.233	0.400	0.944	1.263	**<1e-6**	**<1e-6**
(4,4,6)	**0.015**	**0.121**	0.042	0.347	1.040	1.155	1.039	1.055	1.049	1.068	1.886	4.104	0.984	1.196
(5,5,5)	**0.038**	**0.097**	0.197	0.514	3.892	69.98	4.311	23.66	5.542	8.839	9.553	16.92	0.901	1.007
(5,5,8)	**0.050**	**0.102**	0.111	0.177	2.379	2.526	1.205	2.336	1.824	2.481	1.142	1.874	1.002	2.445
(6,6,8)	**0.029**	**0.038**	0.091	0.151	1.605	8.005	1.718	7.783	4.691	39.03	1.398	16.60	1.001	1.008
(6,6,10	**0.018**	**0.113**	0	0.194	2.083	23.57	1.797	4.521	1.888	35.45	2.590	8.784	1.001	1.008
(c) Datasets containing opeartors $\{\sin, \cos, \mathtt{inv}, +, -, \times\}$.														

Our CVGP finds symbolic expressions with the smallest NMSEs.

Motivation: Control Variable Experiments

Can you guess which equation $y = f(x_1, x_2, x_3)$ generate the data shown in the left table?

X_1	X_2	X_3	Y
2.5	1.0	9.5	12
5.8	1.0	7.2	13
1.8	1.0	3.2	5
4.2	-1.0	2.2	-2
9.7	-1.0	1.7	-8
3.0	-1.0	4.0	1
7.1	8.6	3.8	64.9
2.5	2.6	3.1	9.6
8.9	1.1	0	11.8

How about if I only ask you to look into these rows?
$y = x_1 + x_3$

How about these rows?
$y = -x_1 + x_3$?

Maybe the WHOLE equation is:
$y = x_2 x_1 + x_3$?

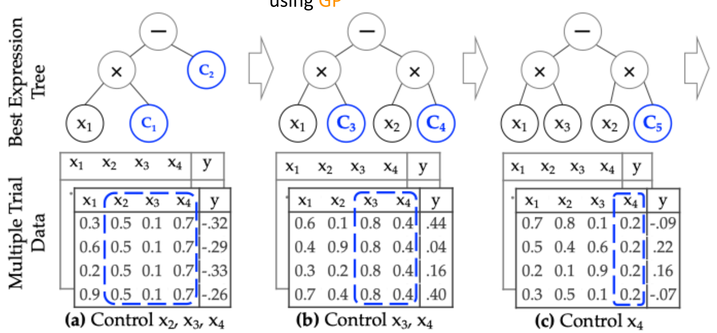
X_2 is controlled!

Orange and blue data are two control variable experiment trials. Control variable experiments simplify symbolic regression!

Method: Control Variable Genetic Programming

Found by GP

Extended from previous reduced-form equations using GP



The data are generated from a data Oracle. The constants are identified using BFGS optimizer on batches of data

Table 2: Ground-truth expression recovery rate.

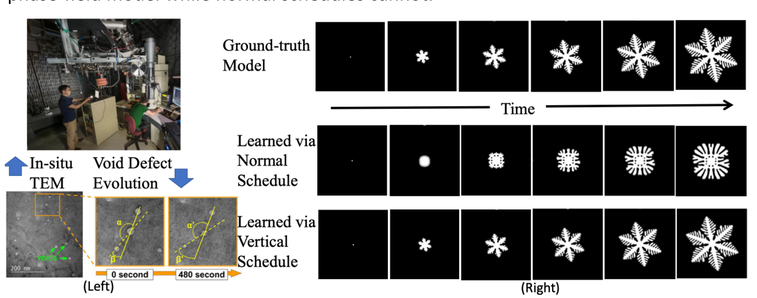
Operator set	Dataset configs	CVGP (ours)	GP
$\{\mathtt{inv}, +, -, \times\}$	(2,1,1)	**64%**	44%
$\{\sin, \cos, +, -, \times\}$		**46%**	22%
$\{\sin, \cos, \mathtt{inv}, +, -, \times\}$		**44%**	32%

Our CVGP finds more correct expressions.

Phase-field model for dendritic solidification in Material Science

In the vertical schedule, first the learning is concentrated on a subset of model parameters. This is done by feeding the model with designed training data in which the remaining parameters do not affect the dynamics of the PDEs.

After this phase, the learning is expanded to all parameters. The right panel of the figure above demonstrate that learning via the vertical schedule is able to identify the correct phase-field model while normal schedules cannot.



Acknowledgement

This research was supported by NSF grants IIS-1850243, CCF-1918327.