International Joint Conference on Artificial Intelligence

Vertical Symbolic Regression via Deep Policy Gradient

Nan Jiang, Md Nasim and Yexiang Xue
Purdue University
{jiang631, mnasim, yexiang}@purdue.edu

IJCAI JEJU 2024

What is Symbolic Regression?

Given a dataset D:

x_1	x_2	x_3	x_4	y
0.3	0.5	0.1	0.7	-0.32
0.6	0.5	0.1	0.7	-0.29
0.2	0.5	0.1	0.7	-0.33
0.9	0.5	0.1	0.7	-0.26

Find a closed-form equation ϕ, like:

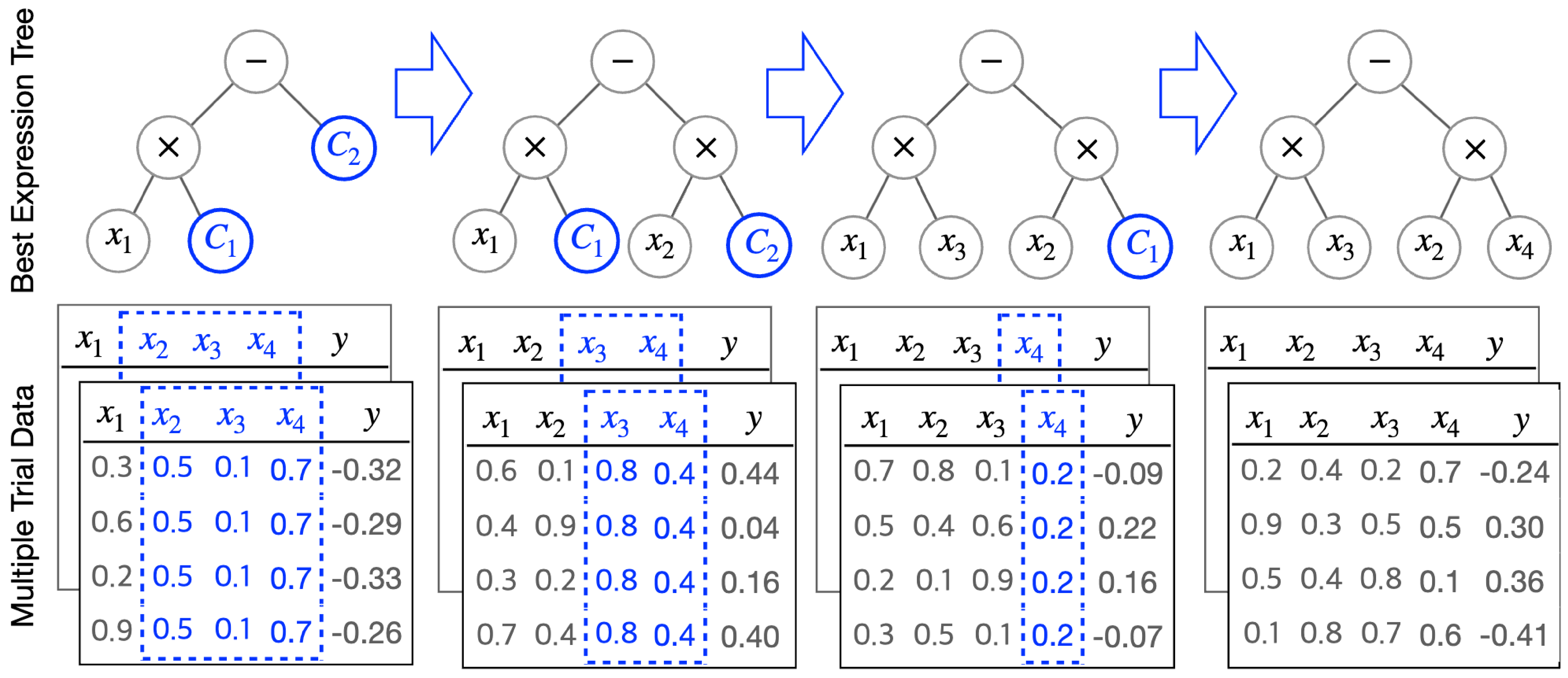
$$\phi = x_1 x_3 - x_2 x_4$$

that best fits the dataset D.

What is Vertical Discovery Path?

Assumption: need a data oracle that can return the controlled variables dataset

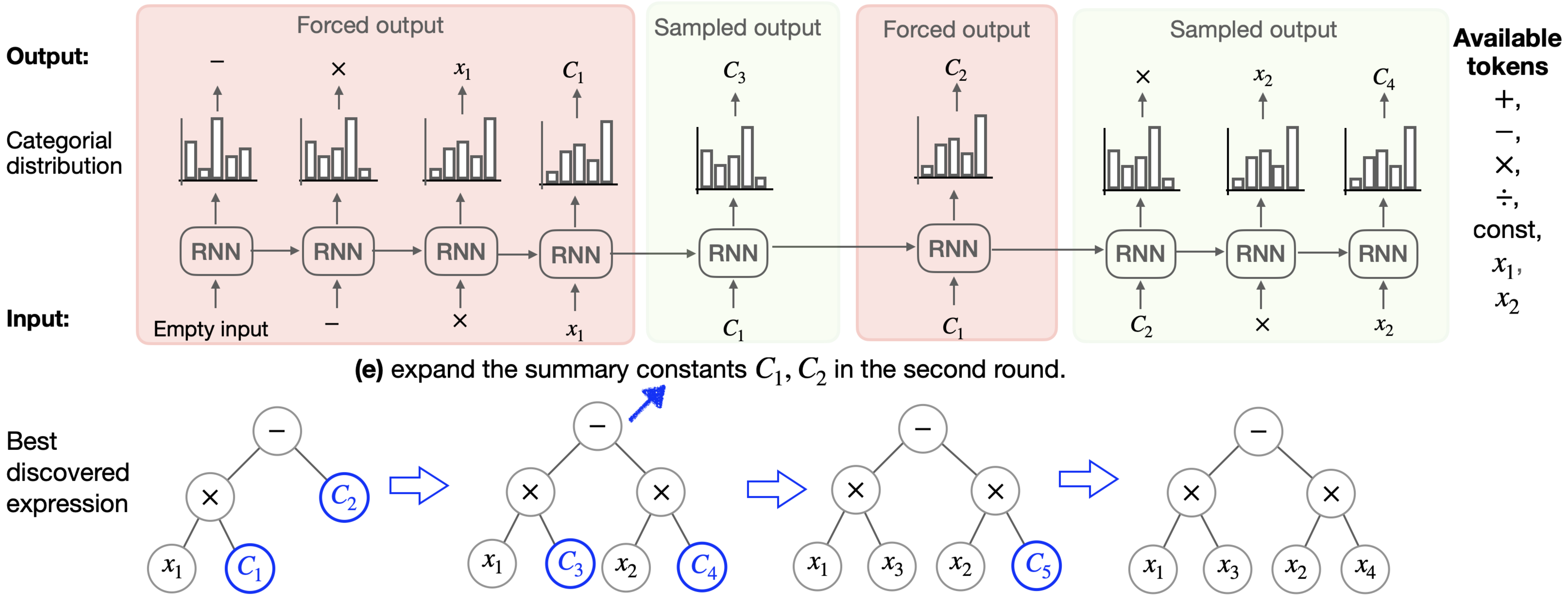
We can iteratively reduce the number of controlled variables.



(a) control x_2, x_3, x_4 **(b)** control x_3, x_4 **(c)** control x_4 **(d)** no control

Challenge in integrating with deep RL

Every preorder traversal of the binary tree correspond to an expression.



(e) expand the summary constants C_1, C_2 in the second round.

(a) control x_2, x_3, x_4 **(b)** control x_3, x_4 **(c)** control x_4 **(d)** no control

The constraints enforce the output of RNN output the given token at each step. It has limitations in *passing the gradient* to the parameters of RNN and needs *heavy engineering* of different constraints.

Context-free grammar for expression

A context-free grammar is represented by $(\mathrm{V}, \Sigma, R, S)$.

- Σ is a set of *terminal* symbols, like:

$$\{x_1, x_2, \cdots x_N, const\}.$$

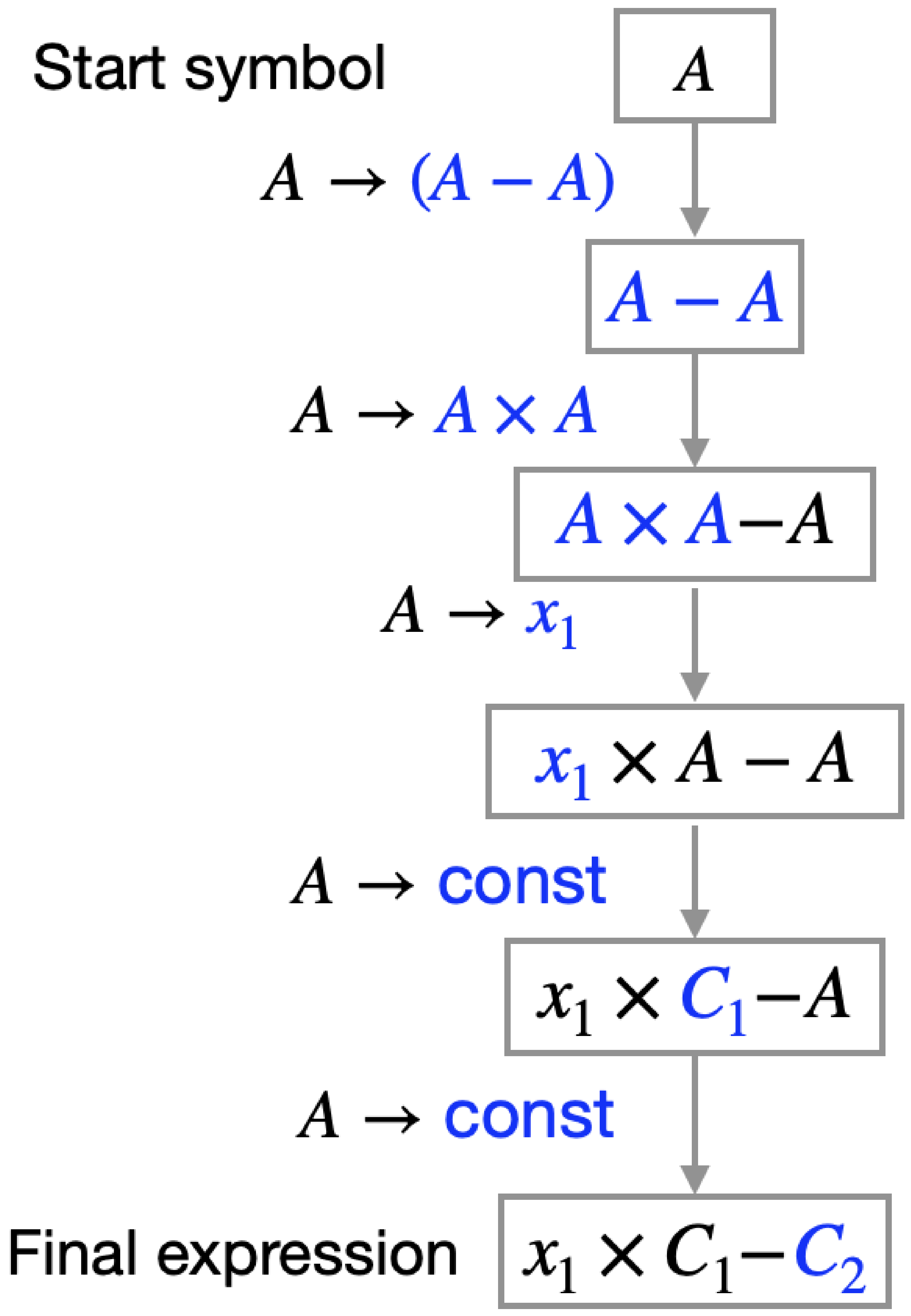
- V is a set of *non-terminal* symbols, like $\{A\}$.

- R is a set of production rules. Left part is replaced with right part.

$$\{A \to (A + A); A \to A - A; A \to A \times A; A \to A \div A\}.$$

- S an ***extended*** start symbol, like:

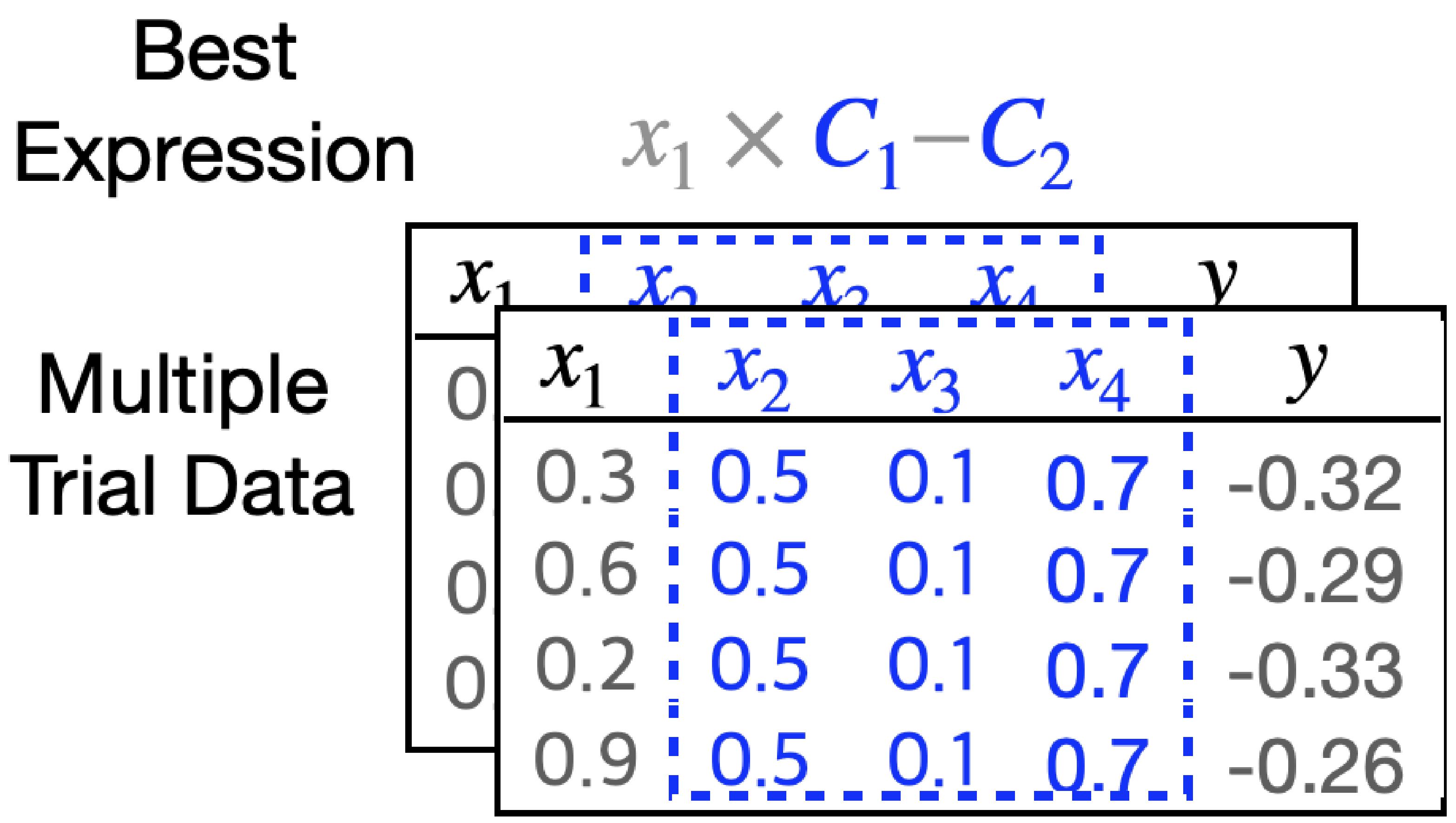
$$\{A, x_1 \times A - A\}.$$



Vertical Symbolic Regression

Assumption: need a data oracle that can return the controlled variables dataset

We can iteratively reduce the number of controlled variables.

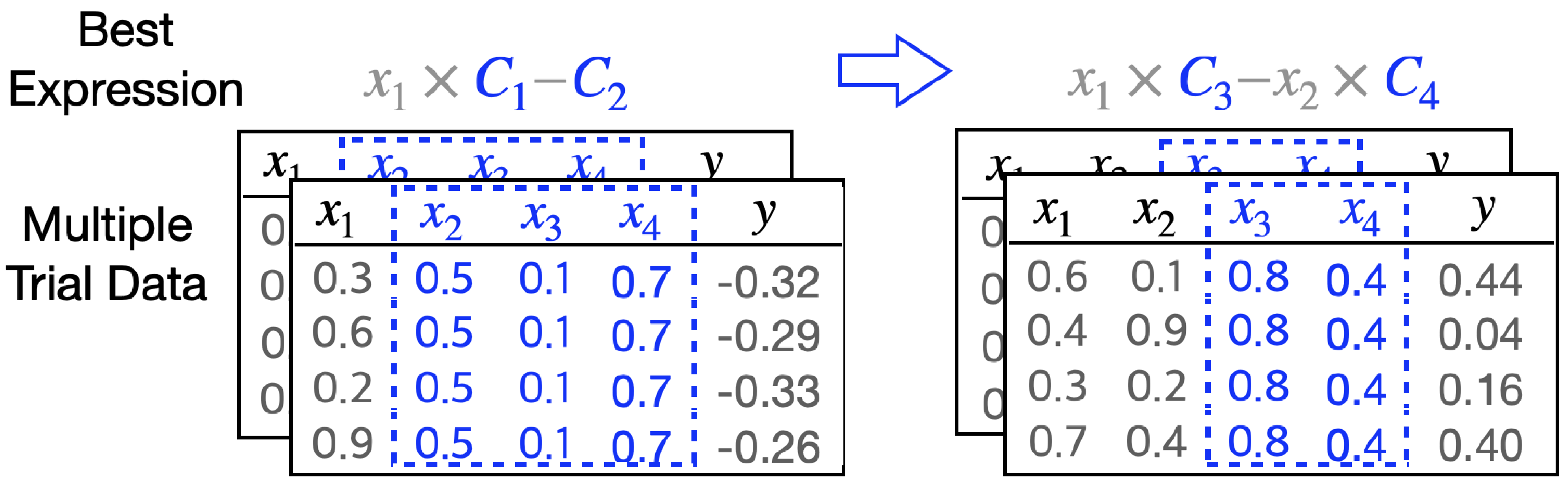


(a) control variables x_2, x_3, x_4

Vertical Symbolic Regression

Assumption: need a data oracle that can return the controlled variables dataset

We can iteratively reduce the number of controlled variables.



(a) control variables x_2, x_3, x_4 **(b)** control variables x_3, x_4

Vertical Symbolic Regression

Assumption: need a data oracle that can return the controlled variables dataset

We can iteratively reduce the number of controlled variables.

Best Expression: $x_1 \times C_1 - C_2$

Multiple Trial Data:

x_1	x_2	x_3	x_4	y
0.3	0.5	0.1	0.7	-0.32
0.6	0.5	0.1	0.7	-0.29
0.2	0.5	0.1	0.7	-0.33
0.9	0.5	0.1	0.7	-0.26

(a) control variables x_2, x_3, x_4

Best Expression: $x_1 \times C_3 - x_2 \times C_4$

x_1	x_2	x_3	x_4	y
0.6	0.1	0.8	0.4	0.44
0.4	0.9	0.8	0.4	0.04
0.3	0.2	0.8	0.4	0.16
0.7	0.4	0.8	0.4	0.40

(b) control variables x_3, x_4

Best Expression: $x_1 \times x_3 - x_2 \times C_5$

x_1	x_2	x_3	x_4	y
0.7	0.8	0.1	0.2	-0.09
0.5	0.4	0.6	0.2	0.22
0.2	0.1	0.9	0.2	0.16
0.3	0.5	0.1	0.2	-0.07

(c) control variables x_4

Vertical Symbolic Regression

Assumption: need a data oracle that can return the controlled variables dataset

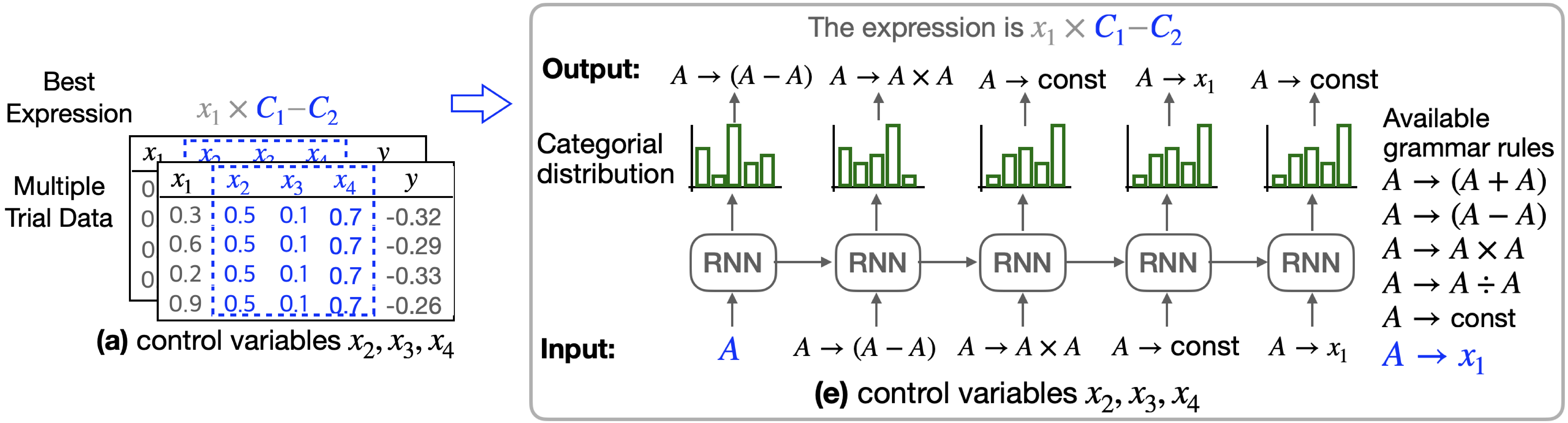
We can iteratively reduce the number of controlled variables.

(a) control variables x_2, x_3, x_4 **(b)** control variables x_3, x_4 **(c)** control variables x_4 **(d)** no variable controlled

Vertical Symbolic Regression

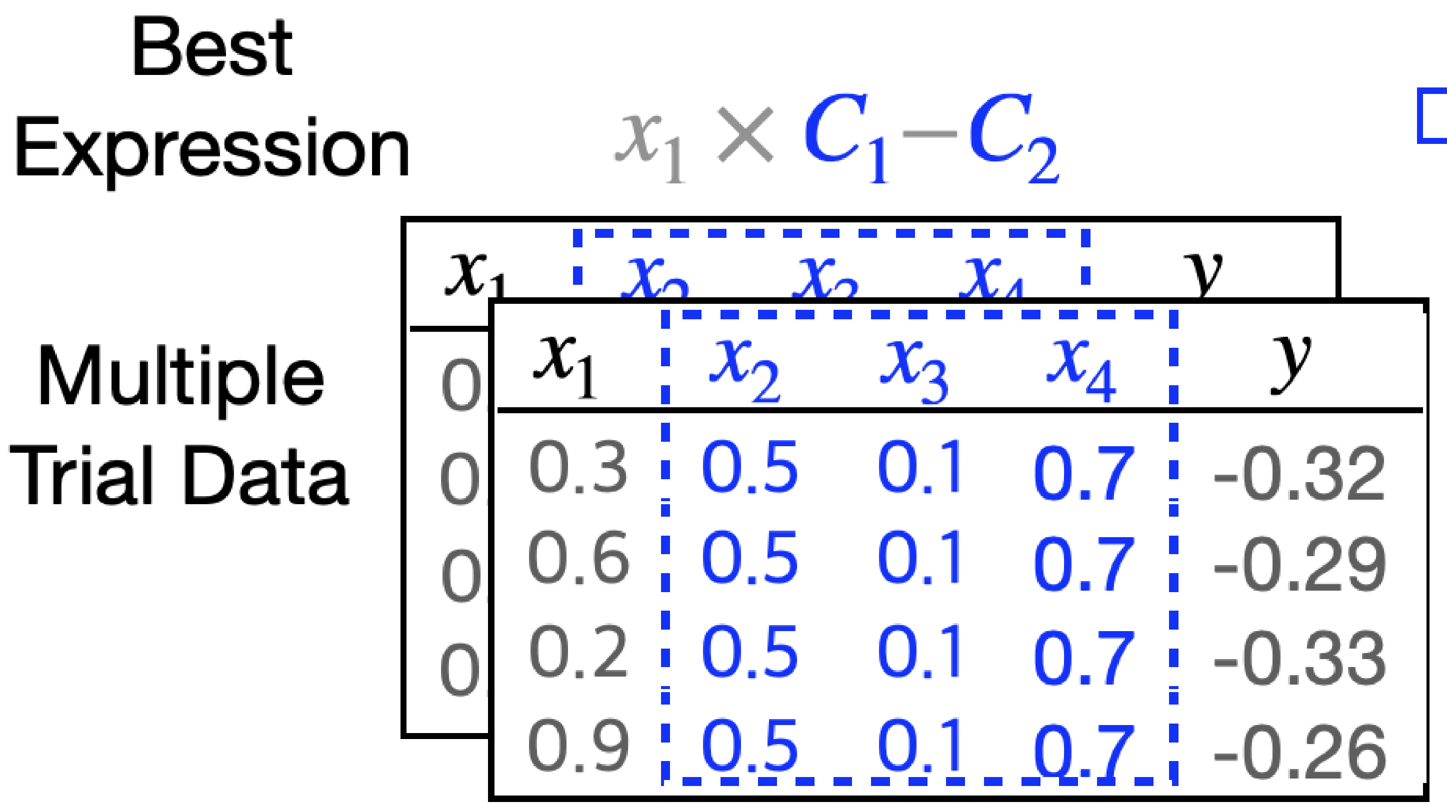
Assumption: need a data oracle that can return the controlled variables dataset

We can iteratively reduce the number of controlled variables.



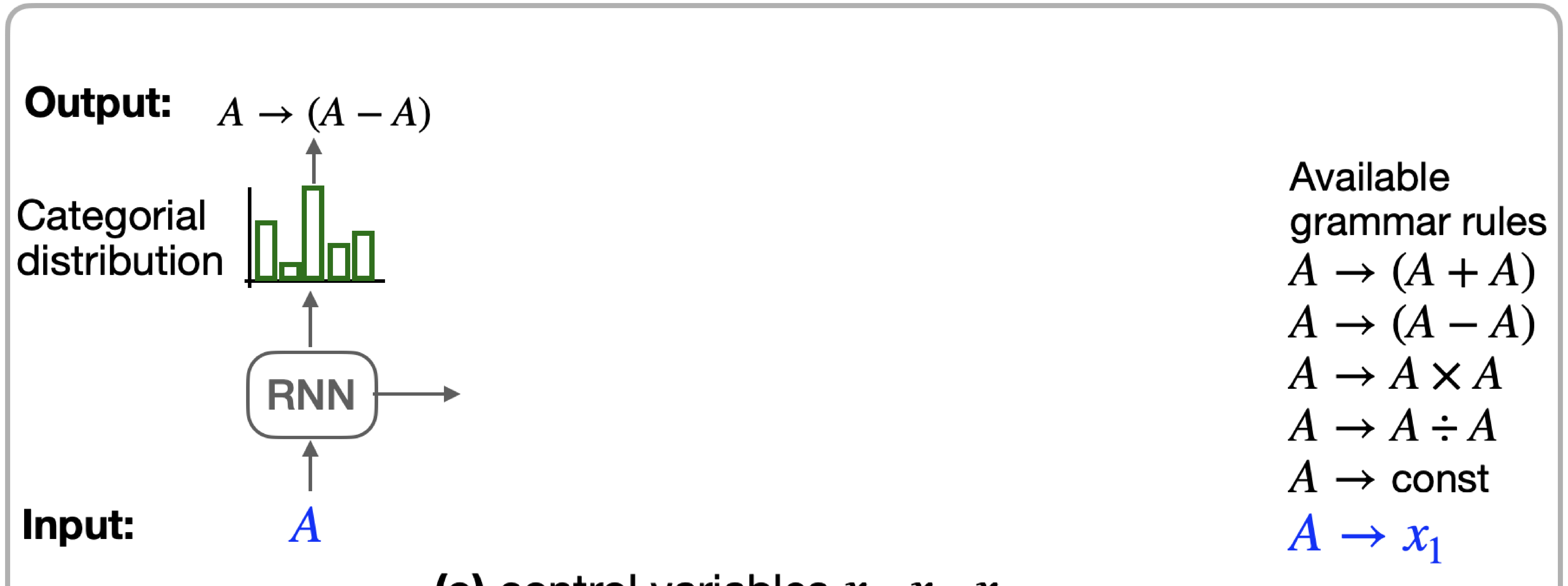
Vertical Symbolic Regression – step 1

$$A - A$$



x_1	x_2	x_3	x_4	y
0.3	0.5	0.1	0.7	-0.32
0.6	0.5	0.1	0.7	-0.29
0.2	0.5	0.1	0.7	-0.33
0.9	0.5	0.1	0.7	-0.26

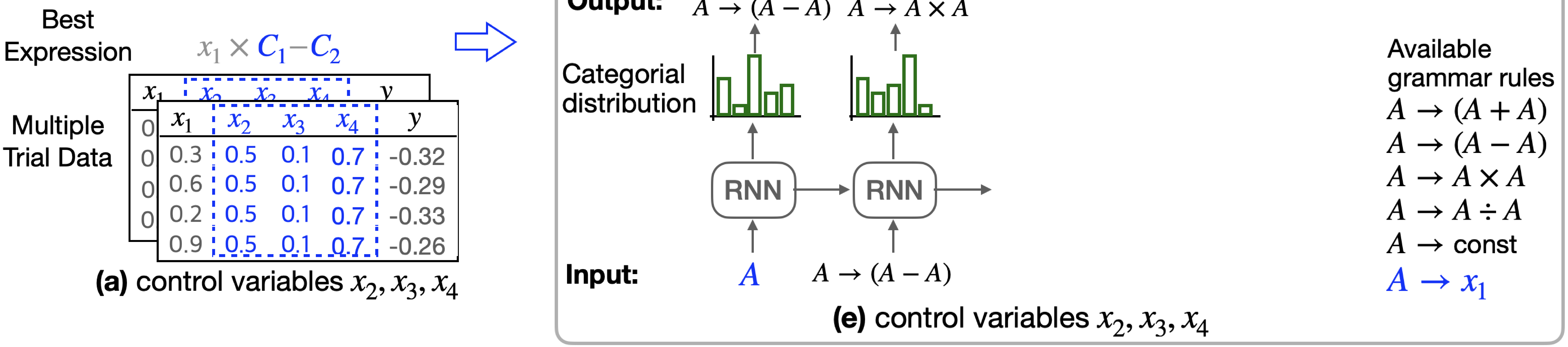
(a) control variables x_2, x_3, x_4



(e) control variables x_2, x_3, x_4

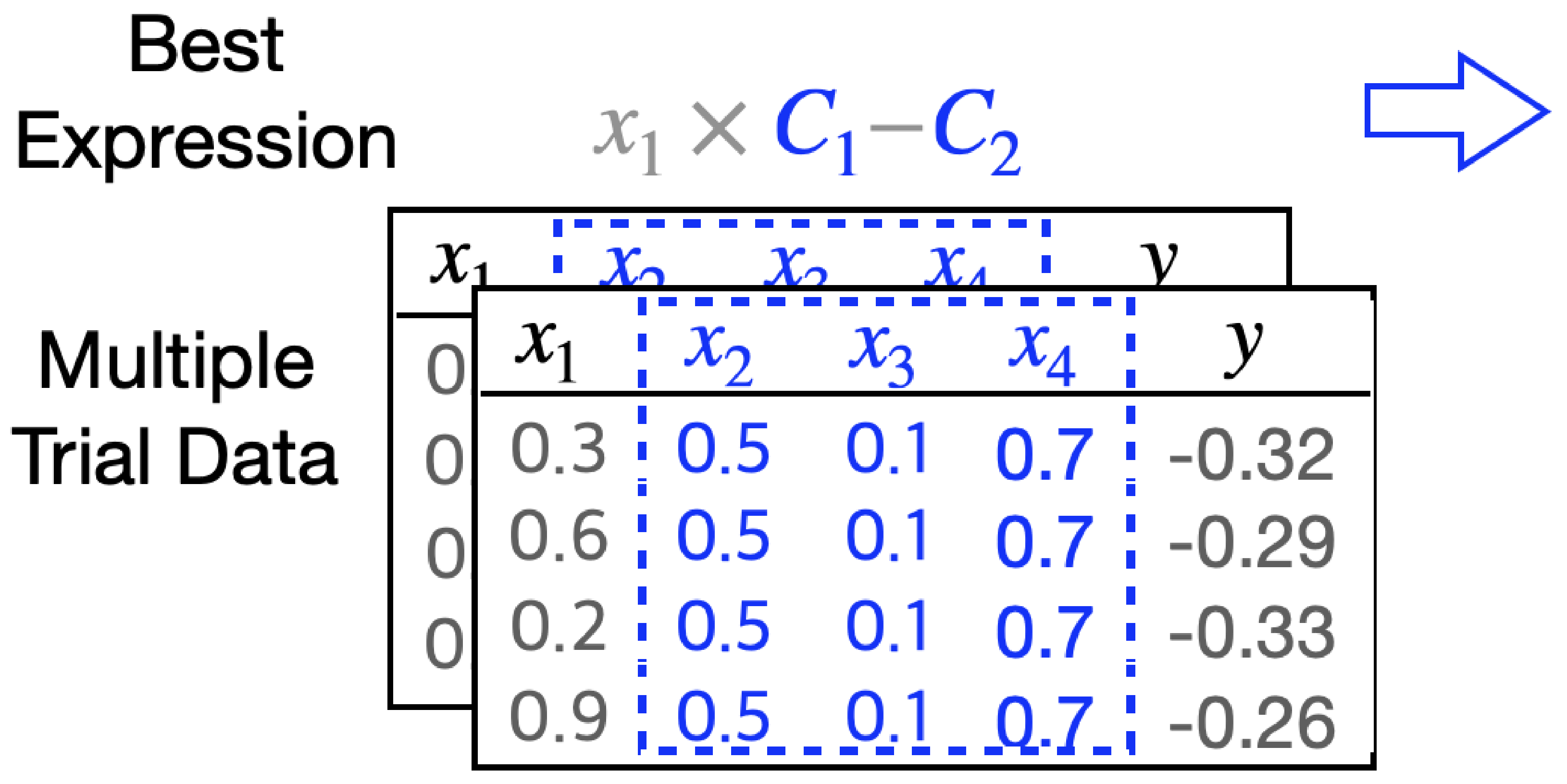
Vertical Symbolic Regression – step 2

$$A \times A - A$$

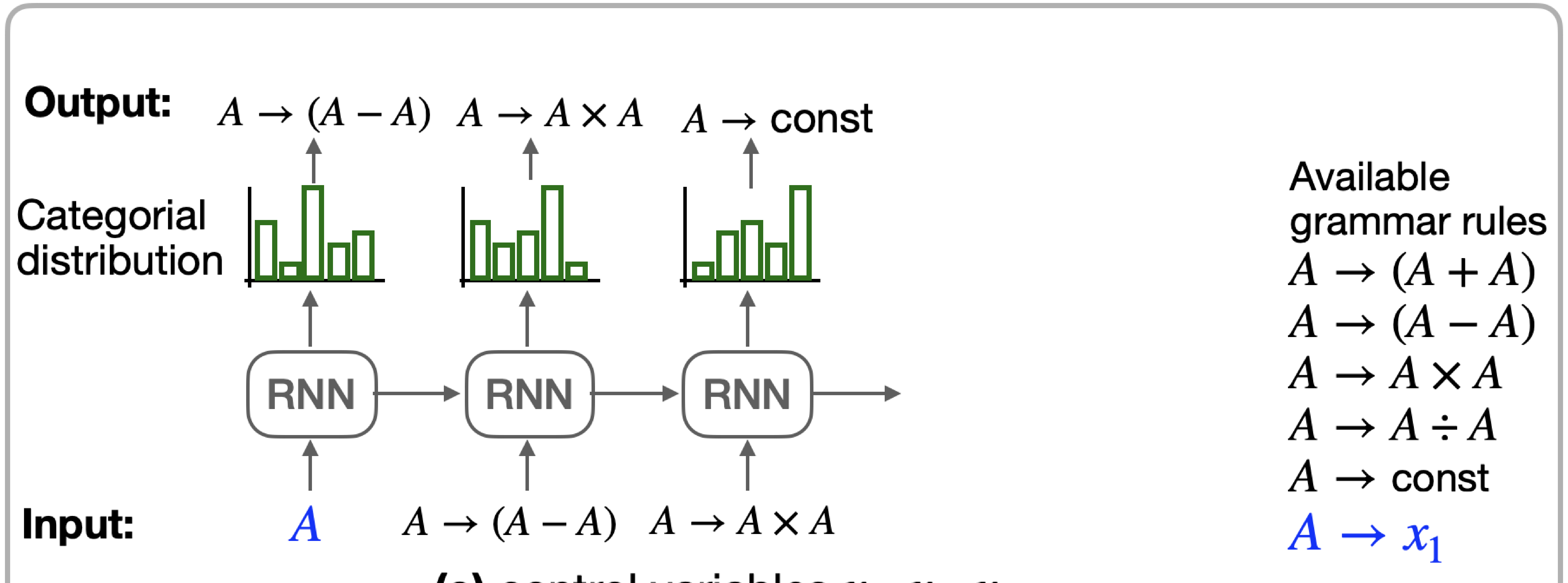


Vertical Symbolic Regression – step 3

$$C_1 \times A - A$$



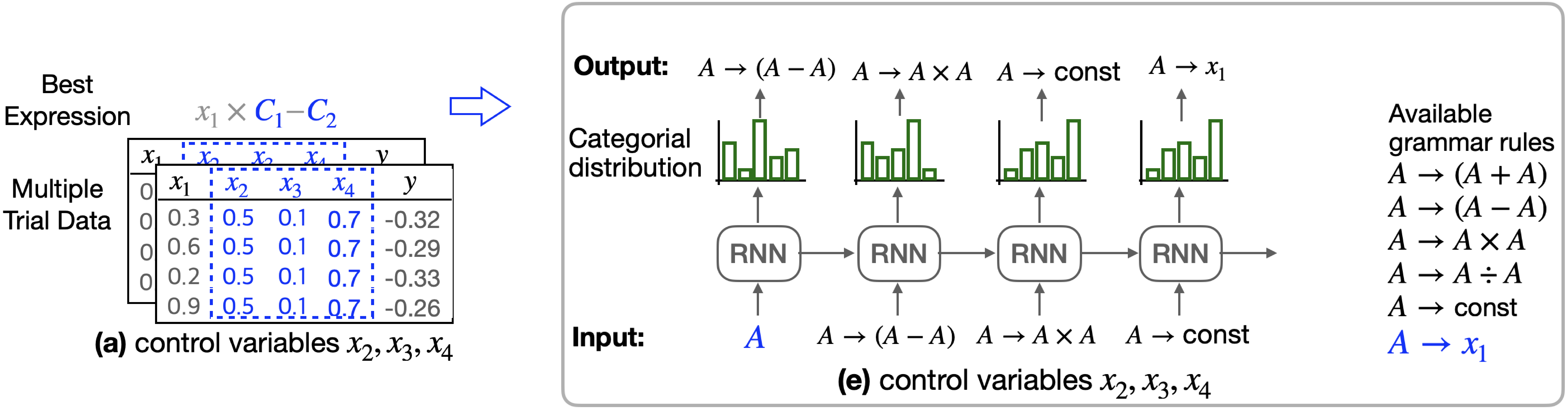
(a) control variables x_2, x_3, x_4



(e) control variables x_2, x_3, x_4

Vertical Symbolic Regression – step 4

$$C_1 \times x_1 - A$$

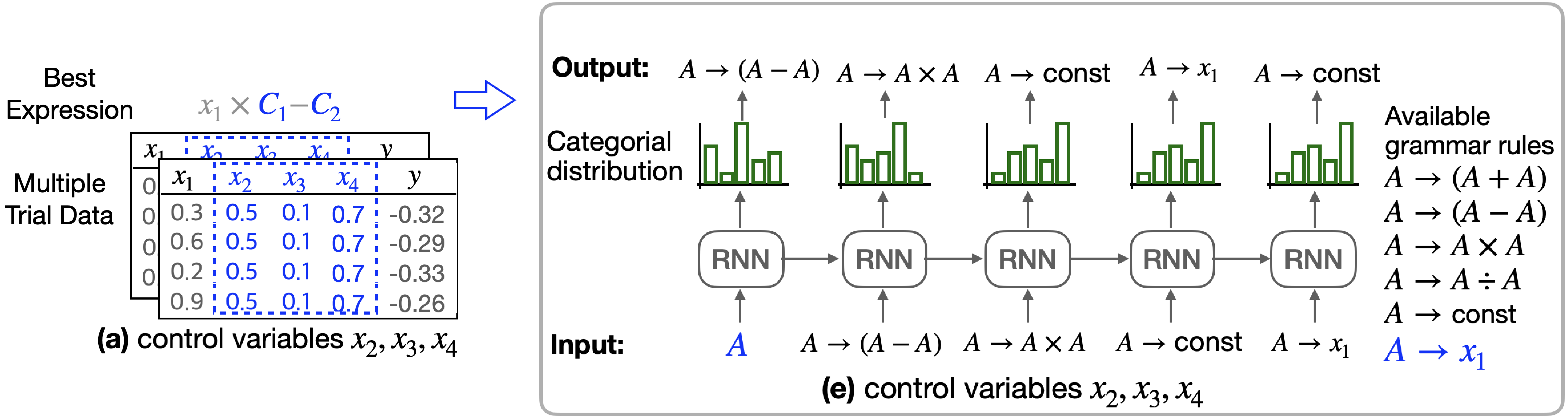


(a) control variables x_2, x_3, x_4

(e) control variables x_2, x_3, x_4

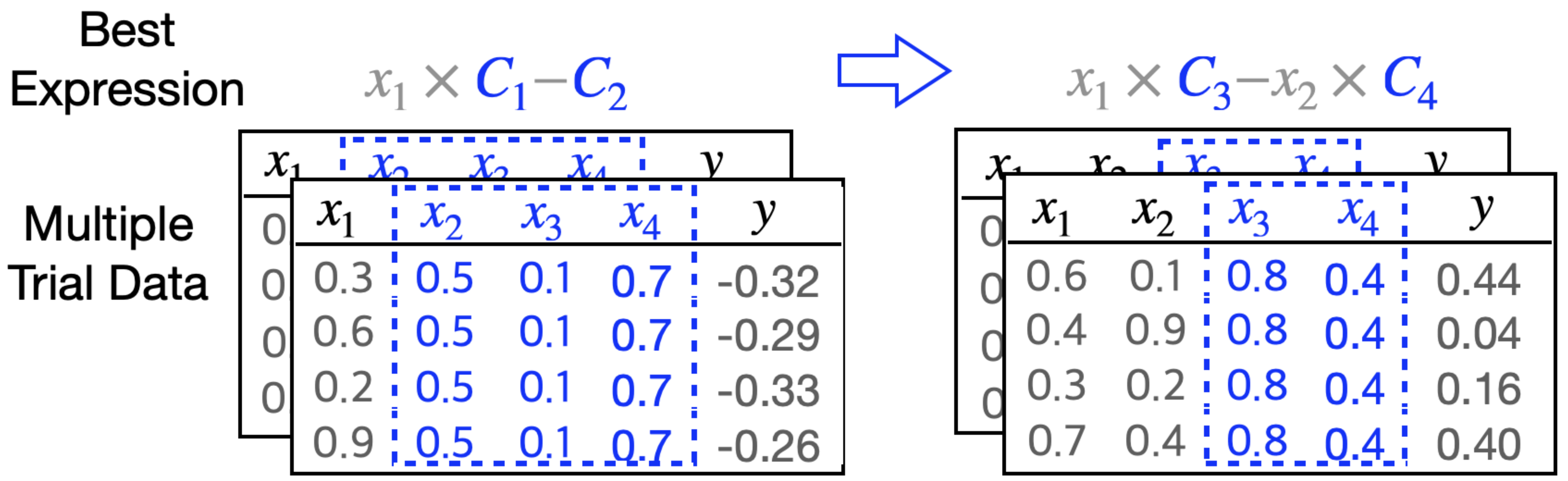
Vertical Symbolic Regression – step 5

$$C_1 \times x_1 - C_2$$

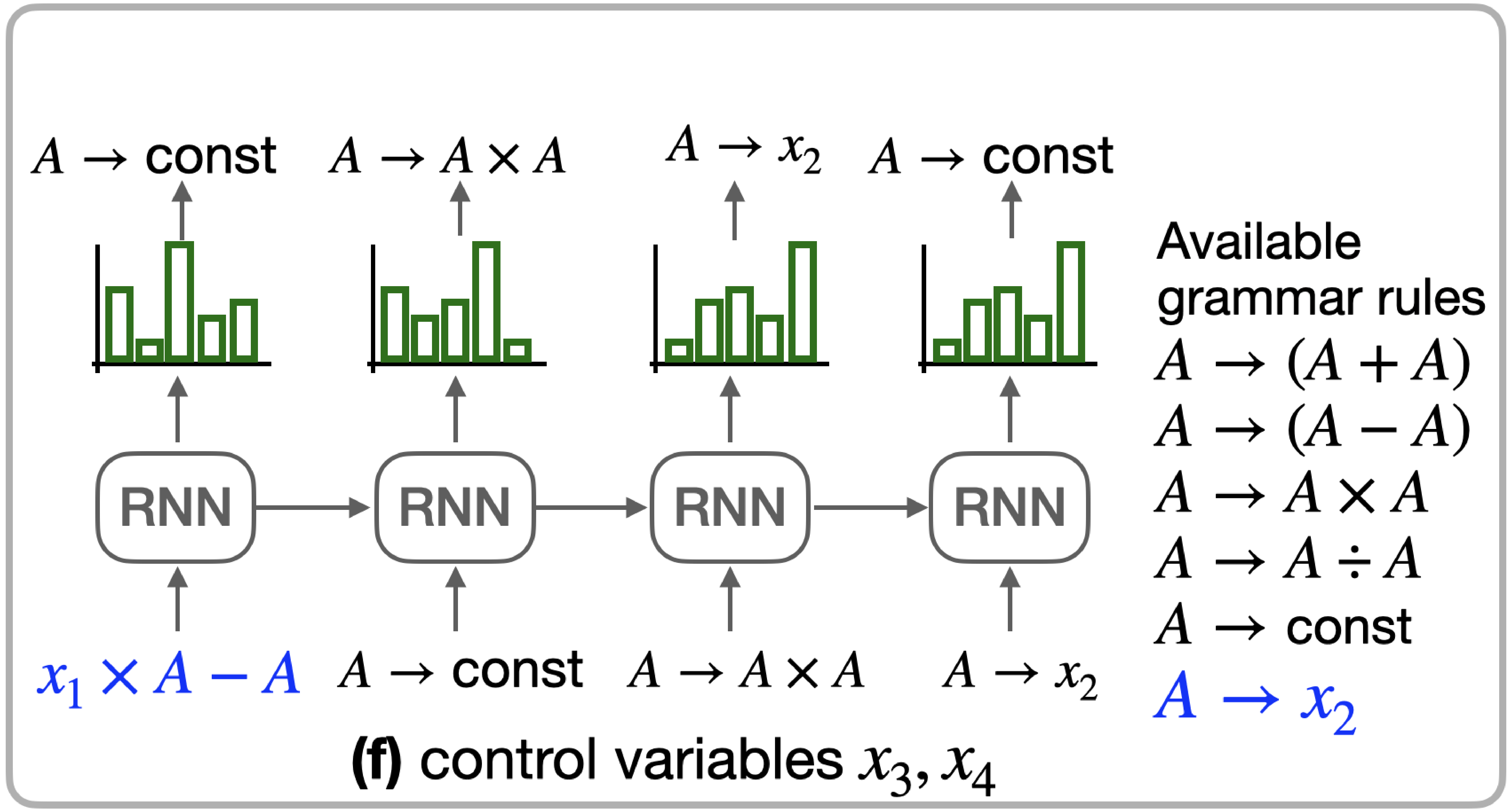


Vertical Symbolic Regression – step 7

The expression is $x_1 \times C_3 - x_2 \times C_4$



(a) control variables x_2, x_3, x_4 **(b)** control variables x_3, x_4



Experiments Analysis

Regression on Algebraic Equations

- Evaluation metric: Median (50%) of NMSE (normalized mean squared error) values.

Our method attains the smallest NMSE values.

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**

Table 1: On selected algebraic equation datasets, median (50%-quartile) of NMSE values of the best-predicted expressions found by all the algorithms. The set of mathematical operator is $O_p = \{+, -, \times, \sin, \cos, \mathtt{const}\}$. The 3-tuples at the top $(\cdot,\cdot,\cdot)$ indicate the number of free variables, singular terms, and cross terms in the ground-truth expressions generating the dataset. O_p stands for the set of allowed operators. "T.O." implies the algorithm is timed out for 48 hours.

Regression on Ordinary 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.

Our method also accelerate the discovery of multivariate ODE.

Conclusion

We integrate vertical discovery with deep neural network.

We propose the use of grammar representation to replace tree representation of expression.

In experiments, we find our method scales better than several baselines to multivariate algebraic equations and ordinary differential equations.

Q & A

https://github.com/jiangnanhugo/VSR-DPG