Vertical AI-driven Scientific Discovery

Anonymous submission

Abstract

Automating scientific discovery has been a grand goal of Artificial Intelligence (AI) and will bring tremendous societal impact if it succeeds. Despite exciting progress, most endeavor in learning scientific equations from experiment data focuses on the *horizontal* discovery paths, i.e., they directly search for the best equation in the full hypothesis space. Horizontal paths are challenging because of the associated exponentially large search space. Our work explores an alternative *vertical* path, which builds scientific equations in an incremental way, starting from one that models data in *control variable experiments* in which most variables are held as constants. It then extends expressions learned in previous generations via adding new independent variables, using new control variable experiments in which these variables are allowed to vary. This vertical path was motivated by human scientific discovery processes. Experimentally, we demonstrate that such vertical discovery paths expedite symbolic regression. It also improves learning physics models describing nano-structure evolution in computational materials science.

Introduction

Automating scientific discovery has been a grand goal of Artificial Intelligence (AI) dating back its founders (Herbert Simon et. al. (Langley et al. 1987; Kulkarni and Simon 1988; Wang et al. 2022)) but remains a holy grail. The underlying societal impact is immense because of its multiplier effect. Indeed, much effort has been made, especially in symbolic equation regression, including search-based methods (Langley 1981; Lenat 1977), genetic programming (Schmidt and Lipson 2009; Virgolin, Alderliesten, and Bosman 2019; Razavi and Gamazon 2022; He et al. 2022), reinforcement learning (Petersen et al. 2021; Scavuzzo et al. 2022; Mundhenk et al. 2021; Petersen et al. 2021), deep function approximation (McConaghy 2011; Chen, Luo, and Jiang 2017; Raissi, Yazdani, and Karniadakis 2020; Raissi, Perdikaris, and Karniadakis 2019; Liu and Tegmark 2021; Xue et al. 2021a; Chen et al. 2018; Jumper et al. 2021; Brunton, Proctor, and Kutz 2016), integrated systems (Valdés-Pérez 1994; King et al. 2004, 2009; Lintott et al. 2008), or simply yet effectively, collecting big datasets (Lintott et al. 2008; Kelling et al. 2012). Most endeavor focuses on *horizontal* discovery paths, i.e., they directly search for the best equation in the full hypothesis space involving all independent variables (red path in Figure 1). The horizontal search can be challenging because of the exponentially large space. After the conventional wisdom of training with larger models and more data has been stretched to its extremity (e.g., GPT-4), what is the next paradigm-changing idea?

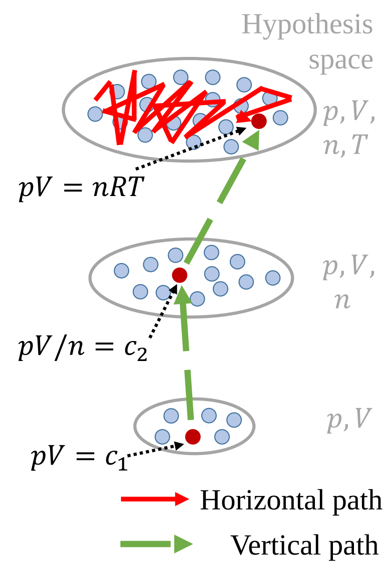

Figure 1: Vertical paths further scale up AI-driven scientific discovery.

Interestingly, the *vertical* paths have been largely overlooked in AI. To discover the ideal gas law $pV = nRT$, scientists first held n (gas amount) and T (temperature) as constants and find p (pressure) is inversely proportional to V (volume). They then studied the relationship between pV and n, T. This led to a vertical discovery path (green path in Figure 1). The first few steps of a vertical path can be significantly cheaper than the horizontal path, because the searches are in reduced spaces involving a small number of independent variables. As a result, vertical discovery has the potential to supercharge state-of-the-art approaches in modeling complex scientific phenomena with more interlocking contributing factors or processes than what current approaches can handle.

This paper demonstrates the power of vertical scientific

X_1	X_2	X_3	Y
2.5	1.0	9.5	12
3.0	-1.0	4.0	1
1.6	3.5	5.2	10.8
1.8	1.0	3.2	5
7.1	8.6	3.8	64.9
1.7	1.0	2.3	4
2.5	2.6	3.1	9.6
8.9	1.1	2.0	11.8
4.2	-1.0	2.2	-2
5.8	1.0	7.2	13
1.6	5.7	1.2	10.3
9.7	-1.0	1.7	-8

(a)

X_1	X_2	X_3	Y
2.5	1.0	9.5	12
1.8	1.0	3.2	5
1.7	1.0	2.3	4
5.8	1.0	7.2	13

$Y=X_1+X_3$

(b)

X_1	X_2	X_3	Y
3.0	-1.0	4.0	1
4.2	-1.0	2.2	-2
9.7	-1.0	1.7	-8

$Y= -X_1+X_3$

(c)

X_1	X_2	X_3	Y
2.5	1.0	9.5	12
3.0	-1.0	4.0	1
1.8	1.0	3.2	5
1.7	1.0	2.3	4
4.2	-1.0	2.2	-2
5.8	1.0	7.2	13
9.7	-1.0	1.7	-8

$Y= X_1X_2 +X_3$

(d)

Figure 2: Motivating example to demonstrate vertical scientific discovery. **(a)** A challenging symbolic regression task. It is difficult to read out the equation $y = f(x_1, x_2, x_3)$ which connects the dependent variable y with the independent variables x_1, x_2, x_3. **(b)** When we focus on studying the relationship of x_1, x_3 and y while holding x_2 at 1, a simple equation $y = x_1 + x_3$ can be discovered. **(c)** $y = -x_1 + x_3$ can be discovered when we hold x_2 at -1. **(d)** Combining (b) and (c), a good candidate equation is $y = x_1x_2 + x_3$, which turns out to be the ground-truth equation.

discovery, in which automated reasoning (e.g., mathematical programming, constraint satisfaction, reinforcement learning, etc) acts as robot scientists to guide the learning process (i.e., pointing out the directions of the green path in Figure 1).

Our first example is in symbolic regression, where the task is to discover symbolic expressions describing experiment data. State-of-the-art approaches in this domain are limited to learning simple expressions. Regressing expressions involving many independent variables still remain out of reach. Motivated by the control variable experiments widely utilized in science, in a recently published paper (Jiang and Xue 2023) we propose **C**ontrol **V**ariable **G**enetic **P**rogramming (CVGP) for symbolic regression over many independent variables. CVGP expedites symbolic expression discovery via customized experiment design, rather than learning from a fixed dataset collected a priori. CVGP starts by fitting simple expressions involving a small set of independent variables using genetic programming, under controlled experiments where other variables are held as constants. It then extends expressions learned in previous generations by adding new independent variables, using new control variable experiments in which these variables are allowed to vary. Experimentally, CVGP outperforms several baselines in learning symbolic expressions involving multiple independent variables.

Our second example is in materials science. Our approach was motivated by tracking and learning the phase-field models describing nano-scale crystalline defect evolution in materials. In a preliminary study, we showed vertical discovery schedules improve the learning of phase-field models for dendritic solidification. 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 remaining parameters do not affect the spatial and temporal dynamics. After this phase, the learning is expanded to all parameters. We demonstrate that the machine learning model is able to discover the ground-truth phase-field model following this vertical schedule, but cannot following the normal schedule (see the Figure in Section).

A Motivating Example

Discovering scientific laws automatically from experiment data has been a grand goal of Artificial Intelligence (AI). Its success will greatly accelerate the pace of scientific discovery. Recently, exciting progress (Schmidt and Lipson 2009; Virgolin, Alderliesten, and Bosman 2019; Guimerà et al. 2020; Petersen et al. 2021; Mundhenk et al. 2021; Petersen et al. 2021; Razavi and Gamazon 2022; He et al. 2022) has been made in this domain, especially taking advantages of the progress in deep neural networks. Consistent strides for higher *throughput* (less time and data required to identify an equation) and better *quality* (equations better fit the data) have been the main drivers in this domain. We notice that almost all prior work follows the horizontal discovery path, which is also the standard machine learning pipeline – first collecting a dataset, then learning the full model, finally evaluating its performance on a separate, yet still fixed test set. Nevertheless, the ***vertical discovery path***, which is heavily utilized by human scientists, is almost forgotten in AI-driven scientific discovery. When studying a complex process involving many interacting subprocesses, scientists always try to isolate each individual process and study their effects separately, via carefully designed control variable experiments. They also use this tool to challenge competing models.

Vertical paths increase the throughput of scientific discovery. Let us verify this assertion from a small human ex-

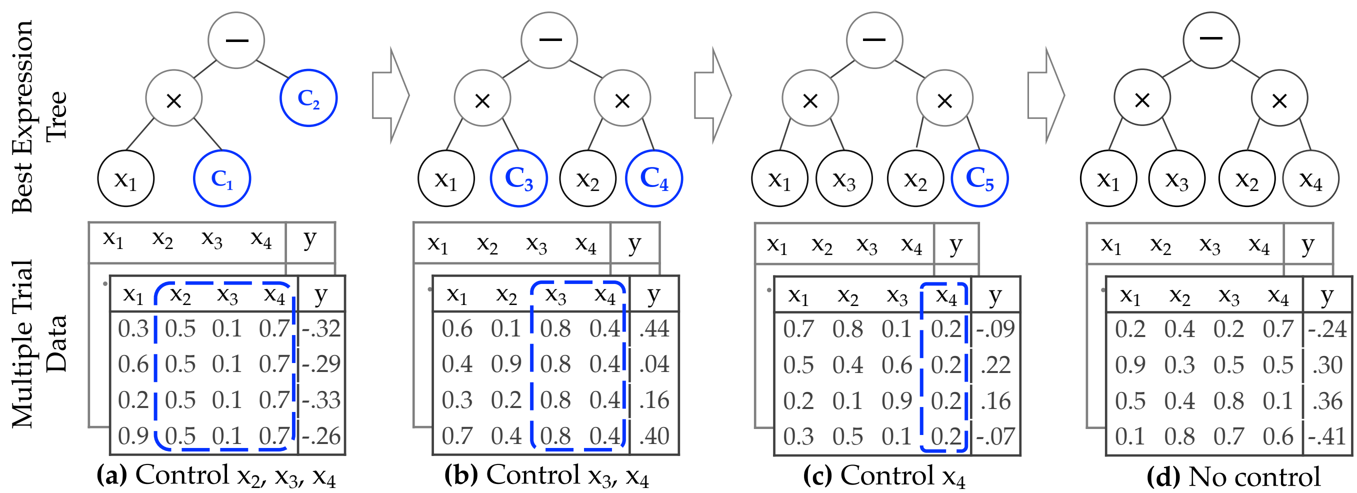

Figure 3: Running example of CVGP. **(a)** Initially, a reduced-form equation $\phi' = C_1 x_1 - C_2$ is found via fitting control variable data in which x_2, x_3, x_4 are held as constants and only x_1 is allowed to vary. **(b)** This equation is expanded to $C_3 x_1 - C_4 x_2$ in the second stage via fitting the data in which only x_3, x_4 are held as constants. **(c,d)** This process continues until the ground-truth equation $\phi = x_1 x_3 - x_2 x_4$ is found. The data generated for control variable experiment trials in each stage are shown at the bottom.

periment. Figure 2 (a) depicts a symbolic regression task where one needs to find a symbolic expression $y = f(x)$ which best maps the input x to the output y. The author ran this experiment in front of hundreds of undergraduate, graduate students, and a few faculty members. Nobody was able to discover the correct equation given the data in (a). However, when the author controlled the value of x_2 in (b) and (c), a majority of the audience were able to identify the equations in both cases. A little bit of additional thinking combining these two equations yields the ground-truth equation in (d). Clearly, control variable experiments in (b) and (c) helped the audience navigate the regression task. This controlled experiment depicts the essence of vertical scientific discovery.

Symbolic Regression via Control Variable Genetic Programming

Our recently proposed **C**ontrol **V**ariable **G**enetic **P**rogramming (CVGP) (Jiang and Xue 2023) implements the vertical scientific discovery process using Genetic Programming (GP) for symbolic regression over many independent variables. The key insight of CVGP is to learn from *a customized set of control variable experiments*; in other words, the experiment data collection adapts to the learning process. This is in contrast to the current learning paradigm of most symbolic regression approaches, where they learn from a fixed dataset collected a priori.

In CVGP, first, we hold all independent variables except for one as constants and learn an expression that maps the single variable to the dependent variable using GP. GP maintains a pool of candidate expressions and improves the fitness of these equations via mating, mutating, and selection over several generations. Mapping the dependence of one independent variable is easy. Hence GP can usually recover the ground-truth reduced-form equation. Then, CVGP frees one independent variable at a time. In each iteration, GP is used to modify the equations learned in previous generations to incorporate the new independent variable, via mating, mutating, and selection. Such a procedure repeats until all the independent variables have been incorporated into the symbolic expression. See figure 3 for the high-level idea of algorithm execution. Theoretically, in the original paper we show CVGP as an incremental builder can reduce the exponential-sized search space for candidate expressions into a polynomial one when fitting a class of symbolic expressions. Experimentally, we show CVGP outperforms a number of state-of-the-art approaches on symbolic regression over multiple independent variables (see Table 1).

Vertical Scientific Discovery in Modeling Nano-structure Evolution in Materials Science

We intend to apply the idea of vertical scientific discovery in learning nano-scale defect evolution for material under extreme conditions. Nano-scale crystalline defects can appear in different forms in these materials. Extreme environments of heat and irradiation can cause these defects to evolve in size and position. As shown in the left panel of the figure above, void shaped defects are captured by transmission electron microscope (TEM) cameras during in-situ radiation experiments. These defects appear in round shapes, and drift in position as demonstrated by the change of angles to respectively, as time progresses. They also change size. These changes can affect the physical and mechanical properties of the material. For this reason, characterizing these defects is essential in designing new materials that can resist adverse environments. Collaborating with materials scientists, we have been analyzing terabytes of in-situ TEM videos of this type and have already made scientific discoveries (Sima

Table 1: Median (50%) and 75%-quantile Normalized Mean Squared Error (NMSE) values of the symbolic expressions found by all the algorithms on several *noisy* benchmark datasets (Gaussian noise with zero mean and standard deviation 0.1 is added). Our CVGP finds symbolic expressions with the smallest NMSEs.

Dataset configs	CVGP (ours) 50%	CVGP (ours) 75%	GP 50%	GP 75%	DSR 50%	DSR 75%	PQT 50%	PQT 75%	VPG 50%	VPG 75%	GPMeld 50%	GPMeld 75%
(4,4,6)	**0.036**	**0.088**	0.038	0.108	1.163	3.714	1.016	1.122	1.087	1.275	1.058	1.374
(5,5,5)	0.076	0.126	**0.075**	**0.102**	1.028	2.270	1.983	4.637	1.075	2.811	1.479	2.855
(5,5,8)	**0.061**	**0.118**	0.121	0.186	1.004	1.013	1.005	1.006	1.002	1.009	1.108	2.399
(6,6,8)	**0.098**	**0.144**	0.104	0.167	1.006	1.027	1.006	1.020	1.009	1.066	1.035	2.671
(6,6,10)	**0.055**	**0.097**	0.074	0.132	1.003	1.009	1.005	1.008	1.004	1.015	1.021	1.126
(a) Datasets containing operators $\{\sin, \cos, \mathtt{inv}, +, -, \times\}$.												
(3,2,2)	**0.098**	**0.165**	0.108	0.425	0.350	0.713	0.351	1.831	0.439	0.581	0.102	0.597
(4,4,6)	**0.078**	**0.121**	0.120	0.305	7.056	16.321	5.093	19.429	2.458	13.762	2.225	3.754
(5,5,5)	**0.067**	**0.230**	0.091	0.313	32.45	234.31	36.797	229.529	14.435	46.191	28.440	421.63
(5,5,8)	**0.113**	**0.207**	0.119	0.388	195.22	573.33	449.83	565.69	206.06	629.41	363.79	666.57
(6,6,8)	**0.170**	**0.481**	0.186	0.727	1.752	3.824	4.887	15.248	2.396	7.051	1.478	6.271
(6,6,10)	**0.161**	**0.251**	0.312	0.342	11.678	26.941	5.667	24.042	7.398	25.156	11.513	28.439
(b) Datasets containing operators $\{\sin, \cos, +, -, \times\}$.												
(3,2,2)	0.049	**0.113**	**0.023**	0.166	0.663	2.773	1.002	1.992	0.969	1.310	0.413	2.510
(4,4,6)	**0.141**	**0.220**	0.238	0.662	1.031	1.051	1.297	1.463	1.051	1.774	1.093	1.769
(5,5,5)	**0.157**	0.438	0.195	**0.337**	1.098	3.617	1.018	5.296	1.012	1.27	1.036	3.617
(5,5,8)	**0.122**	**0.153**	0.166	0.186	1.009	1.103	1.017	1.429	1.007	1.132	1.07	2.904
(6,6,8)	**0.209**	**0.590**	**0.209**	0.646	1.003	1.153	1.047	1.134	1.059	1.302	1.029	3.365
(6,6,10)	0.139	0.232	**0.073**	**0.159**	1.654	3.408	1.027	1.069	1.009	1.654	1.445	2.106
(c) Datasets containing operators $\{\sin, \cos, \mathtt{inv}, +, -, \times\}$.												

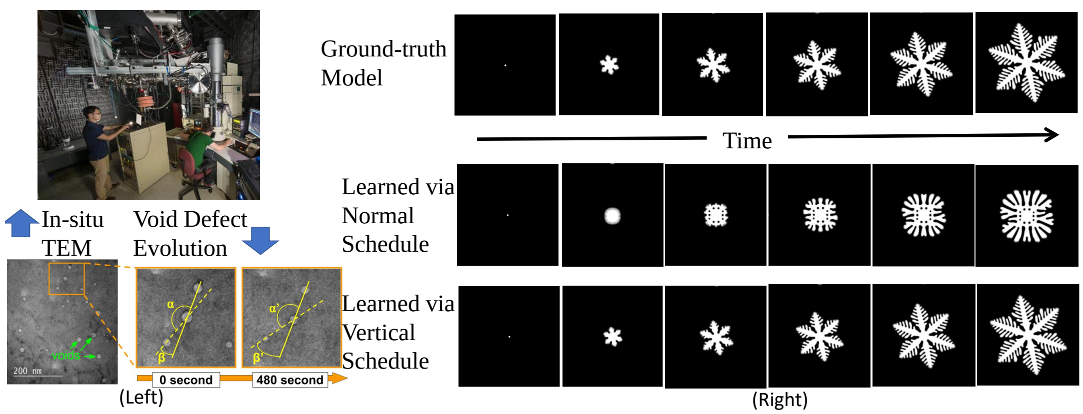

Figure 4: Vertical scientific discovery pipeline can improve learning of physics models behind nano-structure evolution. Left: The Intermediate Voltage Electron Microscopy (IVEM) – Tandem Facility at the Argonne National Laboratory which provides in-situ TEM data. Source: anl.gov. (top) and sample images captured during in-situ radiation experiments (bottom). Right: Learned physics models from vertical scientific discovery pipeline can yield more accurate simulation output close to ground truth compared to baseline approach.

and Xue 2021; Xue et al. 2021b; Nasim et al. 2022, 2023).

As a preliminary study, vertical discovery schedules are used to improve the learning of phase-field models for dendritic solidification. 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.

References

Brunton, S. L.; Proctor, J. L.; and Kutz, J. N. 2016. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15): 3932–3937.

Chen, C.; Luo, C.; and Jiang, Z. 2017. Elite bases regression: A real-time algorithm for symbolic regression. In *ICNC-FSKD*, 529–535. IEEE.

Chen, R. T.; Rubanova, Y.; Bettencourt, J.; and Duvenaud, D. K. 2018. Neural ordinary differential equations. *Advances in neural information processing systems*, 31.

Guimerà, R.; Reichardt, I.; Aguilar-Mogas, A.; Massucci, F. A.; Miranda, M.; Pallarès, J.; and Sales-Pardo, M. 2020. A Bayesian machine scientist to aid in the solution of challenging scientific problems. *Science advances*, 6(5): eaav6971.

He, B.; Lu, Q.; Yang, Q.; Luo, J.; and Wang, Z. 2022. Taylor genetic programming for symbolic regression. In *GECCO*, 946–954. ACM.

Jiang, N.; and Xue, Y. 2023. Symbolic Regression via

Control Variable Genetic Programming. In *Proceedings of the 2023 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, Lecture Notes in Computer Science. Springer.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873): 583–589.

Kelling, S.; Gerbracht, J.; Fink, D.; Lagoze, C.; Wong, W.-K.; Yu, J.; Damoulas, T.; and Gomes, C. P. 2012. eBird: A Human/Computer Learning Network for Biodiversity Conservation and Research. In *Proceedings of the Twenty-Fourth Conference on Innovative Applications of Artificial Intelligence (IAAI)*.

King, R. D.; Rowland, J.; Oliver, S. G.; Young, M.; Aubrey, W.; Byrne, E.; Liakata, M.; Markham, M.; Pir, P.; Soldatova, L. N.; Sparkes, A.; Whelan, K. E.; and Clare, A. 2009. The Automation of Science. *Science*, 324(5923): 85–89.

King, R. D.; Whelan, K. E.; Jones, F. M.; Reiser, P. G.; Bryant, C. H.; Muggleton, S. H.; Kell, D. B.; and Oliver, S. G. 2004. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427(6971): 247–252.

Kulkarni, D.; and Simon, H. A. 1988. The processes of scientific discovery: The strategy of experimentation. *Cognitive science*, 12(2): 139–175.

Langley, P. 1981. Data-driven discovery of physical laws. *Cognitive Science*, 5(1): 31–54.

Langley, P. W.; Simon, H. A.; Bradshaw, G.; and Zytkow, J. M. 1987. *Scientific Discovery: Computational Explorations of the Creative Process*. The MIT Press. ISBN 9780262316002.

Lenat, D. B. 1977. The ubiquity of discovery. *Artificial Intelligence*, 9(3): 257–285.

Lintott, C.; Schawinski, K.; Slosar, A.; Land, K.; Bamford, S.; Thomas, D. I.; Raddick, M.; Nichol, R.; Szalay, A.; Andreescu, D.; Murray, P.; and Berg, J. V. D. 2008. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 389: 1179–1189.

Liu, Z.; and Tegmark, M. 2021. Machine Learning Conservation Laws from Trajectories. *Phys. Rev. Lett.*, 126: 180604.

McConaghy, T. 2011. FFX: Fast, scalable, deterministic symbolic regression technology. In *Genetic Programming Theory and Practice IX*, 235–260. Springer.

Mundhenk, T. N.; Landajuela, M.; Glatt, R.; Santiago, C. P.; Faissol, D. M.; and Petersen, B. K. 2021. Symbolic Regression via Deep Reinforcement Learning Enhanced Genetic Programming Seeding. In *NeurIPS*, 24912–24923.

Nasim, M.; Rayaprolu, S.; Niu, T.; Fan, C.; Shang, Z.; Li, J.; Wang, H.; El-Azab, A.; Xue, Y.; and Zhang, X. 2023. Unraveling the size fluctuation and shrinkage of nanovoids during in situ radiation of Cu by automatic pattern recognition and phase field simulation. *Journal of Nuclear Materials*, 574: 154–189.

Nasim, M.; Zhang, X.; El-Azab, A.; and Xue, Y. 2022. Efficient Learning of Sparse and Decomposable PDEs using Random Projection. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Petersen, B. K.; Landajuela, M.; Mundhenk, T. N.; Santiago, C. P.; Kim, S.; and Kim, J. T. 2021. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. In *ICLR*. OpenReview.net.

Raissi, M.; Perdikaris, P.; and Karniadakis, G. 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378: 686–707.

Raissi, M.; Yazdani, A.; and Karniadakis, G. E. 2020. Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. *Science*, 367(6481): 1026–1030.

Razavi, S.; and Gamazon, E. R. 2022. Neural-Network-Directed Genetic Programmer for Discovery of Governing Equations. *CoRR*, abs/2203.08808.

Scavuzzo, L.; Chen, F. Y.; Chételat, D.; Gasse, M.; Lodi, A.; Yorke-Smith, N.; and Aardal, K. 2022. Learning to branch with Tree MDPs. *CoRR*, abs/2205.11107.

Schmidt, M.; and Lipson, H. 2009. Distilling Free-Form Natural Laws from Experimental Data. *Science*, 324(5923): 81–85.

Sima, C.; and Xue, Y. 2021. LSH-SMILE: Locality Sensitive Hashing Accelerated Simulation and Learning. In *Proceedings of 35th Conference on Neural Information Processing Systems (NeurIPS)*.

Valdés-Pérez, R. 1994. Human/computer interactive elucidation of reaction mechanisms: application to catalyzed hydrogenolysis of ethane. *Catalysis Letters*, 28: 79–87.

Virgolin, M.; Alderliesten, T.; and Bosman, P. A. N. 2019. Linear scaling with and within semantic backpropagation-based genetic programming for symbolic regression. In *GECCO*, 1084–1092. ACM.

Wang, H.; Fu, T.; Du, Y.; Gao, W.; Huang, K.; Liu, Z.; Chandak, P.; Liu, S.; Katwyk, P. V.; Deac, A.; Anandkumar, A.; Bergen, K.; Gomez, C. P.; Ho, S.; Kohli, P.; Lasenby, J.; Leskovec, J.; Liu, T.-Y.; Manrai, A.; Marks, D.; Ramsundar, B.; Song, L.; Sun, J.; Tang, J.; Velickovic, P.; Welling, M.; Coley, C.; Bengio, Y.; and Zitnik, M. 2022. Enabling Scientific Discovery with Artificial Intelligence. *Nature*.

Xue, Y.; Nasim, M.; Zhang, M.; Fan, C.; Zhang, X.; and El-Azab, A. 2021a. Physics Knowledge Discovery via Neural Differential Equation Embedding. In *ECML/PKDD (5)*, volume 12979 of *Lecture Notes in Computer Science*, 118–134. Springer.

Xue, Y.; Nasim, M.; Zhang, M.; Fan, C.; Zhang, X.; and El-Azab, A. 2021b. Physics Knowledge Discovery via Neural Differential Equation Embedding. In *Proceedings of 2021 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*.