Vertical AI-driven Scientific Discovery Integrating Reasoning, Learning and Human Computation

The research objective of this proposal is to explore novel *integrations* of automated reasoning, machine learning and human computation to reduce the time to discover new scientific equations from years relying on scientists' manual effort down to minutes. Automating scientific discovery has been a grand goal dating back the founders of AI (Herbert Simon et. al. [14, 12, 30]) but remains a holy grail. The underlying societal impact is immense because of its multiplier effect. Indeed, much effort has been made, especially in equation regression, including search-based methods [13, 15], genetic programming [27, 29, 25, 5], reinforcement learning [21, 26, 20, 21], deep function approximation [19, 2, 23, 22, 17, 31, 3, 7, 1], integrated systems [28, 10, 9, 16], or simply yet effectively, collecting huge datasets [16, 8]. 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?

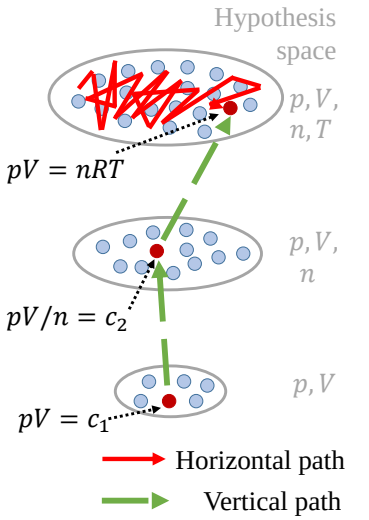

Figure 1: Vertical paths will further scale up AI-driven scientific discovery, after integrating reasoning, learning and human computation in hypothesis forming, experiment design, model learning and diagnosis.

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. Notice vertical discovery requires more than a pre-collected dataset and a clever machine learning algorithm. Instead, it calls for a tight integration of hypothesis forming, experimental designing, model learning and diagnosis.

The PI's CAREER goal is to study the **science to scale up vertical AI-driven scientific discovery**, **integrating automated reasoning, machine learning, and human computation**. Automated reasoning (human computation) acts as robot (citizen) scientists to guide the vertical scientific discovery process. Automated reasoning includes optimization, mathematical programming, constraint satisfaction, reinforcement learning etc. Human computation refers to games that engage both domain experts and the general public in forming good hypotheses, experiment designs, etc.

The proposed research will answer (1) how to identify the optimal vertical discovery path (*hypothesis forming*)? (2) What data best reveal the effect of the variables and processes at study (*experiment design*)? (3) How to identify the equations (*model learning*)? (4) Are there better or alternative models (*model diagnosis*)? In the preliminary work, the PI showed symbolic regression can be accelerated via expert-identified vertical discovery paths. In this proposal, new automated reasoning algorithms and human computation games will be developed to identify vertical paths better than the experts' (*hypothesis forming*), search for experiments that reveal inductive processes of PDEs (*experiment design*), and learn PDE models (*model learning*). A parallel vertical path based on hashing and randomization will be investigated to reason about alternative models (*model diagnosis*), building on the PI's strong record in probabilistic inference with provable guarantees. New solvers for leader-follower games that engage citizen scientists in scientific discovery will also be investigated. The potential of transformative changes from horizontal to vertical AI-driven scientific discovery will establish the PI as a leader and has long-lasting effects beyond five years.

Broader Impacts The goal of broader impact activities is to demonstrate the power of automated reasoning and human computation-empowered learning tools in scaling up scientific discovery in fields of critical national importance, establish common languages and shared research problems and skills between computer and applied-domain scientists and educators, nurture next-generation computer and applied-domain scientists, and increase public literacy and awareness in science. In computational materials science [24, 18, 4, 6, 11], the developed tools will produce more accurate physics models describing nano-scale structure evolution in materials under extreme conditions. They will also refine scalable physics models from their computational-intensive counterparts. The developed games that involve citizen scientists in scientific discovery will increase public participation to scale up scientific discovery.

Integration of Research and Education. The education activities are designed to address the gap of lacking common languages and research interests among computer and materials science students and public participation in science, while leveraging existing institutional resources. Specifically, we propose:

1. Host *"hypothesis forming"* challenge to boost materials knowledge discovery. This challenge asks both computer and materials science students to form new hypotheses and eventually new vertical discovery paths, fusing the information from AI tools.

2. Organize *"criticize models with experiments"* competition. This competition asks teams of computer and materials science students to criticize other teams' models by coming up with competing examples, experiments, etc, using the AI-human integrated loop.

3. Run *"gamify AI-driven scientific discovery"* competition, which asks undergraduate students from Purdue Data-mine to encode the vertical scientific discovery process into games for K-12 students.

4. *Course development and talk series* to provide students from CS and science domains, especially those from minority backgrounds, opportunities to progress jointly in this interdisciplinary domain.

References

[1] Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016.

[2] Chen Chen, Changtong Luo, and Zonglin Jiang. Elite bases regression: A real-time algorithm for symbolic regression. In *ICNC-FSKD*, pages 529–535. IEEE, 2017.

[3] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[4] Carla P Gomes, Bart Selman, and John M Gregoire. Artificial intelligence for materials discovery. *MRS Bulletin*, 44(7):538–544, 2019.

[5] Baihe He, Qiang Lu, Qingyun Yang, Jake Luo, and Zhiguang Wang. Taylor genetic programming for symbolic regression. In *GECCO*, pages 946–954. ACM, 2022.

[6] Qing-Miao Hu and Rui Yang. The endless search for better alloys. *Science*, 378(6615):26–27, 2022.

[7] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[8] Steve Kelling, Jeff Gerbracht, Daniel Fink, Carl Lagoze, Weng-Keen Wong, Jun Yu, Theodoros Damoulas, and Carla P. Gomes. 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)*, 2012.

[9] Ross D. King, Jem Rowland, Stephen G. Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N. Soldatova, Andrew Sparkes, Kenneth E. Whelan, and Amanda Clare. The automation of science. *Science*, 324(5923):85–89, 2009.

[10] Ross D King, Kenneth E Whelan, Ffion M Jones, Philip GK Reiser, Christopher H Bryant, Stephen H Muggleton, Douglas B Kell, and Stephen G Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427(6971):247–252, 2004.

[11] Heather J Kulik and Pratyush Tiwary. Artificial intelligence in computational materials science. *MRS Bulletin*, pages 1–3, 2022.

[12] Deepak Kulkarni and Herbert A Simon. The processes of scientific discovery: The strategy of experimentation. *Cognitive science*, 12(2):139–175, 1988.

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

[14] Patrick W. Langley, Herbert A. Simon, Gary Bradshaw, and Jan M. Zytkow. *Scientific Discovery: Computational Explorations of the Creative Process*. The MIT Press, 02 1987.

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

[16] C. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, Daniel I. Thomas, M. Raddick, R. Nichol, A. Szalay, D. Andreescu, P. Murray, and J. V. D. Berg. 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, 2008.

[17] Ziming Liu and Max Tegmark. Machine learning conservation laws from trajectories. *Phys. Rev. Lett.*, 126:180604, May 2021.

[18] Arun Mannodi-Kanakkithodi and Maria KY Chan. Computational data-driven materials discovery. *Trends in Chemistry*, 3(2):79–82, 2021.

[19] Trent McConaghy. Ffx: Fast, scalable, deterministic symbolic regression technology. In *Genetic Programming Theory and Practice IX*, pages 235–260. Springer, 2011.

[20] T. Nathan Mundhenk, Mikel Landajuela, Ruben Glatt, Cláudio P. Santiago, Daniel M. Faissol, and Brenden K. Petersen. Symbolic regression via deep reinforcement learning enhanced genetic programming seeding. In *NeurIPS*, pages 24912–24923, 2021.

[21] Brenden K. Petersen, Mikel Landajuela, T. Nathan Mundhenk, Cláudio Prata Santiago, Sookyung Kim, and Joanne Taery Kim. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. In *ICLR*. OpenReview.net, 2021.

[22] M. Raissi, P. Perdikaris, and G.E. Karniadakis. 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, 2019.

[23] Maziar Raissi, Alireza Yazdani, and George Em Karniadakis. Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. *Science*, 367(6481):1026–1030, 2020.

[24] ZY Rao, X Wang, J Zhu, XH Chen, L Wang, JJ Si, YD Wu, and XD Hui. Affordable fecrnimncu high entropy alloys with excellent comprehensive tensile properties. *Intermetallics*, 77:23–33, 2016.

[25] Shahab Razavi and Eric R. Gamazon. Neural-network-directed genetic programmer for discovery of governing equations. *CoRR*, abs/2203.08808, 2022.

[26] Lara Scavuzzo, Feng Yang Chen, Didier Chételat, Maxime Gasse, Andrea Lodi, Neil Yorke-Smith, and Karen Aardal. Learning to branch with tree mdps. *CoRR*, abs/2205.11107, 2022.

[27] Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.

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

[29] Marco Virgolin, Tanja Alderliesten, and Peter A. N. Bosman. Linear scaling with and within semantic backpropagation-based genetic programming for symbolic regression. In *GECCO*, pages 1084–1092. ACM, 2019.

[30] Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P. Gomez, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, Debora Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Velickovic, Max Welling, Connor Coley, Yoshua Bengio, and Marinka Zitnik. Enabling scientific discovery with artificial intelligence. *Nature*, 2022.

[31] Yexiang Xue, Md. Nasim, Maosen Zhang, Cuncai Fan, Xinghang Zhang, and Anter El-Azab. Physics knowledge discovery via neural differential equation embedding. In *ECML/PKDD (5)*, volume 12979 of *Lecture Notes in Computer Science*, pages 118–134. Springer, 2021.