Introduction
The need for robotic science is growing rapidly, as exemplified by a central challenge of materials discovery. Advances in technology often require better materials. However, scientists are quickly exhausting the materials that are simpler to make, that is, materials of simple stoichiometry (few elements) and few processing steps. As a result, scientists are driven to explore materials of greater complexity. With each new synthesis or processing parameter, the number of possible materials grows exponentially. This growing number of materials holds great promise but comes with a significant challenge—a rapidly growing number of materials to explore. The traditional Edisonian search for better materials consists of one-by-one materials synthesis, characterization, and data analysis. This approach to materials discovery becomes infeasible as the search space grows. High-throughput methods and machine learning (ML) make it possible to synthesize, characterize, and analyze hundreds of materials in days, but this geometric speedup cannot keep up with the exponential challenge.

The most recent innovation in the search for advanced materials is the use of active learning—the optimal experiment design field of ML. Active learning-based recommendation engines guide experiments both in the laboratory and in silico, accelerating the discovery of novel materials. By integrating active learning with automated tools, closed-loop autonomous physical science (APS) systems—that is, robot scientists—become possible. For these systems, each subsequent materials experiment is selected and executed to maximize knowledge for the user. For example, APS systems have been used in biology to optimize synthesis routes for yeast enzymes and in chemistry to identify thin-film molecular mixtures with improved photoactivity. Additionally, for the first time an APS system has discovered a best-in-class solid-state material—the new best-in-class phase-change memory material. Autonomous techniques have also advanced measurement science,

Figure 1. (a) Image of student projects from a University of Maryland course in machine learning for materials science. Students were split into groups and each group worked with a robot scientist. (b) Image of the robot scientist with a pH sensor.
Scientific societies have begun designing dedicated ML tutorials at workshops and conferences. Some standalone resources exist. We run a boot camp—the Machine Learning for Materials Research. The five-day, hands-on boot camp is now in its seventh year, teaching programming and ML fundamentals through APS skills to students, academic, national laboratory, and industry scientists. We also host the REsource for Materials Informatics (REMI) website, which curates online materials informatics and APS tools including data science Jupyter notebooks.

Although the list of education resources is growing, there is something lacking—a low-cost, easy-to-use APS platform on which APS can be learned, demonstrated, and explored. There is a growing number of commercial APS systems, but they cost hundreds of thousands of dollars. Systems demonstrated in the literature are costly, bespoke, and involve highly complex equipment, making them inappropriate for the classroom. Deneault et al. demonstrated a potential educational tool—a closed-loop autonomous three-dimensional (3D) printer, capable of autonomously optimizing the printed line profile, with a preliminary cost of over $2,500. The authors also discuss the future aim to build a lower cost version for demonstrating physical science experiments.

An education platform must be lower-cost, robust, and easy-to-use for teaching the many needed APS skills. Such a system can also serve industry for APS methods development.

The list of APS skills required of the next-generation workforce is long. Physical science skills include experimental design, materials synthesis, characterization, and analysis. Data science skills are needed to ensure that collected data are findable, accessible, interoperable, and reusable. A next-generation scientist will need to know the basics of software and hardware design, such as designing for safety when dealing with toxic or delicate materials. Control systems knowledge is required to manage APS motion and environmental controls. ML knowledge is required for closed-loop data analysis, prediction, and experiment design. Knowledge of uncertainty quantification and propagation is also key, as proper uncertainty handling improves ML performance including active learning-based experiment design.

Scientific ML—which is also known as inductive bias ML—is another key knowledge domain for the next-generation workforce. Many physical science challenges suffer from a sparsity of data despite having an abundance of prior knowledge encoded in physical laws and heuristics. Scientific ML incorporates this prior knowledge into the ML pipeline to achieve physically meaningful analysis, prediction, and experiment design with greater performance. As a result, scientific ML can greatly reduce the number of APS experiments needed to achieve a user-defined goal. An optimal APS education platform should allow students and professionals to learn and develop novel scientific ML algorithms.

We present the next generation in science education and APS methodology development, LEGOLAS—a LEGO-based Low-Cost Autonomous Science platform. LEGOLAS is an easy-to-use, modular system based on modular toy parts, 3D printed parts, Raspberry Pi components, and aluminum extrusions with a total cost of less than $300, that is, an APS minimum

accelerating x-ray and neutron diffraction at multiple user facilities. Diverse companies now seek to use APS in their R&D pipelines.

A next-generation workforce is needed to fuel the growth of APS. However, with the rapid development of APS, educators have been left behind. Universities are scrambling to integrate next-generation-workforce data science skills into their courses in traditional non-data-centric disciplines.

Figure 2. Image of the applied physical science cycle used in the experiment. Samples are synthesized by mixing acid and base in a well, a sensor measures pH, the data are analyzed, and the next experiment is chosen.

Figure 3. (a) Gaussian process (GP) fit to experiment data. The hidden true function (black dashed), GP mean (blue line), and GP 95% confidence interval (blue region) are indicated. (b) Exploratory acquisition function for Challenge 1. (c) Acquisition function that balances exploration with optimization, a solution for Challenge 2. The next sample to select is indicated by the dashed purple line, with an acid/base ratio of approximately 15 selected for Challenge 1 and approximately 3 for Challenge 2.
viable product. The system was inspired by the work of Gerber et al. \(^9\) to build an education set for teaching chemistry. LEGOLAS was used to teach undergraduates and graduate students during two ML courses at the University of Maryland. Students first learned the fundamentals of APS through lectures and hands-on exercises with Jupyter notebooks. They were then split into groups to work hands-on with the LEGOLAS kits to solve challenges of APS exploration and discovery. The robot scientist was adapted for liquid handling and pH testing, allowing the students to apply their APS knowledge to investigate the relationship between acid-base ratio and the resulting pH, a relationship described (within a certain mixture range) by the Henderson–Hasselbalch (HH) equation. The acid and base used are on the level of vinegar and milk of magnesia, respectively, so the following experiments can be performed with food-safe, household liquids. The students were also provided template APS code, which they were asked to modify to solve the challenges.

**Figure 1** shows a photo from the course and a photo of LEGOLAS with a pH sensor. **Figure 2** presents the LEGOLAS workflow used by the students. ML is used to analyze previous data and active learning is used to determine the next acid-base ratio to investigate. Automated liquid handling creates the desired sample, and automated sensing measures the sample pH. The new data are then used to update the ML model and guide subsequent experiment design, that is, mixing ratio of acid to base. In this article we discuss the course-based projects provided to the students for synthesis–property relationship exploration. We also present more advanced uses for the kit including on-the-fly hypothesis design and validation using probabilistic physical models, that is, identifying the mechanistic rule underlying the studied relationship. Alternatively, for a simpler challenge for younger students, LEGOLAS has also been used to autonomously match a given color through a mounted camera and access to water with red, green, and blue food coloring. We aim to make LEGOLAS available to the community, along with

---

**Figure 4.** (a–c) Bayesian inference and active learning are combined for parameter determination with (a, b) distributions over the parameters and (c) the function. (d–f) Bayesian inference and active learning are combined for model and parameter determination with (d) distributions over possible models (inset: sum over distributions), (e) entropy as a function of ratio, and (f) Bayesian information criteria (BIC) for model determination.
teaching material and a vibrant software ecosystem. A video of LEGOLAS performing the closed-loop pH experiments can be seen at https://youtu.be/TtPM7zXi5kQ.

**Use of the education kit**

The students were asked to solve two challenges: (1) To write a code and operate LEGOLAS, so that it can autonomously identify the relationship between acid-to-base ratio and the resulting pH using the minimum number of experiments. (2) To modify the code so that LEGOLAS identifies the acid-to-base ratio with a pH of 4.5 using the minimum number of experiments. They were also introduced to scientific challenges, such as noisy pH sensor measurements. Each student group was asked to write a Jupyter notebook to address the two challenges. Figure 3 shows a solution to the challenges, where off-the-shelf Gaussian process regression (using a Matern $5/2$ basis function kernel for its flexibility) was fit to the data and (b) paired with an exploratory active learning acquisition function for Challenge 1 and (c) paired with an acquisition function that balances exploration and exploitation for Challenge 2. The two acquisition functions recommend different subsequent experiments.

At the end of the three-week project, students presented their results and their APS strategies and discussed opportunities for future improvements. The students found that with APS, they were able to solve the challenges with less than 10 experiments, compared to an exhaustive study that takes dozens of experiments. Similarly, total experiment time fell from hours to minutes.

Three advanced challenges were provided to an undergraduate student to solve using the LEGOLAS. In the first challenge, the student was provided the HH equation (Equation 1) and asked to use LEGOLAS to find the parameter values in the minimum number of experiments. The student used Bayesian inference and active learning to guide subsequent experiments and focus on the correct parameter values. Bayesian inference uses Bayes rule (Equation 2) and probabilistic sampling to estimate unknown distributions. Here $D$ are the past data and $M(\theta)$ is the model $M$ with parameters $\theta$. Figure 4a and 4b show an example of the learned distributions over the parameters $\alpha$ and $\beta$, and Figure 4c shows the resulting distribution over the HH model. A more advanced version of this was recently implemented for APS neutron scattering to identify magnetic dynamic parameters. 11

$$\text{pH} = \alpha + \beta \log x, x = \frac{[\text{Acid}]}{[\text{Base}]}$$  

Figure 5. Symbolic regression combined with active learning for probabilistic model determination. (a) Example data, (b) output from symbolic regression with five models. The model with the highest score matches the Henderson–Hasselbalch equation with a slight deviation of parameters. MSE, mean-square error.
entropy selection criteria from the previous exercise. The next experiment is performed, data collected, and the cycle is repeated. An example output is given in Figure 5, where five models provide adequate fits to the data. The model with the best score is an extremely close fit to the HH equation.

Summary

LEGOLAS was proven to be an excellent low-cost education platform for teaching APS skills in two courses at the University of Maryland. APS skills taught include automation, closed-loop experimentation, systems control, and software design among others. Students learned and executed autonomous ML and scientific ML at the University of Maryland. APS teaching was also incorporated into the previously mentioned Machine Learning for Materials Research boot camp, expanding the community beyond students to academic, national laboratory, and industry scientists.

References

1. R. Potyrajlo, K. Rajan, K. Stoewe, I. Takeuchi, B. Chisholm, H. Lam, Combinatorial and high-throughput screening of materials libraries: Review of state of the art. ACS Comb. Sci. 13(6), 579 (2011)
2. E. Stach, B. DeCost, A.G. Kusne, J. Hat-trick-Simpers, K.A. Brown, K.G. Reyes, J. Schrier, S. Billinge, T. Buonassisi, I. Foster, C.P. Gomes, J.M. Gregoire, A. Mehta, J. Montoya, E. Olivetti, C. Park, E. Rotenberg, S.K. Saikin, S. Smullin, V. Stanev, B. Maruyama, Autonomous experimentation systems for materials development: A community perspective. Matter 4(9), 2702 (2021)
3. A.G. Kusne, H. Yu, C. Wu, H. Zhang, J. Hattrick-Simpers, B. DeCost, S. Sarker, C. Oses, C. Toher, S. Curtarolo, A.V. Davydov, R. Agarwal, L.A. Bendersky, M. Li, A. Mehta, I. Takeuchi, On-the-fly closed-loop materials discovery via Bayesian active learning. Nat. Commun. 11, 5966 (2020)
4. Nanocenter, University of Maryland, Machine Learning for Materials Research Boot Camp & Workshop on Autonomous Materials Research (n.d.). https://www.nanocenter.umd.edu/events/mlmr/
5. REMF: Resource for Materials Informatics (n.d.). https://pages.nist.gov/remf/
6. J.R. Deneault, J. Chang, J. Myung, D. Hooper, A. Armstrong, M. Pitt, B. Maruyama, Toward autonomous additive manufacturing: Bayesian optimization on a 3D printer. MRS Bull. 46(7), 566 (2021)
7. M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Herina, P.A.C. ’t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Vel- terop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Meza. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018 (2016)
8. B.L. DeCost, J.R. Hattrick-Simpers, Z. Trautt, A.G. Kusne, E. Campo, M.L. Green, Scientific AI in materials science: A path to a sustainable and scalable paradigm. Mach. Learn. Sci. Technol. 1(3), 033001 (2020)
9. L.C. Gerber, A. Calasanz-Kaiser, L. Hyman, K. Voitiuk, U. Patil, I.H. Riedel-Kruse, Liquid-handling Lego robots and experiments for STEM education and research. PLoS Biol. 15(3), e2001413 (2017)
10. C.E. Rasmussen, “Gaussian Processes in Machine Learning,” in Advanced Lectures on Machine Learning, ML Summer Schools 2003, ed. by O. Bousquet, U. von Luxburg, G. Rätsch, Lecture Notes in Computer Science, vol. 3176 (Springer, Berlin, 2003), pp. 63–71
11. A. McDannald, M. Frontzek, A.T. Savici, M. Doucet, E.E. Rodriguez, K. Meuse, J. Opsahl-Ong, D. Samarow, I. Takeuchi, A.G. Kusne, W. Ratcliﬁ, On-the-fly autonomous control of neutron diffracton via physics-informed Bayesian active learning, preprint, arXiv:2108.08918 [cond-mat] (2021)