AutoSciLab: A Self-Driving Laboratory for Interpretable Scientific Discovery

Saaketh Desai[1], Sadhvikas Addamane[1], Jeffrey Y. Tsao[2], Igal Brener[1], Laura P. Swiler[3], Remi Dingreville[1], Prasad P. Iyer[1*]

[1]Center for Integrated Nanotechnologies, Sandia National Laboratories, Albuquerque, NM
[2]Material, Physical and Chemical Sciences Center, Sandia National Laboratories, Albuquerque, NM
[3]Center for Computing Research, Sandia National Laboratories, Center for Computing Research, Albuquerque, NM
*ppadma@sandia.gov

Abstract

Advances in robotic control and sensing have propelled the rise of automated scientific laboratories capable of high-throughput experiments. However, automated scientific laboratories are currently limited by human intuition in their ability to efficiently design and interpret experiments in high-dimensional spaces, throttling scientific discovery. We present AutoSciLab, a machine learning framework for driving autonomous scientific experiments, forming a surrogate researcher purposed for scientific discovery in high-dimensional spaces. AutoSciLab autonomously follows the scientific method in four steps: (i) generating high-dimensional experiments ($\mathbf{x} \in \mathbf{R^D}$) using a variational autoencoder (ii) selecting optimal experiments by forming hypotheses using active learning (iii) distilling the experimental results to discover relevant low-dimensional latent variables ($\mathbf{z} \in \mathbf{R^d}$, with d $\ll$ D) with a 'directional autoencoder' and (iv) learning a human interpretable equation connecting the discovered latent variables with a quantity of interest ($y = f(\mathbf{z})$), using a neural network equation learner. We validate the generalizability of AutoSciLab by rediscovering a) the principles of projectile motion and b) the phase-transitions within the spin-states of the Ising model (NP-hard problem). Applying our framework to an open-ended nanophotonics challenge, AutoSciLab uncovers a fundamentally novel method for directing incoherent light emission that surpasses the current state-of-the-art (Iyer et al. 2023b, 2020).

Introduction

Scientific discoveries over the past centuries have been fueled in large parts by human intuition and ingenuity (Koyré 2013). The traditional scientific discovery process involves hypothesis generation, design of experiments to test hypotheses, and distillation of data into interpretable forms, such as equations. Scientists build upon these learnt equations, often extrapolating with their intuition to guide the next iteration of scientific discovery. However, a critical bottleneck in this scientific discovery process is the reliance on human intuition to generate a valid hypotheses, design experiments to test those hypotheses and interpret results in a large, high-dimensional design space. Within the physical sciences, the intuition bottleneck manifests as a lack of theoretical frameworks, requiring expensive experiments to traverse a high-dimensional space - i.e.,'Blindly looking for a needle in a hay-stack with an expensive rake'.

Machine learning (ML) methods have excelled at aiding researchers in identifying patterns and correlations within dense datasets, (Krizhevsky, Sutskever, and Hinton 2012; Vaswani et al. 2017), even in the physical sciences, (Jumper et al. 2021; Choudhary et al. 2022), demonstrating the ability to execute a few of the necessary steps leading up to interpretable scientific discovery, e.g., active learning (AL) for efficient design of experiments (Ling et al. 2017; Kusne et al. 2020), or deep neural networks to extract high-dimensional correlations in conducted experiments (DeCost et al. 2019; Feng, Zhou, and Dong 2019). However, a complete machine learning framework which realizes interpretable scientific discovery that can augment the human intuition remains an open challenge. Key impediments to the development of such a framework include the ability to design a large search space of candidate experiments significantly beyond current scientific understanding, efficiently discovering optimal candidates in this search space, and realizing interpretable results.

Here, we present a solution titled AutoSciLab, which is a self-driving laboratory that employs machine learning methods to drive automated experiments. AutoSciLab accomplishes the following tasks:

(a) Drives autonomous scientific experiments by hypothesizing the next experiment from a generated space of experiments, without the aid of human intuition

(b) Minimizes the number of scientific experiments required to gain insight into the physical process while accounting for stochastic experiments

(c) Distills the experiments to discover an informative latent space complimenting any prior knowledge

(d) Learns an equation relating the discovered latent space with the physical property of interest

Related Work

Self-driving labs where automated experiments are driven by ML algorithms have recently gained popularity in the physical sciences (Abolhasani and Kumacheva 2023; Seifrid et al. 2022; MacLeod et al. 2020). While significant effort has focused on automating the 'mechanical' tasks of a physical science experiment (robotic sample handling, automated

end-to-end multi-step experimental workflows) (Abolhasani and Kumacheva 2023), the 'cognitive' aspects (data analysis, designing optimal experiments) have largely explored optimization. Self-driving labs have so far not been used for interpretable scientific discovery, especially for understanding scientific principles in high-dimensional search spaces. We now breakdown each component of AutoSciLab, and compare it to prior work:

Generating novel experiments: AutoSciLab employs generative models to generate novel experiments beyond human intuition. Generative models such as Variational Autoencoders (VAEs) (Kingma and Welling 2013) and Generative Adversarial Networks (Goodfellow et al. 2020) have shown excellent promise in generating structures in the physical sciences (Kim et al. 2020; Anstine and Isayev 2023). However, these models have yet to gain popularity within self-driving labs to generate novel experiments.

Efficient design of experiments: Efficient design of experiments via Bayesian optimization or active learning, is a staple of current self-driving labs (Ling et al. 2017; Kusne et al. 2020). Other techniques such as differential evolution (Storn and Price 1997) are also widely used for design of experiments (Zhong et al. 2015; Kim et al. 2021). However, these tools are often used with the intention of optimization and discovery, and not with the goal of interpretable scientific understanding.

Interpreting experiments using machine learning: Explainable and interpretable machine learning techniques, such as physics-informed neural networks (Raissi, Perdikaris, and Karniadakis 2019), explainability metrics (Lundberg and Lee 2017), and equation learner models (Desai and Strachan 2021; Sahoo, Lampert, and Martius 2018; Schmidt and Lipson 2009) have shown great promise in the physical sciences. However, these models have not been used in conjunction with self-driving labs and autonomous experiments.

The AutoSciLab Framework

AutoSciLab drives autonomous scientific experiments to discover an interpretable relationship between a physical quantity of interest $y \in R^m$ and high-dimensional *inputs* $\mathbf{x} \in \mathbf{R^D}$, see Fig. 1. AutoSciLab achieves this by recasting this problem as a discovery of an interpretable relationship between $y \in R^m$ and low dimensional latent space variables $\mathbf{z} \in \mathbf{R^d}$, with $d \ll D$. AutoSciLab then aims to learn $y = f(\mathbf{z})$ in a symbolic (equation) form. This recasting leverages the manifold hypothesis (Narayanan and Mitter 2010; Fefferman, Mitter, and Narayanan 2016), which states that high-dimensional data often resides in a low-dimensional manifold. The manifold hypotheses is often true in the physical sciences, where high-dimensional inputs often have specific, physics-driven mechanisms to affect a low-dimensional set of variables, which end up governing the quantity of interest y. To discover this relationship, AutoSciLab begins by reducing the high-dimensional *inputs* into a low-dimensional latent dimension $\mathbf{z}' \in \mathbf{R^d}$ using a latent-space based generative model, such that sampling the latent dimension $\mathbf{z}' \in \mathbf{R^d}$ generates novel experiments beyond the human intuition bottleneck that constraints experiment (hypothesis) generation. AutoSciLab then employs active learning to efficiently design experiments by creating hypotheses in the latent space $\mathbf{z}' \in \mathbf{R^d}$, designing experiments that optimally find the relationship between $\mathbf{z}'$ and y, i.e., finding $y = f(\mathbf{z}')$ by (for instance) maximizing $f(\mathbf{z}')$. Before this relationship can be distilled into an interpretable equation for scientific understanding, we need a mechanism to incorporate prior physical intuition into our relationship. AutoSciLab achieves this by transforming the low-dimensional latent variables $\mathbf{z}'$ describing the inputs $\mathbf{x}$ into the set of variables $\mathbf{z} \in \mathbf{R^d}$ controlling the physical process f. This transformation is achieved with a 'directional' autoencoder which is a conventional autoencoder with an additional regularization term that introduce correlations between $\mathbf{z} \in \mathbf{R^d}$ and a set of physics-informed variables known to affect f. Finally, AutoSciLab learns the $y = f(\mathbf{z})$ in symbolic (equation form) using a neural network based equation learner that we developed. We now describe each aspect in detail.

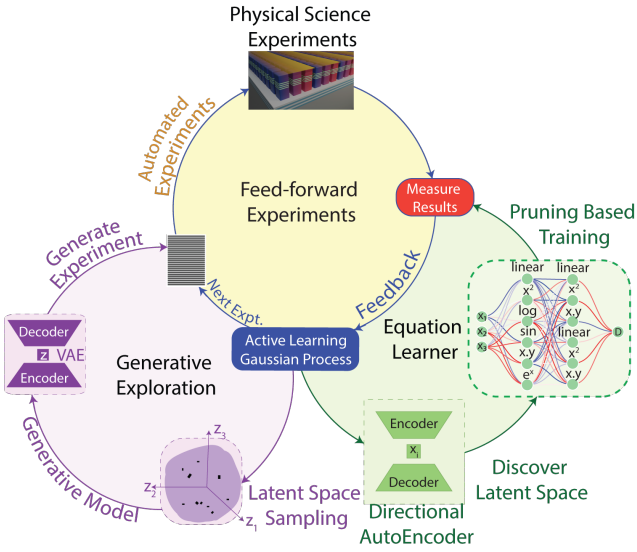

Figure 1: **AutoSciLab**. Automated 'experiments' are driven by an AL agent sampling the latent space of a generative model (variational autoencoder, VAE) (yellow/purple bubble). 'Experiment' here can refer to a physical laboratory measurement, or a model/simulation of a process. The set of experiments run by the AL agent are distilled using a directional autoencoder to discover a relevant latent space of interest (green bubble). The symbolic relationship between the relevant latent space variables and the physical property of interest is learnt using a neural network equation learner that uses pruning based on connection strength.

Generating Novel Experimental Candidates

We use a VAE to generate high-dimensional experiments $\mathbf{x} \in \mathbf{R^D}$. Given a training set $\mathbf{X}$ of candidate experi-

ments $\{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}\}$, with $\mathbf{x_i} \in \mathbf{R^D}$, we train a VAE to maximize $p(\mathbf{X})$, the likelihood of generating experiments similar to, but also beyond those in the training set. The VAE consists of two pieces: an encoder that reduces the experiments into a latent dimension $\mathbf{z'} \in \mathbf{R^d}$, learning the probability distribution $Q(\mathbf{z'}|\mathbf{x})$, and a decoder that reconstructs the experiment, learning the probability distribution $P(\mathbf{x}|\mathbf{z'})$. Maximizing $p(\mathbf{X})$ is equivalent to minimizing the Evidence Lower Bound (ELBO) loss, $\mathcal{L}_{\text{vae}}$ (Kingma and Welling 2013):

$$\mathcal{L}_{\text{vae}} = \mathrm{E}_{\mathbf{z'}\sim\mathrm{Q}}[\log \mathrm{P}(\mathbf{x}|\mathbf{z'}) - \mathrm{D}_{\mathrm{KL}}[\mathrm{Q}(\mathbf{z'}|\mathbf{x})||\mathrm{P}(\mathbf{z'})]] \quad (1)$$

where the first term is a reconstruction error, and the second term is a KL divergence that measures the difference between the learned latent space distribution $Q(\mathbf{z'}|\mathbf{x})$, and a prior latent space distribution, assumed here to be Gaussian with zero mean unit covariance $P(\mathbf{z'}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Sampling the learnt latent space of the trained VAE allows us to generate experiments beyond the training set $\mathbf{X}$. See **Appendix Section S2** (Desai et al. 2024) for details on VAE architectures and training sets.

Efficiently Identifying Promising Experiments

While the VAE excels at generating experiments beyond those studied before, the generation process is not directly tied to the physical quantity of interest (y). To find experimental features most relevant to the underlying scientific phenomenon, we employ active learning, finding 'optimal' experiments in the latent in the latent space $\mathbf{z'} \in \mathbf{R^d}$ spanning features of experimental inputs. The active learning task can be formulated as: $\max_{\mathbf{z'}\in\mathbf{Z'}} f(\mathbf{z'})$. We begin with an initial, small database of experiments $\mathrm{Y_{init}}$, represented by latent space embeddings $\mathbf{z'_{init}}$, and their associated physical quantity of interest y. Using this initial dataset $\mathrm{Y_{init}} = \{(\mathbf{z'_1}, \mathrm{y_1}), (\mathbf{z'_2}, \mathrm{y_2}), ..., (\mathbf{z'_n}, \mathrm{y_n})\}$, We can now train a Gaussian process model to learn $\mathrm{y}(\mathbf{z'}) = \mathcal{GP}(\mu(\mathbf{z'}), \mathrm{K}(\mathbf{z'}, \mathbf{z''}))$, where $\mu(\mathbf{z'})$ is a mean function and $\mathrm{K}(\mathbf{z'}, \mathbf{z''}))$ is a kernel representing the covariance in y between two experiments embedded as $\mathbf{z'}$ and $\mathbf{z''}$. The posterior distribution across every point in the latent space $\hat{\mathbf{z}}$ is now $y(\hat{\mathbf{z}}) = \mathcal{N}(\hat{\mu}(\hat{\mathbf{z}}), \hat{\sigma}(\hat{\mathbf{z}}))$, where (Rasmussen, Williams et al. 2006):

$$\hat{\mu}(\hat{\mathbf{z}}) = \mu(\hat{\mathbf{z}}) + \mathrm{K}(\hat{\mathbf{z}}, \mathbf{z'})\mathrm{K}(\mathbf{z'}, \mathbf{z'})^{-1}(\mathbf{y}(\mathbf{z'}) - \mu(\mathbf{z'})) \quad (2)$$

$$\hat{\sigma}(\hat{\mathbf{z}}) = \mathrm{K}(\hat{\mathbf{z}}, \hat{\mathbf{z}}) - \mathrm{K}(\hat{\mathbf{z}}, \mathbf{z'})\mathrm{K}(\mathbf{z'}, \mathbf{z'})^{-1}\mathrm{K}(\mathbf{z'}, \hat{\mathbf{z}}) \quad (3)$$

Given a posterior distribution, an acquisition function $\mathrm{I}(\mathbf{z'})$ can now hypothesize the next best experiment to conduct. Each experiment conducted by the active learning agent is a hypothesis on features of experiments that would result in optimal properties of interest, based on prior experiments. Since AutoSciLab focuses on scientific discovery, we are interested in not only identifying promising experiments, but also learning about the underlying physical phenomenon. Thus an acquisition function that balances exploration and exploitation, such as Expected Improvement (EI), should enable us to define the optimal relationship $y = f(\mathbf{z'})$. To establish a baseline, we compare EI to an upper confidence bound (UCB) acquisition function, where the hyperparameter λ of the acquisition function is tailored towards exploitation. The two acquisition functions can be written as:

$$\mathrm{I_{UCB}}(\mathbf{z'}) = \mu(\mathbf{z'}) + \lambda\sigma(\mathbf{z'}) \quad (4)$$

$$\mathrm{I_{EI}}(\mathbf{z'}) = \mathrm{E}_{\mathbf{z'}}(\max\{0, \mathrm{I}(\mathbf{z'}) - \mathrm{I}(\mathbf{z}^*)\}) \quad (5)$$

where $\mathbf{z}^*$ represents the current best value for the AF. The next best experiment, as identified by the acquisition function, can then be performed physically (in a laboratory, or using a model/simulation). This new data point $(\mathbf{z'_{n+1}}, \mathbf{y_{n+1}})$ can now be added to the dataset $\mathrm{y_{init}}$ and the process continues until an optimum is reached, or a set experimental budget is exhausted. **See Appendix Section S3** (Desai et al. 2024) for further details on active learning, and a differential evolution algorithm used as baseline for comparison.

Incorporating Prior Knowledge To Discover a Relevant Low-Dimensional Latent Space

Of the search space defined by the VAE latent space, only a fraction of points, such as those sampled by the active learning agent, contribute meaningfully to our understanding of the relationship between features of experimental inputs, and their observed physical properties. To learn this new latent space $\mathbf{z} \subset \mathbf{z'}$, we distill the results of experiments explored by active learning in the previous step. The set of high-dimensional experiments explored by the active learning are obtained by using the VAE's decoder $P(\mathbf{x}|\mathbf{z'})$ on the set of latent space points explored by the active learning. These high-dimensional experiments $\mathbf{x}$ are now encoded into a new latent space $\mathbf{z} \in \mathbf{R^d}$, recognizing that for highest interpretability, we would like to encapsulate in the latent space any prior knowledge of correlations between experimental inputs ($\mathbf{x}$) and quantities of interest (y). We learn the new latent space by training a 'directional' autoencoder, (Pati and Lerch 2019) (dAE), where some directions in latent space are explicitly designed to correlate with features of experiments known to physically affect the measured property. This correlation is enforced as a regularization in the distance between latent space variable (z_i), and a physics-based attribute (a):

$$\mathcal{L}_{\text{dAE}} = ||\mathbf{x}_{\text{recon}} - \mathbf{x}_{\text{gt}}||_2^2 + \mathcal{L}_{\text{dist}}(D_z, D_a) \quad (6)$$

$$\mathcal{L}_{dist}(D_z, D_a) = ||\tanh(D_z) - \mathrm{sgn}(D_a)||_2^2 \quad (7)$$

where the loss function of the directional autoencoder is a conventional reconstruction loss, and an additional regularization term L_{dist}, defined from prior work (Pati and Lerch 2019), with $D_z(i,j) = \mathbf{z_i} - \mathbf{z_j}$ and $D_a(i,j) = a_i - a_j$. **See Appendix Section S4** (Desai et al. 2024) for more details. Correlations between the latent space learnt by the 'directional' autoencoder, and physical quantities deemed relevant by subject matter experts provide us insight into the learnt latent space variables.

Learning Correlations as Equations

Given a subset of latent space variables $\mathbf{z}$ relevant to predicting the physical quantity of interest y, we now develop a neural network based equation learner (nn-EQL) that learns the correlation between $\mathbf{z}$ and y, using a dataset $\mathrm{Y} = \{(\mathbf{z_1}, \mathrm{y_1}), (\mathbf{z_2}, \mathrm{y_2}), ..., (\mathbf{z_n}, \mathrm{y_n})\}$. Our equation learner

is a customized neural network where each neuron in a layer has a specific activation function inspired by functional forms seen in the physical sciences (e.g., sin, x^2, $x_i x_j$) (Desai and Strachan 2021; Sahoo, Lampert, and Martius 2018). The equation learning process consists of three stages: (i) Perform a preliminary fit to the dataset Y to determine a baseline acceptable error, (ii) Systematically remove a fraction p of neurons in each layer based on their contribution to neuron activations in the next layer, followed by re-training until the baseline acceptable error, or similar, is reached, and (iii) Equation readout of the pruned network, where the connectivity in the final sparse network with limited weights and activation functions is expressed as an equation. Note that our pruning scheme is based on connection strength, while conventional pruning schemes using weight-based pruning (LeCun, Denker, and Solla 1989; Han et al. 2015; Sahoo, Lampert, and Martius 2018). This is due to our use of complex, non-monotonic, activation functions, where small weight values can be transformed into large activation values (e.g. $cos(x)$). Thus, instead of pruning the smallest weights in each layer, we prune the 'smallest contributing' connection in each layer, defining a connection ij's contribution as $W_{ij}x_i$, where W_{ij} is the weight connecting neuron i to a neuron j in the next layer, and x_i is neuron i's value. **See Appendix Section S5** (Desai et al. 2024) for more training details.

Benchmarking Components in AutoSciLab

Each component of AutoSciLab is benchmarked on simple tasks to demonstrate their applicability to scientific domains. The design of experiments using active learning is benchmarked on a simple one-dimensional search in one of our exemplars, showing that we can rediscover a known result (see **Appendix Section S3** (Desai et al. 2024)). The neural network equation learner is benchmarked on equations found in symbolic regression benchmarks (La Cava et al. 2021; Udrescu and Tegmark 2020).

Data set	True/Discovered equation
Lotka-Volterra interspecies dynamics	$\dot{x} = 3x - 2xy - x^2$ $\dot{x} = 3.063x - 2.02xy - x^2 - 0.16$
Van der Pol Oscillator	$\dot{x} = 10(y - \frac{1}{3}(x^3 - x))$ $\dot{x} = 10y - 3.153x^3 + 3.249x$
Magnetic moment of an electron in an orbit	$\mu = qvr$ $\mu = qvr$

Table 1: Benchmarking the **nn-EQL**

We find that our model accurately discovers each of these equations (Table 1), showing our ability to discover terms in equations commonly seen in physical systems. The generative model (VAE) uses conventional architectures, and we thus choose to quantify VAE generative capabilities individually for each application. As demonstrated in each exemplar, the VAE has clear capabilities to reconstruct experiments from the training set, as well as go beyond the training set and suggest novel experiments.

Exemplars, Experiments, and Results

We demonstrate the capabilities of AutoSciLab on three exemplar problems with varying difficulties (a) Projectile motion (b) Two-dimensional Ising spin system (c) An open ended nanophotonics problem. The first two problems have known solutions, validating the AutoSciLab framework while demonstrating the generalizability and range of problems that can be handled by the self-driving lab framework. Code for the experiments and results is available in the Appendix.

Exemplar 1: Projectile Motion

Problem description: Projectile motion describes the motion of a particle launched with an initial velocity at a specific angle, following Newtonian mechanics. The problem can be described as the evolution of the height y of the particle, as a function of time t, given an initial velocity u, and angle θ. That is, $y = f(t; u, \theta)$. Each such projectile is defined by a maximum height $H(u) = \frac{u_y^2}{2g}$, where $g = 9.8\frac{m}{s^2}$, and u_y denotes the y-component of the velocity.

Objective: Demonstrate that AutoSciLab can re-discover the relationship between the maximum height (H) of a projectile on the initial velocity (u_y).

AutoSciLab framing of task: In the AutoSciLab framework, we use the VAE to generate a variety of trajectories $y = f(t)$. We develop a training set consisting of quadratic, cubic, and quintic trajectories, taking the general form $y = \alpha t^2 + \beta t^3 + \gamma t^5$. The trained VAE thus generates a wide variety of realistic and unrealistic trajectories $y(t)$, from which the active learning agent now has to select realistic projectile trajectories. Here we utilize the fact that realistic trajectories must have a constant acceleration g, defining the active learning objective to be to minimize the difference between the second derivative of a candidate trajectory $y(t)$ and g. From the set of trajectories selected by the active learning, we can now use AutoSciLab's directional autoencoder learn a new latent space which correlates strongly with the initial velocity u_y, incorporating prior physical knowledge that H depends on u_y. Finally, we can use AutoSciLab's equation learner network to learn $H(u_y)$ in the form of a human-readable equation.

Result: We demonstrate that AutoSciLab accurately rediscovers the equation describing the maximum height attained by the projectile as a function of initial velocity, see Fig. 2. The VAE used in the AutoSciLab framework is able to generate trajectories that resemble projectile motion and pseudo-projectiles (where the acceleration is not constant), see Fig. 2(a). Using this trained VAE, active learning successfully identifies regions in latent space that represent trajectories with constant acceleration $\sim 10\ \frac{m}{s^2}$, i.e., realistic trajectories, see Fig. 2(b), with colored dots showing the region in latent space identified by active learning. AutoSciLab's directional autoencoder can now distill the trajectories explored by the active learning into a new latent space with one variable that correlates strongly with the initial velocity u_y, see Fig. 2(c), capturing prior knowledge that H depends on u_y. Finally, AutoSciLab's **nn-EQL** learns an interpretable equation relating height of the projectile (H) and latent vari-

able z. The learnt relationship $H(z)$ can be accurately translated to the known relationship $H(u_y) = \frac{u_y^2}{2g}$.

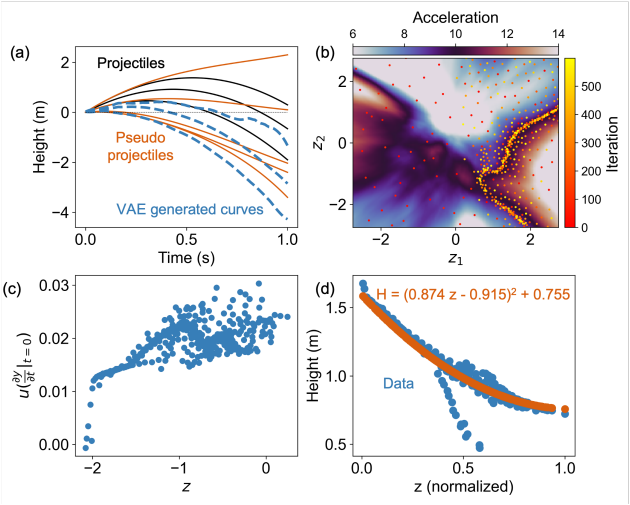

Figure 2: **Rediscovering projectile motion**. (a) Projectile height y as a function of time t. (b) Active learning efficiently finds points with acceleration $\sim$ 10 m/s^2. (c) Correlation between the y-component of the initial velocity u_y and latent space variable z learnt by the directional autoencoder, for trajectories identified by the active learning to have a constant acceleration g. (d) Maximum height (H) vs z, learnt equation overlaid.

Exemplar 2: Spin Dynamics of the Two-Dimensional Ising System

Problem description: As a complex benchmarking problem for AutoSciLab, we choose to re-discover the spin-dynamics of the Ising system as function of temperature. The Ising system consists of discrete variables called spins, which can be in one of two states (+1 or -1) arranged on a two dimensional lattice where each spin interacts with its neighbors. The Ising system is often used as an exemplar for a system that can be defined using simple physical principles, but exhibits complex emergent properties such as an effective magnetization, and a phase transitions (Onsager 1944). Determining the ground state of the Ising model, particularly when considering additional complexities such as external magnetic fields or disordered interactions (as in spin glass models), is known to be an NP-hard problem (Barahona 1982). By addressing this computationally challenging problem, we demonstrate the effectiveness of the AutoSciLab framework in tackling high-dimensional NP-hard scientific discovery tasks.

Objective: Our objective is to re-discover the relationship between the equilibrium average spin state (magnetization M) and the temperature T, known to be $M(\beta) = (1 - sinh(2\beta J)^{-4}))^{1/8}$ (Onsager 1944), where $\beta = \frac{1}{kT}$ is the inverse temperature, k is the Boltzmann constant , and J is the strength of interactions between spins. We assume k and J to be equal to 1 without loss of generality.

AutoSciLab framing of task: The equilibrium properties of the Ising system are obtained in practice using time-consuming Monte Carlo simulations that generate samples from a known equilibrium distribution of spins states $p(E) \propto \exp(\frac{-E}{kT})$, where E is the energy associated with a spin state. We generate arbitrary spin states by initializing a random set of binary spins, defining an average magnetization M_{init}, at a defined temperature T (equivalently β). Starting from sub-optimal random spin states with M_{init} significantly different from M will require long Monte Carlo simulations to reach an equilibrium set of states. We thus utilize the AL component of AutoSciLab to discover optimal spin states M_{init} that minimize the time to reach equilibrium t_{sol} when employing Monte Carlo simulations. Once optimal spin states are learnt, we can run a short Monte Carlo simulation to sample from the equilibrium distribution $p(E)$, quickly computing equilibrium magnetization M. The neural network equation learner component of AutoSciLab can then learn the relationship between M and T.

Results: We find that AutoSciLab re-discovers the relationship between equilibrium magnetization M and temperature T, see Fig. 3(a). We first find that AL efficiently finds optimal initial spin states M_{init} such that the time to reach equilibrium using standard Monte Carlo approaches (t_{sol}) is minimized, see Fig. 3(b). Note that by defining spin states in terms of magnetization M_{init}, we remove the need for a generative model to generate random spin states (simple random number generators can generate spin states), as well as the need for a directional autoencoder that encodes the known physical insight that the most relevant quantity that describes spin states and their properties is the average spin state, i.e., the magnetization M. We thus skip directly to the equation learner, demonstrating that we can learn the relationship between M and T, see Fig. 3(c). To achieve this, we run short Monte Carlo simulations starting from the optimal spin state learnt by the AL agent (M_{init}). We find that we can achieve equilibrium states using an order-of-magnitude fewer number of Monte Carlo steps compared to an arbitrarily random initialization (i.e., $t_{sol}^{AL} \ll t_{sol}^{random}$). The magnetization M achieved after the short Monte Carlo simulation, and the temperature T are collected in a dataset $Y = (M_1, T_1), (M_2, T_2)...(M_n, T_n)$. For ease of training the nn-EQL, we transform the dataset to be $Y' = (y_1, \beta_1), (y_2, \beta_2)...(y_n, \beta_n)$, where $\beta = \frac{1}{kT}$ and $y_i = (\frac{1}{1-M^8})^{1/4}$. The equation to be learnt is thus $y = sinh(2\beta)$, and we find that our neural network equation learner finds $y = 1.01sinh(1.96\beta) + 0.1\beta - 0.1$, which is close to the ground truth.

Exemplar 3: Open-Ended Nanophotonics Problem

Problem description: As an open-ended, real-world application, we employ AutoSciLab to realize high-efficiency steering of incoherent emission. The light emission from thermal lamps (incandescent bulbs, light emitting diodes (LEDs), black-body radiation, etc. are characterized as incoherent since their wavefront of emission remains random in both space and time (Pichler, Daley, and Zoller 2010). This random wavefront prevents traditional phased-array optical

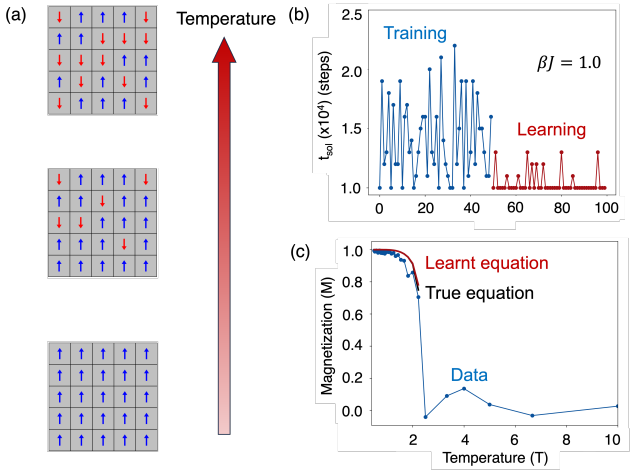

Figure 3: **Rediscovering the spin-dynamics of the Ising spin system.** a) Spin-state (s) represented on a grid showing the effects of increasing the temperature. b) Active learning at a fixed temperature ($T \propto (\beta J)^{-1}$) c) M vs T, showing overlap between the true equation (black) and the learnt equation (red) from the active learning results (blue).

elements (McManamon et al. 1996), used conventionally for lasers (and other coherent sources), to be used to redirect incoherent light. Dynamic semiconductor (GaAs) metasurfaces made up of a sub-wavelength array of tunable optical resonators have provided a route to steer incoherent light by embedding the light emitters within the resonators based on the pump pattern projected onto the metasurface. Previous results have demonstrated that the spatial (x) intensity gradients ($b \sim \frac{\partial \mathbf{x}}{\partial x}$) of the pump enable us to steer the light emission from the metasurface through momentum matching principles ($k_m = \frac{\partial \phi}{\partial x} \propto \frac{\partial \mathbf{x}}{\partial x}$) (Iyer et al. 2023b). However designing the optical pump pattern to re-direct the light into a desired direction, i.e., the inverse problem, remains a challenge. In other words, an open challenge is to maximize the directivity of emission $\mathrm{D_e} = \frac{f(\theta_i;\mathbf{x})}{\Sigma_j f(\theta_j;\mathbf{x})}$, where $f(\theta_i;\mathbf{x})$ is the emission towards an angle θ_i, given pump pattern $\mathbf{x}$.

Objective: Our objective is to discover the dependence of directivity $\mathrm{D_e}$ on features of pump patterns imposed on metasurfaces to efficiently steer incoherent emission.

AutoSciLab framing of task: In the AutoSciLab framework, the VAE generates arbitrary pump patterns $\mathbf{x}$ given a training set $\mathbf{X}$ of pump patterns $\{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}\}$, with $\mathbf{x_i} \in \mathbf{R^D}$. The active learning agent searches the latent space of the VAE to maximize directivity, formulated as: $\max_{\mathbf{z'} \in \mathbf{Z'}} f(\mathbf{z'})$. **Appendix Section S3** (Desai et al. 2024) shows proof of concept results for AL on noisy experimental data. We inspect the pump patterns generated by the VAE and the active learning, which have high directivity, and encode them into a new latent space $\mathbf{z'} \in \mathbf{R^d}$, using the directional autoencoder piece of AutoSciLab. Specifically, we encode prior physical knowledge: the local slope of the pump pattern $b = \frac{\partial \mathbf{x}}{\partial x}$, and the local curvature of the pump pattern $a = \frac{\partial^2 \mathbf{x}}{\partial x^2}$ that affect directivity. Finally, we learn an interpretable relationship between features of pump patterns $\mathbf{z}$ and their associated directivity $\mathrm{D_e}$, using AutoSciLab's equation learner.

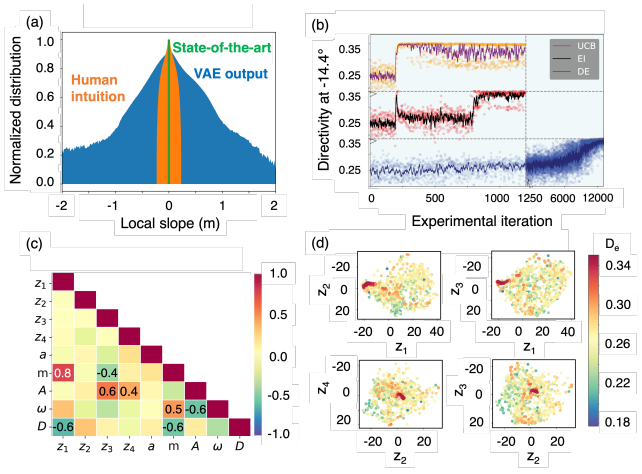

Figure 4: **Discovering novel relationships in the nanophotonics domain**. (a) Generative capacity of the VAE, quantified as the normalized distribution (log scale) of the local slope (b) in pump patterns. (b) Directivity as a function of experimental iteration. Dots represent each experiment, and curves reperesent moving averages. (c) Spearman correlations between the discovered latent space and physically relevant pump pattern characteristics. (d) Correlating the latent space with Directivity, incorporating prior knowledge.

Results: We demonstrate that AutoSciLab can successfully tackle the open-ended nanophotonics problem of steering incoherent emission resulting in a discovery of a novel principle relating pump pattern features to directivity in an equation, see Fig. 4. Fig. 4(a) quantifies the ability of our VAE to generate pump patterns beyond those studied so far, an aspect critical to defining the search space over which we understand the phenomenon of incoherent emission steering. Fig. 4(a), demonstrates the generative capability of the VAE, where the variety (quantified as the local slope distribution) of the pump patterns is larger than the state-of-the-art and human intuition based training set of the VAE by at least two orders of magnitude (Iyer et al. 2023a). A wide variation in local slopes indicates numerous changes in the pump pattern, going significantly beyond previous the state of the art, and proposing pump patterns that could potentially yield a directivity never seen before. See **Appendix Section S2** (Desai et al. 2024) for examples of pump patterns generated by the VAE, and their translation to patterns imposed on the physical metasurface. We then find that AL, searching over the latent space of the VAE, discovers pump patterns with directivity values that are 3-4x higher than in the training set, across multiple emission angles, see Fig. 4(b). Note that ground-truth data here is obtained by evaluating a neural network surrogate model, but the AL scheme is also able to identifying promising pump patterns when

obtaining ground-truth data directly from the noisy experiment, see **Appendix Section S3** (Desai et al. 2024). Here we focus on the noiseless surrogate model as it allows us to compare multiple acquisition functions, as well as multiple downstream algorithms in a principled manner, focusing on the generalizability of our framework rather than focusing on physics of the experimental noise. We demonstrate that AL (with EI and UCB acquisition functions) reduces the required number of experiments by an order of magnitude (see 4(b)) to discover the pump patterns with high directivity when compared to differential evolution (Storn and Price 1997). Fig. 4(b) demonstrates that AL has the capability to efficiently hypothesize the next experiment - with only a limited training set - maximizing the physical property with minimal experiments in a self-driving lab framework. See **Appendix Section S3** for more AL results (Desai et al. 2024).

We analyze the experiments selected by the EI AF to guide our understanding of the physical process to learn interpretable relationships between features in the pump pattern (input) and directivity of emission (output). To learn these interpretable relationships, we convert the set of latent space points $\mathbf{z}$ explored by the AL into a 'relevant' set of pump patterns $\mathbf{x}'$. These high dimensional pump patterns are condensed into a new, low-dimensional latent space $\mathbf{z}$ such that dimension z_1 correlates with the local slope b, see Fig. 4(c). **Appendix Section S4** (Desai et al. 2024) documents repeatability and generality of AutoSciLab for other emission angles.

We formalize the insights described so far in Fig. 4 as an equation governing incoherent light emission from a metasurface. Using our nn-EQL, we learn an interpretable equation between latent space variables $\mathbf{z}$ and directivity $\mathrm{D_e}$.

$$\mathrm{D_e} = 0.0467z_1^2 - 0.0265z_2^2 - 0.175z_1 - 0.0955z_1z_2 + 0.22z_2 + 2.707 \quad (8)$$

where z_1 and z_2 are the first two variables in the learned 'relevant' latent space $\mathbf{z}$. (See Appendix for additional analysis describing the physical implications of the equation to the process of emission). This structure-property relationship discovered here goes beyond our current understanding of steering incoherent light, based on Fourier transforms.

Impact of AutoSciLab

The goal of AutoSciLab is to achieve interpretable scientific discovery in high-dimensional spaces. Ideally, AutoSciLab would (a) reduce the number of experiments needed for a scientific discovery (b) Propose a wider, more varied set of experiments than human intuition (c) Select and conduct experiments with valuable information needed for scientific discovery. Table 2 documents estimates of each aspect mentioned above, for multiple exemplars. The metrics in Table 2 quantify the number of experiments, the 'variety' of experiments proposed (in terms of the bounds of the design space), and the 'value' of experiments as the average 'performance' (time to solution, directivity etc.) of the set of experiments conducted. We expect AutoSciLab to increase $\frac{\mathrm{N_{human}}}{\mathrm{N_{auto}}}$ (by lowering $\mathrm{N_{auto}}$), increase $\frac{\mathrm{Variety_{auto}}}{\mathrm{Variety_{human}}}$, and increase $\frac{\mathrm{Value_{auto}}}{\mathrm{Value_{human}}}$. The 'Gain Factor' is a product of these three terms, with higher values being better. For more details on these calculations, please refer to **Appendix Section S6** (Desai et al. 2024). We find the gain factor from using AutoSciLab to be significantly large in problems with large design spaces, or problems with unknown underlying physics that require a large number of experiments with limited intuition. For small design spaces, AutoSciLab performs similarly to human intuition, which is expected. Additionally, AutoSciLab, as a scientific discovery framework, attains 'level 4 autonomy' in the six levels of autonomy for scientific discovery described in (Kramer et al. 2023), positioning our work as a groundbreaking effort to address this currently unmet challenge. AutoSciLab also integrates multiple aspects of scientific discovery as defined in recent work (Langley 2024). Specifically, AutoSciLab supports 'Inducing Numeric Laws' via the nn-EQL, 'Experimentation and Observation' with the active learning and the generative model (VAE), as well as 'Measuring and Identifying Variables' with the directional autoencoder. AutoSciLab's applicability to material science, as demonstrated with the open-ended nanophotonics exemplar also addresses the need for autonomous scientific discovery in this field, as identified by the above work.

Problem	Design Space	Method	Experiments			Gain Factor
			Number	Variety	Value	
Projectile Motion	2	Human	10	0.65	1	0.97
		Auto	500	36	0.88	
Ising Model	1024	Human	10	1	1	100
		Auto	10	1	100	
Photonics	8×10^6	Human	1M	0.1	1	2×10^6
		Auto	1000	100	2	

Table 2: Quantifying the gains from using AutoSciLab

Limitations of AutoSciLab

While the AutoSciLab framework is developed to be an autonomous research agent performing experiments in the lab, this initial demonstration is informed and constrained by human researchers in a few ways. The ML framework expects automated experiments, which can limit the type of experiments that can be handled by this framework. Furthermore, the variety of experiments generated by the VAE depends on the training set, which is currently defined by human intuition. The directional autoencoder is designed to incorporate prior subject knowledge in the latent space, and currently assumes that this prior knowledge is accurate, which may not be true for cutting edge research. Lastly, the equation learner is initialized with a dictionary of activation functions provided by the researcher, again assuming this set of functions to be valid. The requirements and inclusion of prior knowledge into the AutoSciLab framework can bias the explorative directions of the experiments, while also acting as guardrails to limit experiments with bounded outcomes.

Conclusion

We demonstrate a self-driving lab framework capable of interpretable scientific discoveries over a wide range of scientific phenomena. We leverage an AL agent, searching over the low-dimensional latent space of a generative model (VAE), to efficiently discover optimal experiments, across multiple domains such as projectile motion, the Ising spin system, and an open ended nanophotonics problem. The sampled experiments were distilled using a directional autoencoder to discover the relevant latent variables as a subspace of the VAE latent space, while encoding human intuition based variables in the latent space. Correlations discovered between latent space variables and human-intuition based variables are formalized as human-interpretable physical relationships, using neural-network equation learners. AutoSciLab successfully re-discovers known physics in the projectile motion and the Ising spin system problem, and discovers novel steering principles in the nanophotonics problem, discovering a new method to steer incoherent light emission from metasurfaces. We also demonstrate that as the dimensionality of the design space increases, the gain in experimental cost enabled by AutoSciLab also scales. We envision this scientific discovery from AutoSciLab to form a novel pathway to realize clean energy sources (LEDs and thermal lamps), with the potential for use in applications such as AR/VR and holographic displays. From the perspective of scientific discovery, we have developed an ML framework to hypothesize and generate candidate experiments, efficiently search the experimental design space, discover the relevant latent variables while encoding prior knowledge, and finally, condense experimental knowledge in the form a human readable equation. These steps capture the scientific method, and we envision that AutoSciLab can be generalized to most physical science domains with expensive and noisy experiments, to learn physical relationships as interpretable equations.

Acknowledgments

The work was in part supported by Sandia National Laboratories' Laboratory Directed Research and Development program, and in part by the Center for Integrated Nanotechnologies, an Office of Science user facility operated for the U.S. Department of Energy. This article has been authored by an employee of National Technology & Engineering Solutions of Sandia, LLC under Contract No. DE-NA0003525 with the U.S. Department of Energy (DOE).

References

Abolhasani, M.; and Kumacheva, E. 2023. The rise of self-driving labs in chemical and materials sciences. *Nature Synthesis*, 2(6): 483–492.

Anstine, D. M.; and Isayev, O. 2023. Generative models as an emerging paradigm in the chemical sciences. *Journal of the American Chemical Society*, 145(16): 8736–8750.

Barahona, F. 1982. On the computational complexity of Ising spin glass models. *Journal of Physics A: Mathematical and General*, 15(10): 3241.

Choudhary, K.; DeCost, B.; Chen, C.; Jain, A.; Tavazza, F.; Cohn, R.; Park, C. W.; Choudhary, A.; Agrawal, A.; Billinge, S. J.; et al. 2022. Recent advances and applications of deep learning methods in materials science. *npj Computational Materials*, 8(1): 59.

DeCost, B. L.; Lei, B.; Francis, T.; and Holm, E. A. 2019. High throughput quantitative metallography for complex microstructures using deep learning: A case study in ultrahigh carbon steel. *Microscopy and Microanalysis*, 25(1): 21–29.

Desai, S.; Addamane, S.; Tsao, J. Y.; Brener, I.; Swiler, L. P.; Dingreville, R.; and Iyer, P. P. 2024. AutoSciLab: A Self-Driving Laboratory For Interpretable Scientific Discovery. *arXiv preprint arXiv:2412.12347*.

Desai, S.; and Strachan, A. 2021. Parsimonious neural networks learn interpretable physical laws. *Scientific reports*, 11(1): 12761.

Fefferman, C.; Mitter, S.; and Narayanan, H. 2016. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4): 983–1049.

Feng, S.; Zhou, H.; and Dong, H. 2019. Using deep neural network with small dataset to predict material defects. *Materials & Design*, 162: 300–310.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Han, S.; Pool, J.; Tran, J.; and Dally, W. 2015. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28.

Iyer, P. P.; DeCrescent, R. A.; Mohtashami, Y.; Lheureux, G.; Butakov, N. A.; Alhassan, A.; Weisbuch, C.; Nakamura, S.; DenBaars, S. P.; and Schuller, J. A. 2020. Unidirectional luminescence from InGaN/GaN quantum-well metasurfaces. *Nature Photonics*, 14(9): 543–548.

Iyer, P. P.; Desai, S.; Addamane, S.; Dingreville, R.; and Brener, I. 2023a. Learning incoherent light emission steering from metasurfaces using generative models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3770–3777.

Iyer, P. P.; Karl, N.; Addamane, S.; Gennaro, S. D.; Sinclair, M. B.; and Brener, I. 2023b. Sub-picosecond steering of ultrafast incoherent emission from semiconductor metasurfaces. *Nature Photonics*, 1–6.

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.

Kim, C.; Batra, R.; Chen, L.; Tran, H.; and Ramprasad, R. 2021. Polymer design using genetic algorithm and machine learning. *Computational Materials Science*, 186: 110067.

Kim, S.; Noh, J.; Gu, G. H.; Aspuru-Guzik, A.; and Jung, Y. 2020. Generative adversarial networks for crystal structure prediction. *ACS central science*, 6(8): 1412–1420.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Koyré, A. 2013. *The Astronomical Revolution: Copernicus-Kepler-Borelli*. Routledge.

Kramer, S.; Cerrato, M.; Džeroski, S.; and King, R. 2023. Automated scientific discovery: from equation discovery to autonomous discovery systems. *arXiv preprint arXiv:2305.02251*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Kusne, A. G.; Yu, H.; Wu, C.; Zhang, H.; Hattrick-Simpers, J.; DeCost, B.; Sarker, S.; Oses, C.; Toher, C.; Curtarolo, S.; et al. 2020. On-the-fly closed-loop materials discovery via Bayesian active learning. *Nature communications*, 11(1): 5966.

La Cava, W.; Orzechowski, P.; Burlacu, B.; de França, F. O.; Virgolin, M.; Jin, Y.; Kommenda, M.; and Moore, J. H. 2021. Contemporary symbolic regression methods and their relative performance. *arXiv preprint arXiv:2107.14351*.

Langley, P. 2024. Integrated Systems for Computational Scientific Discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 22598–22606.

LeCun, Y.; Denker, J.; and Solla, S. 1989. Optimal brain damage. *Advances in neural information processing systems*, 2.

Ling, J.; Hutchinson, M.; Antono, E.; Paradiso, S.; and Meredig, B. 2017. High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates. *Integrating Materials and Manufacturing Innovation*, 6: 207–217.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

MacLeod, B. P.; Parlane, F. G.; Morrissey, T. D.; Häse, F.; Roch, L. M.; Dettelbach, K. E.; Moreira, R.; Yunker, L. P.; Rooney, M. B.; Deeth, J. R.; et al. 2020. Self-driving laboratory for accelerated discovery of thin-film materials. *Science Advances*, 6(20).

McManamon, P. F.; Dorschner, T. A.; Corkum, D. L.; Friedman, L. J.; Hobbs, D. S.; Holz, M.; Liberman, S.; Nguyen, H. Q.; Resler, D. P.; Sharp, R. C.; et al. 1996. Optical phased array technology. *Proceedings of the IEEE*, 84(2): 268–298.

Narayanan, H.; and Mitter, S. 2010. Sample complexity of testing the manifold hypothesis. *Advances in neural information processing systems*, 23.

Onsager, L. 1944. Crystal statistics. I. A two-dimensional model with an order-disorder transition. *Physical Review*, 65(3-4): 117.

Pati, A.; and Lerch, A. 2019. Latent space regularization for explicit control of musical attributes. In *ICML Machine Learning for Music Discovery Workshop (ML4MD), Extended Abstract, Long Beach, CA, USA*.

Pichler, H.; Daley, A.; and Zoller, P. 2010. Nonequilibrium dynamics of bosonic atoms in optical lattices: Decoherence of many-body states due to spontaneous emission. *Physical Review A*, 82(6): 063605.

Raissi, M.; Perdikaris, P.; and Karniadakis, G. E. 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.

Rasmussen, C. E.; Williams, C. K.; et al. 2006. *Gaussian processes for machine learning*, volume 1. Springer.

Sahoo, S.; Lampert, C.; and Martius, G. 2018. Learning equations for extrapolation and control. In *International Conference on Machine Learning*, 4442–4450. PMLR.

Schmidt, M.; and Lipson, H. 2009. Distilling free-form natural laws from experimental data. *science*, 324(5923): 81–85.

Seifrid, M.; Pollice, R.; Aguilar-Granda, A.; Morgan Chan, Z.; Hotta, K.; Ser, C. T.; Vestfrid, J.; Wu, T. C.; and Aspuru-Guzik, A. 2022. Autonomous chemical experiments: Challenges and perspectives on establishing a self-driving lab. *Accounts of Chemical Research*, 55(17): 2454–2466.

Storn, R.; and Price, K. 1997. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11: 341–359.

Udrescu, S.-M.; and Tegmark, M. 2020. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16): eaay2631.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhong, Y. K.; Fu, S. M.; Ju, N. P.; Chen, P. Y.; and Lin, A. 2015. Experimentally-implemented genetic algorithm (ExpGA): toward fully optimal photovoltaics. *Optics Express*, 23(19): A1324–A1333.