

# Deep Learning and Symbolic Regression for Discovering Parametric Equations

Michael Zhang<sup>\*1</sup>, Samuel Kim<sup>\*1a</sup>, Peter Y. Lu<sup>2</sup>, Marin Soljačić<sup>2b</sup>

**Abstract**—Symbolic regression is a machine learning technique that can learn the governing formulas of data and thus has the potential to transform scientific discovery. However, symbolic regression is still limited in the complexity and dimensionality of the systems that it can analyze. Deep learning on the other hand has transformed machine learning in its ability to analyze extremely complex and high-dimensional datasets. We propose a neural network architecture to extend symbolic regression to parametric systems where some coefficient may vary but the structure of the underlying governing equation remains constant. We demonstrate our method on various analytic expressions, ODEs, and PDEs with varying coefficients and show that it extrapolates well outside of the training domain. The neural network-based architecture can also integrate with other deep learning architectures so that it can analyze high-dimensional data while being trained end-to-end. To this end we integrate our architecture with convolutional neural networks to analyze 1D images of varying spring systems.

**Index Terms**—Symbolic regression, deep learning, neural network, parametric, PDE, varying coefficient, high-dimensional

## I. INTRODUCTION

**D**ISCOVERING the governing equations of nature is key to many scientific disciplines. Many complex systems can be described by mathematical equations, ranging from Hooke’s law for harmonic oscillators to Maxwell’s equations for electrodynamics. While scientists have often spent years developing insights to discover these equations, machine learning has become alluring in its potential to tackle and automate extremely complex tasks. For example, deep learning in recent years has been able to create images from captions [1] and predict a protein’s 3D structure [2] better and faster than humans. However, deep learning models are often black-box, making it difficult to gain scientific insight from these techniques. Thus, in order to make deep learning widely applicable for scientific discovery, we covet methods that are interpretable so that scientists can extract meaningful information from complex datasets.

Symbolic regression is a machine learning technique that finds an analytical mathematical expression that describes the data, thus resulting in an interpretable model. Symbolic regression is often implemented through genetic programming, which searches through the space of mathematical expressions while ensuring that the equation is viable through various

heuristics [3]. The equations are pieced together through basic building blocks known as primitive functions, which include constants and simple functions (e.g. addition, multiplication, sine). Ref. [4], one of the most popular earlier works in this direction, demonstrated how symbolic regression could discover equations of motions including Hamiltonians and Lagrangians for various physical systems. However, these approaches do not scale well to high-dimensional problems and often require numerous hand-built heuristics and rules.

One type of complexity we explore in this work are datasets described by parametric equations in which the underlying equation may stay the same but coefficients may vary along one or more dimensions. For example, the diffusion constant may vary over time or space as the system governed by the diffusion equation evolves. Various approaches have been proposed to discover parametric PDEs, including genetic algorithms combined with averaging over local windows [5], linear regression with kernel smoothing over adjacent coefficients [6], and group sparsity on SINDy (Sparse Identification of Nonlinear Dynamical systems) [7].

There have also been numerous approaches at introducing the power of deep learning into symbolic regression to enable it for more complex tasks. For example, AI-Feynman checks for a number of physics-inspired invariances and symmetries using both hand-built rules and neural networks to simplify the data [8]. Neural network autoencoders have been combined with SINDy to enable equation discovery on high-dimensional dynamical systems [9]. PDE-Net 2.0 incorporates a symbolic network to discover PDEs using convolutional networks with constrained filters [10]. Ref. [11] incorporates a symbolic network with a neural network encoder to discover ODE and PDE systems from partial observations. Ref. [12] performs traditional symbolic regression on graph neural network weights in a 2-step process to discover the dynamics of many-body systems.

In particular, a neural network architecture was proposed that can perform symbolic regression by replacing the activation functions with primitive functions [13], [14]. Furthermore, ref. [15] showed how this architecture can be integrated into other deep learning architectures including convolutional networks and recurrent networks to perform symbolic regression on high-dimensional and dynamical systems, while allowing the entire architecture to be trained end-to-end through backpropagation. Ref. [16] further extended this for recursive programs, implicit functions, and image classification.

In this work we extend the approach from ref. [15] and enable neural network-based symbolic regression to parametric equations that may have varying coefficients. We propose two

<sup>\*</sup>These authors contributed equally to this work.

<sup>1</sup>Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA

<sup>2</sup>Department of Physics, Massachusetts Institute of Technology, Cambridge, MA, USA

<sup>a</sup>E-mail: samkim@mit.edu

<sup>b</sup>E-mail: soljacic@mit.edu

novel architectures, the stacked EQL network (SEQL) and the parametric EQL network (PEQL), which can each discover parametric equations with various advantages. We demonstrate our method on various analytic equations, PDEs, and a high-dimensional dataset consisting of images of particles.

## II. EQL NETWORK

The EQL network is a neural network architecture that can perform symbolic regression by replacing the nonlinear activation functions with primitive functions. It was initially proposed in [13], [14] and further expanded in [15]. In Section II-A we briefly review the base EQL architecture for symbolic regression, while more details can be found in ref. [15]. We also propose several modifications to the EQL network that improve its training behavior. In Sections II-D and II-E we propose 2 variants of the EQL architecture that can discover parametric equations. Note that in our discussion and notation in this section, we assume that the coefficients are parameterized with respect to *time* as this provides a convenient intuition applicable to many systems. However, the parameterization could also be with respect to other quantities (e.g. space).

### A. Base Architecture

The EQL network architecture closely mirrors a fully-connected neural network. The output of the  $i^{\text{th}}$  layer can be described by

$$\mathbf{g}^{(i)} = \mathbf{W}^{(i)} \mathbf{h}^{(i-1)} \quad (1)$$

$$\mathbf{h}^{(i)} = f(\mathbf{g}^{(i)}) \quad (2)$$

where  $\mathbf{W}$  is a weight matrix,  $f$  is the activation function, and  $\mathbf{h}_0 = \mathbf{x}$  is the input data. The activation function for the final layer is typically linear, so the output of the neural network with  $L$  hidden layers is  $y = \mathbf{W}^{(L+1)} \mathbf{h}^{(L)}$ .

While conventional neural networks typically use activation functions such as ReLU or sigmoid, the EQL network uses a set of primitive functions, where each component of  $g$  may go through a different primitive function and where a primitive function may take multiple inputs. The network is trained using the same techniques as conventional neural networks, i.e. stochastic gradient descent, and once it is trained, the discovered equation can simply be read off of the weights.

### B. Sparsity

To ensure the interpretability of symbolic regression, we need to force the system to learn the simplest expression that describes the data. In genetic programming-based approaches, this is typically done by limiting the number of terms in the expression. For the EQL network, we enforce this through the use of sparsity regularization on the network weights such that as many of the weights are set to 0 as possible. While [15] primarily uses a smoothed  $L_{0.5}$  regularization, in this work we use a relaxed form of  $L_0$  regularization [17]. We briefly review the details here, and refer the reader to refs. [15] and [17] for more details.

The weights of the neural network are reparameterized as

$$\mathbf{W} = \tilde{\mathbf{W}} \odot \mathbf{z}$$

where  $\mathbf{z}$  has the same dimensions as  $\mathbf{W}$  and can be interpreted as a gate variable, and the multiplication is component-wise. Ideally each element of  $\mathbf{z}$  is a binary “gate” such that  $z \in \{0, 1\}$ . However, this is not differentiable and so we allow  $z$  to be a stochastic variable drawn from the hard concrete distribution:

$$\begin{aligned} u &\sim \mathcal{U}(0, 1) \\ s &= \text{sigmoid}([\log u - \log(1 - u) + \log \alpha] / \beta) \\ \bar{s} &= s(\zeta - \gamma) + \gamma \\ z &= \min(1, \max(0, \bar{s})) \end{aligned}$$

where  $\alpha$  is a trainable variable that describes the location of the hard concrete distribution, and  $\beta, \zeta, \gamma$  are hyperparameters that describe the distribution. In the case of binary gates, the regularization penalty would simply be the sum of  $\mathbf{z}$  (i.e., the number of non-zero elements in  $\mathbf{W}$ ). However, in the case of the hard concrete distribution, we can calculate an analytical form for the expectation of the regularization penalty over the distribution parameters. The sparsity regularization loss is then

$$\mathcal{L}_R = \sum_j \text{sigmoid}\left(\log \alpha_j - \beta \log \frac{-\gamma}{\zeta}\right)$$

where  $j$  is indexing through all of the weight components. While ref. [17] applies group sparsity to the rows of the weight matrices with the goal of computational efficiency, we apply parameter sparsity (to individual elements) with the goal of simplifying the expression in symbolic regression.

The advantage of  $L_0$  regularization is that it enforces sparsity without placing a penalty on the magnitude of the weights by placing a penalty on the expected number of non-zero weights. Additionally, it lends itself to a straightforward definition of group sparsity across time-steps as we will see in Section II-D. In our experiments, we use the hyperparameters for the  $L_0$  regularization suggested by ref. [17].

### C. Skip Connections

In this work, we also add skip connections to the EQL network to introduce an inductive bias towards simpler equations while simultaneously enabling the discovery of more complex equations. The most well-known type of skip connections were introduced in ResNets, which take the output of a layer and add it to the layer ahead with the goal of allowing gradient information to efficiently propagate through many layers and enabling extremely deep architectures [18]. While these would be feasible to implement in the EQL network, they would serve to increase the complexity of the equation as information flows through the network. In contrast, we turn to the skip connections introduced by DenseNets which concatenates, rather than sums, the output of the previous layer with that of the next layer [19]. More specifically, we modify Equation 2 as:

$$\mathbf{h}^{(i)} = \left[ f(\mathbf{g}^{(i)}); \mathbf{h}^{(i-1)} \right] \quad (3)$$

Skip connections introduce a slight inductive bias towards learning simpler functions, since functions can route “directly” to the output without needing to go through the identity primitive function of successive layers. Additionally, skip

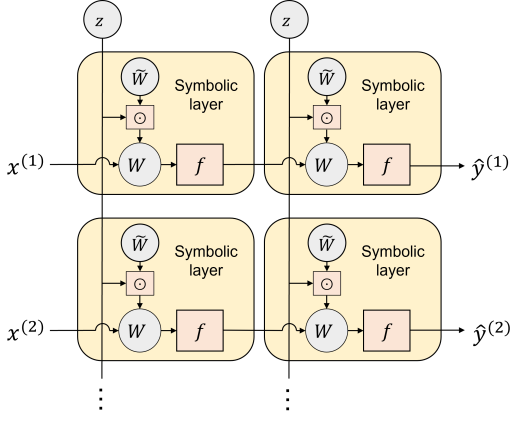


Fig. 1. Architecture of the stacked EQL network. Note that the indexing  $x^{(j)}$  is for the time step, rather than for the data in each time step. Each horizontal row represents an EQL network for each time step. The gate  $\mathbf{z}$  is shared across time steps.

connections minimize instabilities during training that can arise as a result of gradients exploding as they pass through the primitive functions. Thus, skip connections allow us to train EQL networks with more layers, which in turn can learn more complex equations.

#### D. Stacked Architecture (SEQL)

The first extension we propose to analyze parametric equations is to train a separate EQL network for each time step, an architecture that we call the stacked EQL (SEQL) network. Suppose we have a dataset

$$\mathcal{D} = \left\{ \left\{ x^{(i,j)}, y^{(i,j)} \right\}_{i=1}^{N_t} \right\}_{j=1}^{N_t} \quad (4)$$

where  $N_t$  is the number of time steps and  $N^{(j)}$  is the number of data points in that time step (note that  $N^{(j)}$  does not need to be constant across time steps). For layer  $i$  of the SEQL network, we can construct  $N_t$  separate weight matrices,  $\{\tilde{\mathbf{W}}^{(i,j)}\}_{j=1}^{N_t}$ .

If we naively train  $N_t$  separate EQL networks, then it is possible that each network may learn a different equation. Additionally, we would cut the volume data by a factor of  $N_t$  for each network, thus throwing out the remainder of the data. To counteract this, we enforce that the different networks learn the same equation by implementing group sparsity through weight sharing of the gate variable  $\mathbf{z}$ . For the  $i^{\text{th}}$  layer of the  $j^{\text{th}}$  time step, we modify Eq. 2 as:

$$\mathbf{h}^{(i,j)} = f \left( \left( \tilde{\mathbf{W}}^{(i,j)} \odot \mathbf{z}^{(i)} \right) \mathbf{h}^{(i-1,j)} \right) \quad (5)$$

For an architecture with  $L$  hidden layers, there are  $(L+1) \cdot N_t$  weight matrices and  $L+1$  gate matrices.

Another modification we make to the architecture is weight regularization across time steps to introduce an inductive bias towards smoothness in the coefficients. We use  $L_2$  regularization loss between adjacent time steps. Looking at just a single

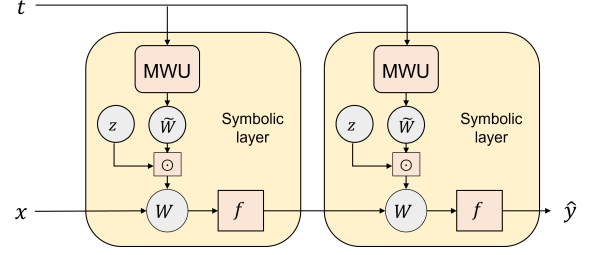


Fig. 2. Architecture of the parameterized EQL network. The linear output layer is not shown here for simplicity.

element  $w_{k,l}$  of  $\mathbf{W}$  in a single layer for notational simplicity, the  $L_2$  loss is simply

$$L_{S,k,l} = \sum_{j=1}^{N_t-1} \left( w_{k,l}^{(j+1)} - w_{k,l}^{(j)} \right)^2 \quad (6)$$

and the total regularization loss is

$$\mathcal{L}_S = \sum_{i,k,l} L_{S,k,l}^{(i)} \quad (7)$$

where  $i$  is indexing the layer. This regularization pushes coefficients in adjacent time-steps closer together and can more effectively counteract noisy datasets.

#### E. Parameterized Architecture (PEQL)

An alternative approach to learning parametric systems is to parameterize the weights so that they are a function of time,  $\mathbf{W}(t)$ . While a number of models can be used to parameterize the weights, we use a fully-connected neural network as it is an extremely flexible model that can handle functions with discontinuities and can be trained with backpropagation, allowing the entire system to be trained end-to-end. We call this fully-connected neural network the meta-weight unit (MWU). This parameterized EQL (PEQL) architecture is shown in Figure 2.

This idea is similar to that of hypernetworks, in which a network is used to generate the weights of another network [20]. The general idea using a network to parameterize or interact with the weights of another network has been most notably leveraged for meta-learning [21]–[24], and has also been applied to a variety of other architectures, including the Neural ODE [25] and HyperPINN [26].

The PEQL has a separate MWU in each layer (including the linear output layer) which takes time  $t$  as an input and outputs the weight matrix  $\tilde{\mathbf{W}}^{(i)}$  for that layer. The gate variables  $\mathbf{z}$  are not modified and are thus not a function of the parametric variable. As a result, all of the “time steps” share the same sparsity regularization allowing us to forego any further modifications to implement group sparsity.

The advantage of this architecture is that it does not need to replicate the EQL network itself, thus saving on computational resources (especially for large  $N_t$ ). The architecture can also make predictions on a continuous domain of  $t$  and does not need to restrict the data to fixed points in time. More specifically, rather than viewing the dataset as Equation 4, we have greater flexibility and can view the dataset as

$$\mathcal{D} = \left\{ x^{(i)}, y^{(i)}, t^{(i)} \right\}_{i=1}^N \quad (8)$$

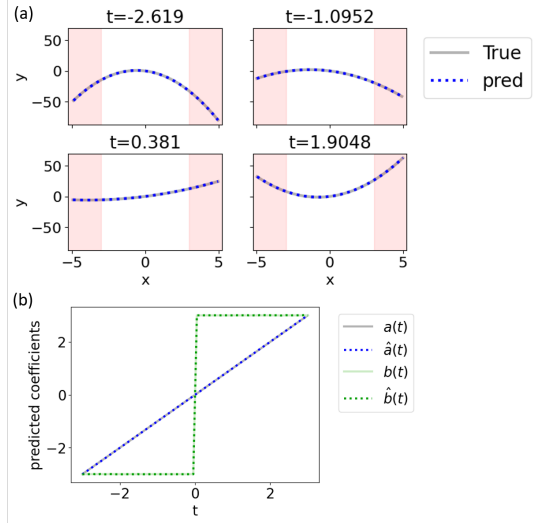


Fig. 3. Results of the SEQL network for learning  $f_3(x, t) = t \cdot x^2 + 3 \operatorname{sgn}(t) \cdot x$ . (a) Predictions for select values of  $t$ . Outputs with  $|x| > 3$  (highlighted in red) are extrapolated. (b) Learned coefficient functions.

Although we do not explicitly regularize the functional space of the parametric coefficients, neural networks tend to generalize well despite typically being overparameterized, which is a topic of significant interest [27]–[30]. In practice, this means that the predictions of neural networks for regression tasks tend to be smooth, and so the function of the parametric coefficient will also tend to be smooth.

### III. RESULTS

We now look at several different problem settings with parametric quantities that can be analyzed by our system. For simplicity, we highlight some of the results here, and the remainder can be found in the appendix.

#### A. Analytic Expressions

To verify the ability of the PEQL and SEQL networks to discover parametric equations, we benchmark the networks on analytical expressions in Table I, where  $\operatorname{sgn}$  is the *sign* function (also known as the *signum* function).

TABLE I  
PARAMETRIC EQUATIONS FOR BENCHMARKING

Label	Equations
$f_1$	$t \cdot x$
$f_2$	$t \cdot x^2 + 3 \sin(t) \cdot x$
$f_3$	$t \cdot x^2 + 3 \operatorname{sgn}(t) \cdot x$
$f_4$	$\sin\left(\frac{5+t}{2} \cdot x\right)$

Here we show results for benchmarks  $f_3$  and  $f_4$ , and the remainder of the results can be found in Appendix B. While we train the networks on data drawn from the domain  $x \in [-3, 3]$ , we evaluate the networks on a wider domain  $x \in [-5, 5]$  to test extrapolation performance.

Figure 3 shows the results for learning  $f_3$  using the SEQL network. Figure 3(a) plots the values of the true function versus the predicted function for various values of  $t$ . The

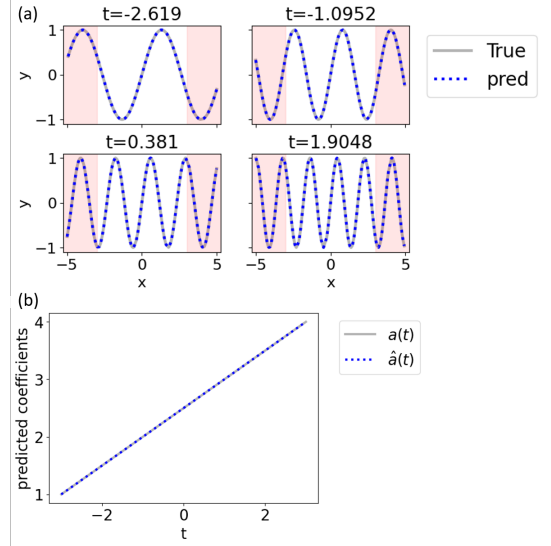


Fig. 4. Results of the PEQL network for learning  $f_4(x, t) = \sin\left(\frac{5+t}{2} \cdot x\right)$ . (a) Predictions for select values of  $t$ . Outputs with  $|x| > 3$  (highlighted in red) are extrapolated. (b) Learned coefficient functions.

prediction matches the true function extremely well not only in the training regime, but also in the test regime, demonstrating that the neural network is able to extrapolate. The extracted equations for these time steps are shown in Table II. We see that the network has successfully discovered the function  $\hat{f}_3 = a(t)x^2 + b(t)x + \epsilon(t)$  where  $a(t)$  and  $b(t)$  are the parametric coefficients and  $\epsilon$  is a small number that can either be eliminated with further training or ignored upon inspection. The true versus predicted parametric coefficients match extremely closely, as seen in Figure 3(b). Note that the SEQL network is able to learn the discontinuous  $\operatorname{sgn}$  function without any apparent smoothing at  $t = 0$ . Discontinuous coefficients would be difficult to learn using methods that rely on local averaging or smoothing, such as Ref. [5], [6]. We can also see in Table II that the PEQL network performs similarly well as the SEQL network.

We also show the results of learning  $f_4$  using the PEQL network in Figure 4. The network has learned the equation  $\hat{f}_4 = \sin(a(t)x)$  where  $a(x)$  is plotted in Figure 4(b). Again, the predictions match the true function extremely well across time steps and outside of the training regime. Although sinusoidal functions are typically difficult to learn through linear regression techniques, the PEQL network is able to learn this function across multiple spatial frequencies.

Note that because the varying coefficient is inside the  $\operatorname{sgn}$  and  $\sin$  functions for  $f_3$  and  $f_4$ , respectively, methods such as from refs. [6] or [7] that rely on linear regression techniques would not be able to discover these types of equations. However, the multi-layer architecture of the SEQL and PEQL networks allow for the varying coefficient to be inside nested functions, enabling discovery of much more complex parametric equations.

TABLE II  
LEARNED EQUATIONS ON SELECT  $t$  VALUES FOR THE FUNCTION  $f_3(t, x) = t \cdot x^2 + 3 \operatorname{sgn}(t) \cdot x$ .

$t$	True	SEQL	PEQL
-2.619	$-2.62x^2 - 3.00x$	$-2.62x^2 - 3.00x - 0.04$	$-2.63x^2 - 3.02x - 0.03$
-1.095	$-1.10x^2 - 3.00x$	$-1.10x^2 - 3.00x - 0.02$	$-1.11x^2 - 3.00x - 0.01$
0.381	$0.38x^2 + 3.00x$	$0.38x^2 + 3.00x + 0.01$	$0.39x^2 + 3.01x$
1.905	$1.90x^2 + 3.00x$	$1.91x^2 + 3.00x + 0.02$	$1.91x^2 + 3.02x + 0.03$

TABLE III  
LEARNED EQUATIONS ON SELECT  $x$  VALUES FOR THE ADVECTION-DIFFUSION EQUATION.

$x$	True	SEQL	PEQL
-4.375	$-0.89u - 0.79u_x + 0.10u_{xx}$	$-0.86u - 0.77u_x + 0.11u_{xx}$	$-0.86u - 0.76u_x + 0.11u_{xx}$
-1.875	$0.89u - 2.21u_x + 0.10u_{xx}$	$0.83u - 2.14u_x + 0.07u_{xx}$	$0.86u - 2.16u_x + 0.09u_{xx}$
0.625	$-0.89u - 0.79u_x + 0.10u_{xx}$	$-0.86u - 0.77u_x + 0.11u_{xx}$	$-0.86u - 0.77u_x + 0.11u_{xx}$
3.125	$0.89u - 2.21u_x + 0.10u_{xx}$	$0.83u - 2.14u_x + 0.07u_{xx}$	$0.88u - 2.17u_x + 0.10u_{xx}$

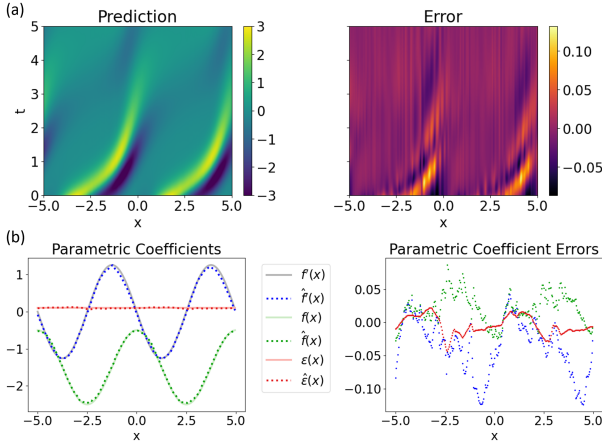


Fig. 5. Results for learning the advection-diffusion equation using the PEQL network. (a) Prediction values and errors of  $u_t$ . (b) Predicted coefficient functions and prediction errors.

## B. PDEs

We now look at two of the PDE datasets investigated in ref. [31]. In ref. [31], the partial differential terms (e.g.  $u_x, u_{xx}$ ) and their combinations (e.g.  $uu_x$ ) were pre-computed and fed into SINDy to discover the governing PDE. Here we pre-compute the individual partial differential terms, but we do not explicitly pre-compute the combinations.

1) *Advection-Diffusion Equation*: The advection-diffusion equation describes numerous physical transport systems and has been applied to describe the movement of pollutants, reservoir flow, heat, and semiconductors. We use an adaptation of the equation that includes a spatially-dependent velocity field, as in [31]:

$$u_t = f'(x)u + f(x)u_x + \epsilon u_{xx}. \quad (9)$$

where  $f(x) = -1.5 + \cos\left(\frac{2\pi x}{5}\right)$  and  $\epsilon = 0.1$ . Note that the parametric quantities vary with respect to space rather than time.

Table III shows the equations that the SEQL and PEQL have learned after training for few instances of  $x$ . We see that both networks have learned an equation of the form  $\hat{u}_t = \hat{f}'(x)u + \hat{f}(x)u_x + \hat{\epsilon}(x)u_{xx}$ , and have thus successfully discovered the equation structure. The predicted  $\hat{u}_t$  along with the learned

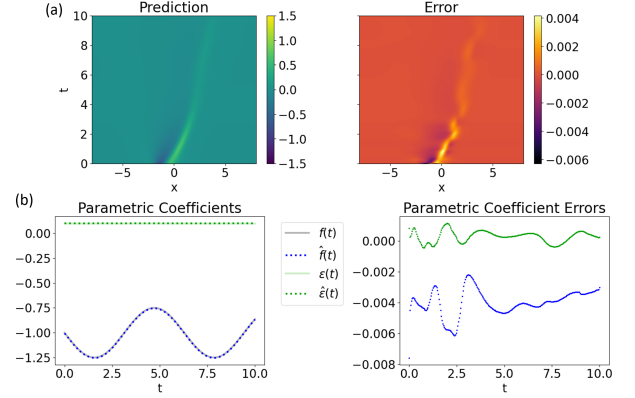


Fig. 6. Results for learning Burgers' equation using the SEQL network. (a) Predicted vs. actual values of  $u_t$ . (b) Predicted coefficient functions and prediction errors.

parametric coefficients ( $\hat{f}'(x)$ ,  $\hat{f}(x)$ ,  $\hat{\epsilon}(x)$ ) are shown in Figure 5 for the PEQL network. The predicted values match the actual values very closely. Again, note that the predicted coefficients by the fully-connected neural network are smooth as a function of  $x$  despite the lack of explicit regularization.

2) *Burgers' Equation*: Burgers' equation is an important differential equation originally proposed to model turbulent flow but has been applied to other processes such as traffic flow and boundary layer behavior. Here we analyze Burgers' equation with an oscillating coefficient for the non-linear term, as in [31]:

$$u_t = f(t)uu_x + \epsilon u_{xx}. \quad (10)$$

where  $f(t) = -\left(1 + \frac{\sin(t)}{4}\right)$  and  $\epsilon = 0.1$ .

As before, both the SEQL and PEQL networks are able to accurately discover the correct equation. We see in Figure 6 that the SEQL network is able to accurately predict the function and the parametric coefficients. Note that while ref. [31] needs to pre-compute product terms of the individual spatial derivatives such as  $uu_x$  and  $u_x^2 u_{xxx}$  and then perform a linear regression over these terms, the SEQL is able to learn non-linear relations on its own using the multiplication primitive function. So the SEQL is only given  $u$ ,  $u_x$ ,  $u_{xx}$ , ... as inputs, but is able to learn the form of the nonlinear PDE.



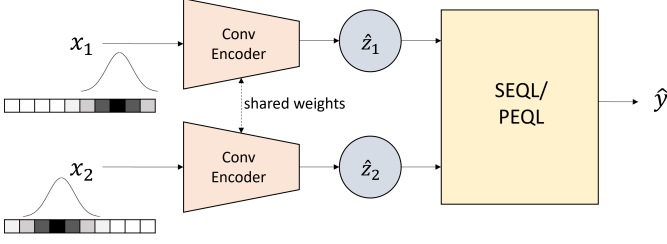


Fig. 7. The combined architecture used for high-dimensional system tasks involving a convolutional encoder followed by an EQL network.

### C. Spring System

Finally, we demonstrate the ability of the parametric EQL networks to perform symbolic regression on structured, high-dimensional data by integrating with other deep learning architectures and training end-to-end.

We consider a dataset that consists of pairs of 1D images of point particles that interact through a spring-like force. The input data is a 1D grayscale image with 64 pixels which represents a 1D spatial domain  $\psi \in [-4, 4]$ . Each image contains a single particle, represented by a Gaussian with mean centered at its position  $\psi_i$  and a fixed variance of 0.1. We look at two different targets for symbolic regression: the spring force  $F = -k(t)(\psi_2 - \psi_1)$  and the spring energy  $E = \frac{k(t)}{2}(\psi_2 - \psi_1)^2$ , where  $k(t) = \frac{5-t}{2}$ . The spring constant decreases over time, which we can imagine is representative of a spring degrading with use.

To approach this problem, we use the architecture shown in Figure 7. Each image is fed into a separate encoder, where the two encoders share the same weights. The encoder consists of 2 convolutional layers followed by 3 fully-connected layers and a batch normalization layer. The encoders each output a single-dimensional latent variable  $\hat{z}_1, \hat{z}_2$ , which are then fed into the parametric EQL network (which can be either the PEQL or the SEQL). The batch normalization layer serves to constrain the range of the latent variable so that the EQL network does not need to scale to arbitrarily-sized inputs when training end-to-end. The EQL network has a single scalar output, which is trained to match either the spring force or the spring energy. The entire network is trained end-to-end and is only shown the inputs and the output, but must learn an appropriate representation  $\hat{z}_i$ . While there are no constraints on the latent representation  $\hat{z}_i$ , we expect it to have a one-to-one mapping to the true position of the particle,  $\psi_i$ .

For all tests, 512 training data points with  $\psi_1, \psi_2 \in [-3, 3]$  were sampled for each of 128 fixed values of  $t \in [-3, 3]$ . To evaluate the extrapolation ability of these architectures, training data points were restricted to pairs with  $|\psi_2 - \psi_1| \leq 4$ , while no such restriction was imposed on testing data. In addition, we compare against a baseline test of a model consisting of the same encoder architecture with a dense ReLU network replacing EQL network. We call this baseline the **ReLU network**.

Results for learning the spring force is shown in Figure 8. Both the SEQL network and the ReLU network successfully predict the force inside the training domain, but only the SEQL network is able to extrapolate outside of the training regime

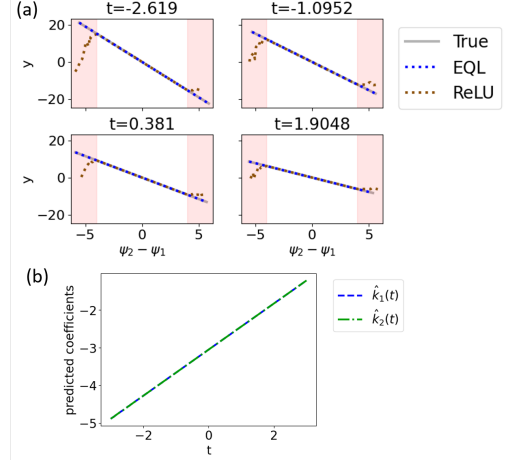


Fig. 8. Results for learning the spring force  $F$  using the SEQL network. (a) Predictions for select values of  $t$ . Outputs with  $|\psi_2 - \psi_1| > 4$  (highlighted in red) are extrapolated. (b) Coefficient functions in the equation  $\hat{F}(t, \hat{z}_1, \hat{z}_2) = \hat{k}_1(t) \cdot \hat{z}_1 - \hat{k}_2(t) \cdot \hat{z}_2$  learned by the SEQL network.

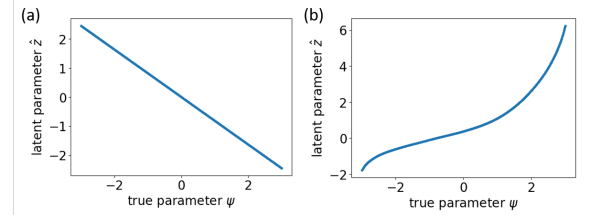


Fig. 9. Latent variable encodings for the function  $f(t, \psi_1, \psi_2) = -\frac{5-t}{2} \cdot (\psi_2 - \psi_1)$  learned by (a) the convolutional SEQL network and (b) the ReLU network.

whereas the ReLU network completely fails to extrapolate. Additionally, the EQL network learns the governing equation as shown in Table IV, with the learned parametric coefficient plotted in Figure 8(b). Note that the SEQL network learns the expression  $\hat{F} = \hat{k}_1(t)\hat{z}_1 - \hat{k}_2(t)\hat{z}_2$ , where we do not necessarily have  $\hat{k}_1 = \hat{k}_2$ . Upon inspection, however, we see that  $\hat{k}_1(t) \approx \hat{k}_2(t)$  and so the SEQL network has discovered an approximately equal expression to what we expect.

Additionally, while the SEQL network discovers an equation in terms of  $\hat{z}_{1,2}$ , it also learns a linear mapping of the latent variable to the true position as shown in Figure 9(a). While there is no explicit constraint or regularization placed on the latent space, because the EQL network must learn to use the latent variable to form the equation, the end-to-end training of the architecture forces the mapping to be an analytical transformation of the original variable, which in this case is a linear mapping. In contrast, the latent variable mapping for the ReLU network is shown in Figure 9(b). While it is one-to-one, it is not linear since there is no bias to make the mapping linear. Using this linear mapping, we can perform a linear regression to find the approximate relationship between  $\hat{z}$  and  $\psi$  and reconstruct the discovered equation in terms of  $\hat{\psi}$ , which is shown in the right-most column of Table IV.

We see similar results for the spring potential data, this time using the PEQL network, in Figures 10 and Table V. Again, the PEQL network is able to extrapolate outside of the training

TABLE IV

LEARNED EQUATIONS OF THE STACKED NETWORK ON SELECT  $t$  VALUES FOR THE SPRING FORCE FUNCTION  $F(t, \psi_1, \psi_2) = -\frac{5-t}{2} \cdot (\psi_2 - \psi_1)$  IN THE LATENT SPACE AND TRANSFORMED TO THE ORIGINAL PARAMETER SPACE.

$t$	True	Learned Latent	Learned Transformed
-2.619	$-3.81(\psi_2 - \psi_1)$	$-4.66\hat{z}_1 + 4.66\hat{z}_2$	$3.82\hat{\psi}_1 - 3.82\hat{\psi}_2$
-1.095	$-3.05(\psi_2 - \psi_1)$	$-3.72\hat{z}_1 + 3.72\hat{z}_2$	$3.05\hat{\psi}_1 - 3.05\hat{\psi}_2$
0.381	$-2.31(\psi_2 - \psi_1)$	$-2.82\hat{z}_1 + 2.82\hat{z}_2$	$2.31\hat{\psi}_1 - 2.31\hat{\psi}_2$
1.905	$-1.55(\psi_2 - \psi_1)$	$-1.89\hat{z}_1 + 1.89\hat{z}_2$	$1.55\hat{\psi}_1 - 1.55\hat{\psi}_2$

TABLE V

LEARNED EQUATIONS OF THE PARAMETERIZED NETWORK ON SELECT  $t$  VALUES FOR THE FUNCTION  $E(t, \psi_1, \psi_2) = \frac{5-t}{4} \cdot (\psi_2 - \psi_1)^2$  IN THE LATENT SPACE AND TRANSFORMED TO THE ORIGINAL PARAMETER SPACE.

$t$	True	Learned Latent	Learned Transformed
-2.619	$-1.90(\psi_2 - \psi_1)^2$	$6.59\hat{z}_1^2 + 6.59\hat{z}_2^2 - 13.18\hat{z}_1\hat{z}_2 + 0.02$	$1.91\hat{\psi}_1^2 + 1.91\hat{\psi}_2^2 - 3.82\hat{\psi}_1\hat{\psi}_2 + 0.02$
-1.095	$-1.52(\psi_2 - \psi_1)^2$	$5.27\hat{z}_1^2 + 5.27\hat{z}_2^2 - 10.55\hat{z}_1\hat{z}_2 + 0.01$	$1.53\hat{\psi}_1^2 + 1.53\hat{\psi}_2^2 - 3.06\hat{\psi}_1\hat{\psi}_2 + 0.01$
0.381	$-1.16(\psi_2 - \psi_1)^2$	$4.01\hat{z}_1^2 + 4.01\hat{z}_2^2 - 8.02\hat{z}_1\hat{z}_2 + 0.01$	$1.16\hat{\psi}_1^2 + 1.16\hat{\psi}_2^2 - 2.33\hat{\psi}_1\hat{\psi}_2 + 0.01$
1.905	$-0.77(\psi_2 - \psi_1)^2$	$2.64\hat{z}_1^2 + 2.64\hat{z}_2^2 - 5.28\hat{z}_1\hat{z}_2 + 0.01$	$0.77\hat{\psi}_1^2 + 0.77\hat{\psi}_2^2 - 1.53\hat{\psi}_1\hat{\psi}_2 + 0.01$

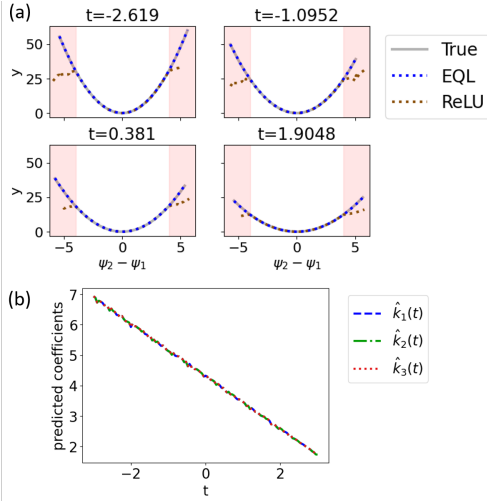


Fig. 10. Results for learning the spring energy  $E$  using the PEQL network. (a) Predictions for select values of  $t$ . Outputs with  $|\psi_2 - \psi_1| > 4$  (highlighted in red) are extrapolated. (b) Coefficient functions in the equation  $f(t, \hat{z}_1, \hat{z}_2) = \hat{k}_1(t) \cdot \hat{z}_1^2 + \hat{k}_2(t) \cdot \hat{z}_2^2 - 2\hat{k}_3(t) \cdot \hat{z}_1\hat{z}_2$  learned by the PEQL network.

regime whereas the ReLU network fails to extrapolate. Note that in this case, the PEQL learns the equation  $\hat{E}(t, \hat{z}_1, \hat{z}_2) = \hat{k}_1(t)\hat{z}_1^2 + \hat{k}_2(t)\hat{z}_2^2 - 2\hat{k}_3(t)\hat{z}_1\hat{z}_2 + \epsilon(t)$  where  $\hat{k}_1 \approx \hat{k}_2 \approx \hat{k}_3$  and  $\epsilon$  is small. Thus, the PEQL network has discovered the correct equation.

#### IV. DISCUSSION

We have proposed two different variants of the EQL network—the stacked architecture (SEQL) and the parameterized architecture (PEQL)—to enable neural network-based symbolic regression of parametric systems where coefficients may vary. We have demonstrated our system on parametric analytic equations and PDEs, as well as a dataset encoded as images. Our method has the potential to combine the power of deep learning and symbolic regression to enable scientific discovery on complex and high-dimensional datasets.

We note that in our experiments we used analytic expressions for the varying coefficients for simplicity. However, our

method is not constrained to these types of expressions, and the parametric coefficient can more generally be any arbitrary function. Thus, our method can be applied to systems that we know are partially governed by an analytic equation, but partially governed by some other mechanism that may be too complex or noisy to capture. This is similar in spirit to methods for solving PDEs that replace part of the equation with a neural network, often to correct for discretization errors [32], [33].

As far as the comparison of the two architectures goes, all results for both architectures can be found in the Appendix. For a moderate number of time steps (e.g.  $N_t < 512$ ) the SEQL has fewer parameters than the PEQL; despite this, however, the PEQL trains on each minibatch  $3.7\times$  faster than the SEQL on the analytic equations for our settings of hyper-parameters and network sizes. This is likely because the limiting factor is the computation of the activation functions, which must be processed separately for each component of  $h$  (whereas in a conventional neural network the use of a single activation function is able to take advantage of vectorization optimizations). For a larger number of time steps, (e.g.  $N_t > 512$ ), the PEQL is more memory-efficient as well since the SEQL parameters scale linearly with the number of time steps. Thus, the PEQL is able to scale to larger datasets. In terms of the data format, prior methods rely on gridded data [5], [31] while both the SEQL and the PEQL allow a variable grid along the varying dimension. The PEQL architecture takes this flexibility a step further in that it is able to interpolate in time and make predictions at arbitrary time points, whereas the stacked architecture is fixed to certain time points. On the other hand, we find that the stacked architecture is less sensitive to the random initialization and converges more quickly to the solution. Thus, we have a tradeoff between performance and flexibility. One possible direction for future work to bridge this gap is to introduce different learning rate schedules for the EQL network and the MWU in the parameterized architecture, as the EQL network typically requires large learning rates to escape local minima and converge, whereas large learning rates may be detrimental to the MWU.

The parametric architectures presented here can be viewed as implementing functional regularization. Functional regular-

ization, which imposes regularization on the learned function rather than on the parameters, is attractive as it is much more intuitive and can lead to more natural methods for tasks such as continual learning [34], [35]. It has been explored in neural networks through regularizing the predictions on batch of the data [34] and through defining the prior over functions rather than weights in the case of Bayesian neural networks [36], [37]. In the case of the EQL network, the coefficients of the resulting equation are typically very simple functions (oftentimes the identity function) of the weights themselves. This means that in practice, the  $L_2$  smoothing regularization in the stacked EQL network architecture often implicitly applies to the *function space*, even though we are explicitly applying the regularization in the *weight space*. In the case of the parametric EQL architecture, the output of fully-connected neural networks will tend to be smooth due to modern training methods such as stochastic gradient descent (which is a deep topic of great interest in the literature), and so the MWU itself acts as a regularization on the function space of the EQL network.

#### ACKNOWLEDGMENTS

We would like to thank Rumen Dangovski, Anka Hu, and Amber Li for insightful discussions and work on related projects. This work is supported in part by the MIT UROP program, the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, <http://iaifi.org/>), the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program, and the Air Force Office of Scientific Research under the award number FA9550-21-1-0317. Research was sponsored by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

#### APPENDIX A TRAINING DETAILS

All neural network architectures are implemented in TensorFlow [38]. The network is trained using the RMSProp optimizer, and the following loss function:

$$\mathcal{L} = \frac{1}{N} \sum (\hat{y}_i - y_i)^2 + \lambda L_r, \quad (11)$$

where  $N$  is the mini-batch size,  $\lambda$  is the regularization weight, and  $L_r$  is the total regularization. For the parameterized architecture  $L_r = \mathcal{L}_R$  is simply the sparsity regularization, while the stacked architecture has an additional term  $L_r = \mathcal{L}_R + \theta \mathcal{L}_S$  to incorporate the smooth weight regularization described in II-D.

For both learning rate and regularization weight schedules, we use a one cycle policy, as shown in Figure 11. We start off

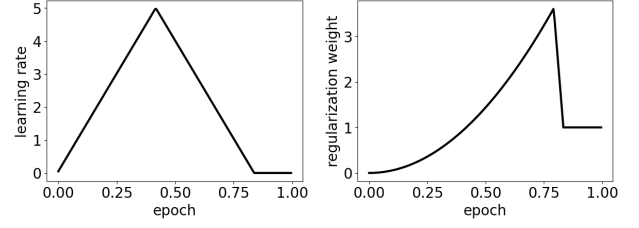


Fig. 11. (Left) Learning rate and (right) regularization weight schedules during training relative to `base_lr` and `base_rw`.

with a small learning rate and regularization to ensure the EQL network settles into a stable configuration containing many different terms such that the network weights do not explode. The learning rate is ramped up to allow the EQL network escape local minima in search of global minima, and the regularization is likewise increased to pare down the number of terms. Finally, we expect the EQL network to have learned the correct equation structure partway through training, and so we decrease learning rate and regularization to fine-tune the weights and optimize primarily for MSE.

To extract the learned equation from the trained EQL network, we can simply multiply the weights by the primitive functions using symbolic mathematics. We implement this using SymPy, which can automatically simplify the expression [39]. Additionally, we use a thresholding procedure in the final expression where we drop terms where the coefficient is smaller than a threshold, which we set to 0.01.

#### APPENDIX B ADDITIONAL RESULTS

##### Analytic Expression

For all tests, 512 training data points with  $x \in [-3, 3]$  are sampled for each of 128 fixed values of  $t \in [-3, 3]$  for a total of  $512 \cdot 128 = 65536$  training examples. To test generalization, the parametric EQL architectures are evaluated on test data points with  $x \in [-5, 5]$ .

Due to sensitivity of the parametric EQL architectures to the random initialization of network weights, 80 trials were run for each function. In practice, the networks only need to learn the correct equation once over a reasonable number of trials, since it is possible to construct a validation method that selects the best equation from a set of learned equations. For all the results in this paper, we simply select the trial with the lowest generalization error. Other considerations that can be integrated in the validation process are equation simplicity and prior beliefs about the equation form, for example.

The results for the analytical expressions  $f_1$ ,  $f_2$ , and  $f_4$  are shown in Table VI ( $f_3$  is shown in Table II in the main text). Both the SEQL and PEQL match the true equations very closely. To compare the SEQL and PEQL, we also show the errors of the predicted parametric coefficients on  $f_2$  and  $f_3$  in Figure 12. We see that the PEQL tends to have a larger prediction error than the SEQL.



TABLE VI  
LEARNED EQUATIONS FOR ANALYTIC EXPRESSIONS.

Benchmark	$t$	True	SEQL	PEQL
$f_1(t, x) = t \cdot x$	-2.619	$-2.62x$	$-2.62x$	$-2.62x - 0.02$
	-1.095	$-1.10x$	$-1.10x$	$-1.10x$
	0.381	$0.38x$	$0.38x$	$0.38x$
	1.905	$1.90x$	$1.91x$	$1.91x$
$f_2(t, x) = t \cdot x^2 + 3 \sin(t) \cdot x$	-2.619	$-2.62x^2 - 1.50x$	$-2.62x^2 - 1.51x - 0.08$	$-2.62x^2 - 1.50x + 0.01$
	-1.095	$-1.10x^2 - 2.67x$	$-1.09x^2 - 2.68x - 0.08$	$-1.10x^2 - 2.66x + 0.02$
	0.381	$0.38x^2 + 1.12x$	$0.38x^2 + 1.12x + 0.01$	$0.38x^2 + 1.12x$
	1.905	$1.90x^2 + 2.83x$	$1.90x^2 + 2.85x + 0.06$	$1.90x^2 + 2.83x$
$f_4(t, x) = \sin\left(\frac{5+t}{2} \cdot x\right)$	-2.619	$\sin(1.19x)$	$\sin(1.19x)$	$\sin(1.19x)$
	-1.095	$\sin(1.95x)$	$\sin(1.95x)$	$\sin(1.95x)$
	0.381	$\sin(2.69x)$	$\sin(2.69x)$	$\sin(2.69x)$
	1.905	$\sin(3.45x)$	$\sin(3.45x)$	$\sin(3.45x)$

TABLE VII  
LEARNED EQUATIONS FOR BURGERS' EQUATION.

$x$	True	SEQL	PEQL
0.627	$-1.15uu_x + 0.10u_{xx}$	$-1.15uu_x + 0.10u_{xx}$	$-1.16uu_x + 0.10u_{xx}$
3.137	$-1.00uu_x + 0.10u_{xx}$	$-1.00uu_x + 0.10u_{xx}$	$-1.01uu_x + 0.10u_{xx}$
5.647	$-0.85uu_x + 0.10u_{xx}$	$-0.86uu_x + 0.10u_{xx}$	$-0.85uu_x + 0.10u_{xx}$
8.157	$-1.24uu_x + 0.10u_{xx}$	$-1.24uu_x + 0.10u_{xx}$	$-1.25uu_x + 0.10u_{xx}$

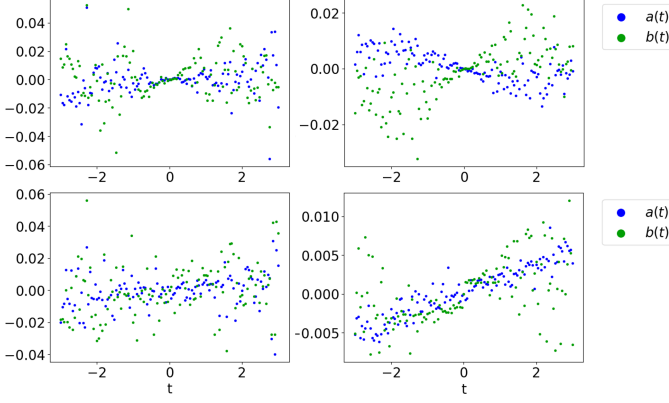


Fig. 12. Prediction errors of the parametric coefficients for the (left) PEQL and the (right) SEQL on the analytical expressions (top)  $f_2$  and (bottom)  $f_3$ .

### Differential Equations

For the advection-diffusion equation, data was sampled from 256 different points in the  $x$ -domain and 512 different points in the  $t$ -domain, for a total of  $256 \cdot 512 = 131072$  examples. The equation is solved numerically using a spectral method on the domain  $x \in [-5, 5]$  and  $t \in [0, 5]$  with  $f(x) = -1.5 + \cos\left(\frac{2\pi x}{5}\right)$  and  $\epsilon = 0.1$  using code from [31].

For the Burgers' equation, data was sampled from 512 different points in the  $x$ -domain and 256 different points in the  $t$ -domain, for a total of  $512 \cdot 256 = 131072$  examples. The equation was solved numerically using a spectral method on the domain  $x \in [-8, 8]$  and  $t \in [0, 10]$  using code from [31]. Similar to the analytic expression experiments, 80 trials were run for each equation and the trial with the lowest generalization error was selected.

Figure 13 displays results for SEQL on the advection-diffusion equation, and Figure 14 displays results for PEQL on Burgers' equation. The learned equations for Burgers' equations are listed in table VII. The SEQL and PEQL

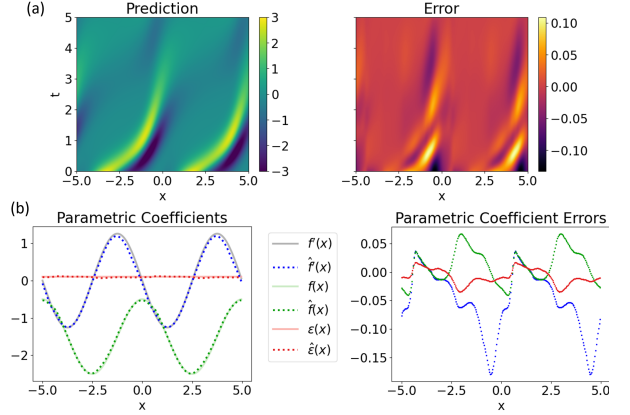


Fig. 13. Results for learning the advection-diffusion equation using the SEQL network. (a) Predicted vs. actual values of  $u_t$ . (b) Predicted coefficient functions and prediction errors.

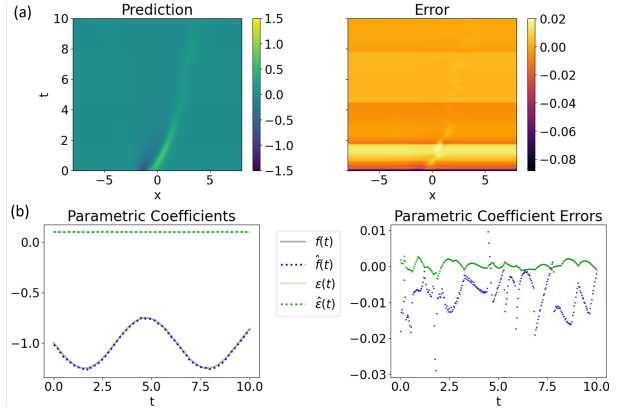


Fig. 14. Results for learning Burgers' equation using the PEQL network. (a) Predicted vs. actual values of  $u_t$ . (b) Predicted coefficient functions and prediction errors.

networks have similar level of errors in learning the parametric coefficients for the advection-diffusion equation. However, for the Burgers' equation, the PEQL network has significantly larger error than the SEQL network, although it is still able to predict the equation accurately.

## REFERENCES

- [1] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.
- [2] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Židek, A. Potapenko *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.
- [3] J. Koza, "Genetic programming as a means for programming computers by natural selection," *Statistics and Computing*, vol. 4, no. 2, pp. 87–112, jun 1994. [Online]. Available: <http://link.springer.com/10.1007/BF00175355>
- [4] M. Schmidt and H. Lipson, "Distilling free-form natural laws from experimental data," *Science (New York, N.Y.)*, vol. 324, no. 5923, pp. 81–5, apr 2009. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/19342586>
- [5] H. Xu, D. Zhang, and J. Zeng, "Deep-learning of parametric partial differential equations from sparse and noisy data," *Physics of Fluids*, vol. 33, no. 3, p. 037132, 2021.
- [6] Y. Luo, Q. Liu, Y. Chen, W. Hu, and J. Zhu, "Ko-pde: Kernel optimized discovery of partial differential equations with varying coefficients," *arXiv preprint arXiv:2106.01078*, 2021.
- [7] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the national academy of sciences*, vol. 113, no. 15, pp. 3932–3937, 2016.
- [8] S.-M. Udrescu and M. Tegmark, "Ai feynman: A physics-inspired method for symbolic regression," *Science Advances*, vol. 6, no. 16, p. eaay2631, 2020.
- [9] K. Champion, B. Lusch, J. N. Kutz, and S. L. Brunton, "Data-driven discovery of coordinates and governing equations," *Proceedings of the National Academy of Sciences*, vol. 116, no. 45, pp. 22 445–22 451, 2019.
- [10] Z. Long, Y. Lu, and B. Dong, "Pde-net 2.0: Learning pdes from data with a numeric-symbolic hybrid deep network," *Journal of Computational Physics*, vol. 399, p. 108925, 2019.
- [11] P. Y. Lu, J. Ariño, and M. Soljačić, "Discovering sparse interpretable dynamics from partial observations," *arXiv preprint arXiv:2107.10879*, 2021.
- [12] M. Cranmer, A. Sanchez Gonzalez, P. Battaglia, R. Xu, K. Cranmer, D. Spergel, and S. Ho, "Discovering symbolic models from deep learning with inductive biases," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 429–17 442, 2020.
- [13] G. Martius and C. H. Lampert, "Extrapolation and learning equations," *arXiv preprint arXiv:1610.02995*, oct 2016. [Online]. Available: <http://arxiv.org/abs/1610.02995>
- [14] S. Sahoo, C. Lampert, and G. Martius, "Learning equations for extrapolation and control," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4442–4450.
- [15] S. Kim, P. Y. Lu, S. Mukherjee, M. Gilbert, L. Jing, V. Čeperić, and M. Soljačić, "Integration of neural network-based symbolic regression in deep learning for scientific discovery," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 9, pp. 4166–4177, 2020.
- [16] A. Costa, R. Dangovski, O. Dugan, S. Kim, P. Goyal, M. Soljačić, and J. Jacobson, "Fast neural models for symbolic regression at scale," *arXiv preprint arXiv:2007.10784*, 2020.
- [17] C. Louizos, M. Welling, and D. P. Kingma, "Learning Sparse Neural Networks through  $\mathcal{L}_0$  Regularization," *arXiv preprint arXiv:1712.01312*, dec 2017. [Online]. Available: <https://arxiv.org/abs/1712.01312>
- [18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [19] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [20] D. Ha, A. Dai, and Q. V. Le, "Hypernetworks," *arXiv preprint arXiv:1609.09106*, 2016.
- [21] M. Andrychowicz, M. Denil, S. Gomez, M. W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, and N. De Freitas, "Learning to learn by gradient descent by gradient descent," *Advances in neural information processing systems*, vol. 29, 2016.
- [22] T. Munkhdalai and H. Yu, "Meta networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2554–2563.
- [23] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," 2016.
- [24] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *arXiv preprint arXiv:2004.05439*, 2020.
- [25] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," *Advances in neural information processing systems*, vol. 31, 2018.
- [26] F. de Avila Belbute-Peres, Y.-f. Chen, and F. Sha, "Hyperpinn: Learning parameterized differential equations with physics-informed hypernetworks," in *The Symbiosis of Deep Learning and Differential Equations*, 2021.
- [27] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, "Deep double descent: Where bigger models and more data hurt," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2021, no. 12, p. 124003, 2021.
- [28] J. Liu, G. Jiang, Y. Bai, T. Chen, and H. Wang, "Understanding why neural networks generalize well through gsnr of parameters," *arXiv preprint arXiv:2001.07384*, 2020.
- [29] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," *arXiv preprint arXiv:2201.03545*, 2022.
- [30] D. Jakubovitz, R. Giryes, and M. R. Rodrigues, "Generalization error in deep learning," in *Compressed sensing and its applications*. Springer, 2019, pp. 153–193.
- [31] S. Rudy, A. Alla, S. L. Brunton, and J. N. Kutz, "Data-driven identification of parametric partial differential equations," *SIAM Journal on Applied Dynamical Systems*, vol. 18, no. 2, pp. 643–660, 2019. [Online]. Available: <https://doi.org/10.1137/18M1191944>
- [32] J. Pathak, M. Mustafa, K. Kashinath, E. Motheau, T. Kurth, and M. Day, "Using machine learning to augment coarse-grid computational fluid dynamics simulations," *arXiv preprint arXiv:2010.00072*, 2020.
- [33] D. Kochkov, J. A. Smith, A. Alieva, Q. Wang, M. P. Brenner, and S. Hoyer, "Machine learning-accelerated computational fluid dynamics," *Proceedings of the National Academy of Sciences*, vol. 118, no. 21, 2021.
- [34] A. S. Benjamin, D. Rolnick, and K. Kording, "Measuring and regularizing networks in function space," *arXiv preprint arXiv:1805.08289*, 2018.
- [35] P. Pan, S. Swaroop, A. Immer, R. Eschenhagen, R. Turner, and M. E. E. Khan, "Continual deep learning by functional regularisation of memorable past," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4453–4464, 2020.
- [36] S. Sun, G. Zhang, J. Shi, and R. Grosse, "Functional variational bayesian neural networks," *arXiv preprint arXiv:1903.05779*, 2019.
- [37] T. G. Rudner, Z. Chen, and Y. Gal, "Rethinking function-space variational inference in bayesian neural networks," in *Third Symposium on Advances in Approximate Bayesian Inference*, 2020.
- [38] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: <https://www.tensorflow.org/>
- [39] A. Meurer, C. P. Smith, M. Paprocki, O. Čertík, S. B. Kirpichev, M. Rocklin, A. Kumar, S. Ivanov, J. K. Moore, S. Singh, T. Rathnayake, S. Vig, B. E. Granger, R. P. Muller, F. Bonazzi, H. Gupta, S. Vats, F. Johansson, F. Pedregosa, M. J. Curry, A. R. Terrel, v. Roučka, A. Saboo, I. Fernando, S. Kulal, R. Cimman, and A. Scopatz, "Sympy: symbolic computing in python," *PeerJ Computer Science*, vol. 3, p. e103, Jan. 2017. [Online]. Available: <https://doi.org/10.7717/peerj-cs.103>