
Lie Point Symmetry Data Augmentation for Neural PDE Solvers

Johannes Brandstetter^{† 1 2} Max Welling¹ Daniel E. Worrall^{‡ 3}

Abstract

Neural networks are increasingly being used to solve *partial differential equations* (PDEs), replacing slower numerical solvers. However, a critical issue is that neural PDE solvers require high-quality ground truth data, which usually must come from the very solvers they are designed to replace. Thus, we are presented with a proverbial chicken-and-egg problem. In this paper, we present a method, which can partially alleviate this problem, by improving neural PDE solver sample complexity—Lie point symmetry data augmentation (LPSDA). In the context of PDEs, it turns out that we are able to quantitatively derive an exhaustive list of data transformations, based on the *Lie point symmetry group* of the PDEs in question, something not possible in other application areas. We present this framework and demonstrate how it can easily be deployed to improve neural PDE solver sample complexity by an order of magnitude.

1. Introduction

Simulation of the physical world around us is an important component of many numerate disciplines. From computational fluid dynamics to molecular modeling (Lelièvre & Stoltz, 2016) or astronomical simulation (Courant et al., 1967), solving the underlying dynamical equations is usually analytically intractable. The equations themselves are regularly expressed as *partial differential equations* (PDEs), solutions of which require efficient and accurate *solvers*. In reality, however, classical solvers are expensive and custom designed per PDE class (Quarteroni & Quarteroni, 2009).

In recent years, with the reascendance of deep learning, it has become popular to learn PDE solvers (Greydanus et al.,

2019; Bar-Sinai et al., 2019; Sanchez-Gonzalez et al., 2020; Thuerey et al., 2021), circumventing the lengthy and often tedious process of solver design. But we are left with a proverbial ‘chicken-and-egg problem’. From where do we obtain the abundant data needed to train said neural solvers? It has to be generated with a classical solver, after all.

It seems the best we can do is to generate ‘groundtruth’ data using a high-quality, slow ‘teacher’ and then clone its behavior with a cheaper neural ‘student’—a form of so-called *model order reduction* (Schilders, 2008). This considered, a natural question is whether we can train student models with less of this ‘expensive’ data. More precisely, how do we improve solver *sample complexity*? One potential route is to exploit symmetries of the loss with respect to input transformations, building in equivariance (Wang et al., 2021). However, the design of equivariant layers is a difficult task.

A much simpler traditional method is data augmentation (Simard et al., 2003). However, knowing how to augment data and to what extent is a routine question faced by ML practitioners. It turns out that PDEs are uniquely special in that it is possible to characterize the full set of permissible augmentations for each PDE. We shall define permissible augmentations as *Lie point symmetries* in Section 2. As such, for learnable PDE solvers, we shall show that we are able to define exactly what kinds of data augmentation we need, placing a typically intuition-based aspect of deep learning pipeline design on a firm mathematical footing.

The contributions of this paper are:

- *Lie point symmetry data augmentation* (LPSDA), a mathematically-grounded treatment of data augmentation for neural PDE solvers. LPSDA comprises the full set of solution-preserving, pointwise, continuous data transformations for any (analytic) PDE.
- Demonstration that LPSDA improves neural PDE solver sample complexity to over an order of magnitude on a SOTA neural PDE solver, across multiple PDEs (of evolution type) and training modalities.
- Ablations and interpolation on the effect of individual symmetries on neural PDE solvers.
- Demonstration that LPSDA is faster and stabler than running a classical solver to produce the same solution trajectories and measurement of the equivariance properties of classical and neural solvers.

¹University of Amsterdam ²Johannes Kepler University Linz
[†]now at Microsoft Research ³Qualcomm AI Research, an initiative of Qualcomm Technologies, Inc [‡]now at Deepmind. Correspondence to: Johannes Brandstetter <brandstetter@ml.jku.at>.

2. Background

Here we cover PDEs and their Lie point symmetries. For a textbook we suggest [Olver \(1986\)](#) as an excellent resource or [Oliveri \(2010\)](#) as a more easily digestible review paper.

Partial Differential Equations. A partial differential equation (PDE) Δ specifies a relationship between a *solution* $\mathbf{u} : \mathcal{X} \rightarrow \mathbb{R}^n$ and its derivatives at all points \mathbf{x} in the domain $\mathcal{X} \subset \mathbb{R}^m$. It is convenient to define the *prolongation* of \mathbf{u} as

$$\text{pr}^{(n)}\mathbf{u} = (\mathbf{u}, \mathbf{u}_{\mathbf{x}}, \mathbf{u}_{\mathbf{x}\mathbf{x}}, \dots, \mathbf{u}_{n\mathbf{x}}), \quad (1)$$

where $\mathbf{u}_{\mathbf{x}}$ is all first-order derivatives of \mathbf{u} with respect to the *independent variable* \mathbf{x} , $\mathbf{u}_{\mathbf{x}\mathbf{x}}$ is all unique second-order derivatives of \mathbf{u} (e.g. $\partial_{xx}\mathbf{u}, \partial_{xy}\mathbf{u}, \partial_{yx}\mathbf{u}$), and so forth up until n^{th} order. The range of the prolonged function $\text{pr}^{(n)}\mathbf{u}$ lives in the so-called *jet space* $\mathcal{U}^{(n)}$. PDEs are thus represented as a system of algebraic equations

$$\Delta(\mathbf{x}, \text{pr}^{(n)}\mathbf{u}|_{\mathbf{x}}) = 0, \quad \text{for all } \mathbf{x} \in \mathcal{X}. \quad (2)$$

By way of example, the 1 + 1 dimensional Heat equation in space and time can be written

$$\Delta(\mathbf{x}, \text{pr}^{(2)}u|_{\mathbf{x}}) = u_t - \alpha u_{xx} = 0 \quad (3)$$

for independent variables $\mathbf{x} = (x, t)$ and thermal diffusivity $\alpha > 0$. We use unbolded type for scalar-valued variables. The set $\mathcal{S}_{\Delta} = \{(\mathbf{x}, \mathbf{u}^{(n)}|_{\mathbf{x}}) | \Delta(\mathbf{x}, \mathbf{u}^{(n)}|_{\mathbf{x}}) = 0\} \subset \mathcal{X} \times \mathcal{U}^{(n)}$, where $\mathbf{u}^{(n)} = \text{pr}^{(n)}\mathbf{u}$, can be thought of as a (collection of) surface(s) in *total space* $\mathcal{X} \times \mathcal{U}^{(n)}$, is called an *algebraic subvariety* ([Olver, 1986, p. 79](#)). It is the graph¹ of all prolonged solutions. Thinking in these geometrical terms, we are equipped to define *symmetries* of PDEs.

Lie Point Symmetries of PDEs. If a function \mathbf{u} satisfies Equation 2 at point \mathbf{x} , then \mathbf{u} is called the *solution* of Δ at \mathbf{x} . We leave a discussion on boundary conditions for later. Symmetries are classes of transformations, which map the solution set back into itself. For solution \mathbf{u} we define its image after transformation as $g\mathbf{u}$. g could abstractly represent a rotation, or a shear, or a flip. The set G of all solutions preserving g is called the *symmetry group* of Δ . If $g\mathbf{u}$ represents a transformed \mathbf{u} , then $\text{pr}^{(n)}g \cdot \mathbf{u}^{(n)}$ represents the corresponding transformation of the prolongation $\mathbf{u}^{(n)}$. Then $(\mathbf{x}, \mathbf{u}^{(n)}|_{\mathbf{x}}) \in \mathcal{S}_{\Delta}$ implies $(g\mathbf{x}, \text{pr}^{(n)}g \cdot \mathbf{u}^{(n)}|_{\mathbf{x}}) \in \mathcal{S}_{\Delta}$. Defining $g(\mathbf{x}, \mathbf{u}^{(n)}|_{\mathbf{x}}) = (g\mathbf{x}, \text{pr}^{(n)}g \cdot \mathbf{u}^{(n)}|_{\mathbf{x}})$ we then have

$$g\mathcal{S}_{\Delta} = \mathcal{S}_{\Delta}. \quad (4)$$

$g\mathcal{S}_{\Delta}$ is the set where g is applied to every element of \mathcal{S}_{Δ} .

¹It is unfortunate that this nomenclature is already so overloaded in our field. Graph here refers to the surface one would draw if plotting $\mathbf{u}^{(n)}$ with pencil and paper.

The symmetry group of Δ has the structure of a mathematical group: a set, which is closed under composition, is associative, has an identity element, and for which each element has an inverse. For PDEs, there exists a subgroup², within G , with additional structure; namely, it is a *Lie point symmetry*. Lie point symmetries are Lie groups—classes of smooth transformations with smooth inverse—that act on the $(\mathbf{x}, \mathbf{u}|_{\mathbf{x}})$ representation of a function *pointwise*, as

$$(\mathbf{x}, \mathbf{u}|_{\mathbf{x}}) \xrightarrow{g} (g\mathbf{x}, g\mathbf{u}|_{\mathbf{x}}), \quad \text{for all } g \text{ and } (\mathbf{x}, \mathbf{u}). \quad (5)$$

This maps points to points, hence the name. In particular, g may not depend on derivatives of \mathbf{u} , otherwise it is a *Lie-Bäcklund transformation*, not considered in this paper. Examples of Lie point symmetries are rotations, translations, Galilean boosts, scalings, and shears. A counterexample is a blur, which neither acts pointwise, since it performs averages over neighborhoods, nor is invertible, thus not a group-structured transformation. Figure 1 shows the four Lie point symmetries of the Korteweg-de Vries equation.

The significance of PDE symmetries has a long history dating back to Sophus Lie, who famously developed the theory of Lie groups to better solve ordinary differential equations. Most importantly for us, if we have procured a solution \mathbf{u} (by any means necessary), we can generate a collection of new solutions $\{g\mathbf{u}\}_{g \in G}$ for free, given computing $g\mathbf{u}$ is relatively cheap, which we show in Section 4.7.

One-parameter Lie groups. The Lie point symmetries of a given PDE can be multidimensional and it is thus useful to break this down into modular components. Any d dimensional simply-connected Lie group can be decomposed ([Olver, 1986, Equation 1.40](#)) into a series of *one-parameter transformations* $g_i : \mathbb{R} \rightarrow G$ for $i = 1, \dots, d$, with the property $g_i(\delta)g_j(\epsilon) = g_i(\delta + \epsilon)$ if $i = j$, such that

$$g = g_1(\epsilon_1)g_2(\epsilon_2) \cdots g_d(\epsilon_d). \quad (6)$$

Note that $g_i(\delta)g_j(\epsilon) \neq g_j(\epsilon)g_i(\delta)$ for $i \neq j$. Each g_i is a smooth, group-structured map from the real parameter ϵ_i on to the group, and is thus also referred to as a *one-parameter group*. For us, each one-parameter group will correspond to a different kind of data augmentation we can apply. The *caveat that G is simply-connected* is important. It means we only consider transformations that can be continuously parameterized, for instance, 2D rotations but not reflections.

For each PDE Δ , a complete set of simply-connected Lie point symmetries can be derived in the form of a collection of one-parameter Lie groups ([Olver, 1986, § 2.4](#)). There is extensive literature tabulating these symmetries for a vast array of common PDEs (e.g., [Ibragimov \(1993\)](#)) and for novel PDEs software exists to derive them symbolically (e.g., [Baumann \(1998\)](#)).

²Group within a group

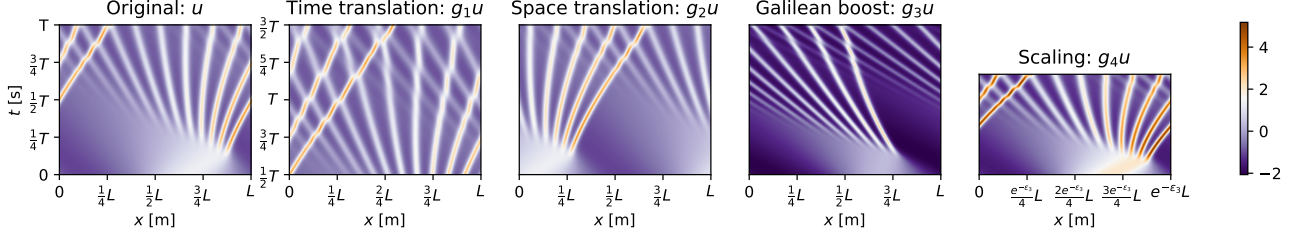


Figure 1. One-parameter Lie point symmetries of the *Korteweg-de Vries equation*. The transformations are: (g_1) time translation, (g_2) space translation, (g_3) Galilean-like boost, and (g_4) scaling. The domain is periodic in the x -direction with units of meters and the time axis has units of seconds. Each plot has been shown for random transformations with shared spatial-scaling and colorspace.

3. Lie Point Symmetry Data Augmentation

Here we show how Lie point symmetries can be used as a data augmentation technique for a learned neural PDE solver. We elucidate our method with a worked example.

Given a PDE Δ , one can look up its Lie point symmetries as one-parameter transformations $\{g_1, \dots, g_d\}$ using, say, [Ibragimov \(1993\)](#) or [Baumann \(1998\)](#). Theory tells us that list is exhaustive, under regularity conditions (Section 3.1). Each symmetry g_i is a function of a single variable ϵ_i , controlling the magnitude of the augmentation. For instance ϵ_i might correspond to a rotation angle or spatial shift. At training time, we sample each ϵ_i independently from an appropriate probability density $p(\epsilon_i)$ for $i = 1, \dots, d$. Each density is chosen using standard techniques, such as cross-validation. At training time, one can then augment the training set \mathcal{D} by sampling a solution $\mathbf{u} \in \mathcal{D}$ and augmenting it as

$$\mathbf{u}' = g_d(\epsilon_d) \cdots g_1(\epsilon_1) \mathbf{u}. \quad (7)$$

In words, we take \mathbf{u} , apply transformation $g_1(\epsilon_1)$, followed by $g_2(\epsilon_2)$, and so forth, where each $\epsilon_1, \epsilon_2, \dots$ is sampled from its own density. How \mathcal{D} is generated is beyond the scope of this paper, but it may consist of known analytical solutions, simulation data, or real-world data. The latter two methods will contain noise meaning that \mathbf{u} does not lie within the solution set, implying that $g\mathbf{u}$ does not either. We assume this noise is sufficiently small, so while $g\mathbf{u}$ may not lie exactly in the solution set, it is situated very close to it.

Worked Example: The Korteweg-de Vries Equation.

By way of example, let's consider the *Korteweg-de Vries* (KdV) equation ([Boussinesq, 1877](#), p. 360), a nonlinear scalar-valued PDE in space and time, made famous in the 1960s ([Zabusky & Kruskal, 1965](#)) for exhibiting *solitons*, solitary weakly-interacting “wave-pulses”. Example solutions can be found in Figure 1. The KdV equation is

$$\Delta((x, t), u^{(3)}) = u_t + uu_x + u_{xxx} = 0. \quad (8)$$

Notable is the nonlinear convective term uu_x and the dispersive term u_{xxx} , which in traditional solvers require careful attention. We solve this equation on a domain $[0, L]$

for times $[0, T]$. We impose periodic boundary conditions $u(0, t) = u(L, t)$ and initial conditions $u(x, 0) = u_0(x)$ for all x and t . Note $\mathcal{X} = [0, L] \times [0, T]$. We discretize \mathcal{X} uniformly on a grid $\mathbf{X} = \{(x_i, t_j)\}_{i,j}$ of N_x spatial points and N_t timesteps. A typical neural PDE solver takes as input initial conditions $\mathbf{u}_0 = \{u_0(\mathbf{x}) | \mathbf{x} \in \mathbf{X}_0\}$ evaluated at $\mathbf{X}_0 = \{\mathbf{x} \in \mathbf{X} | t = 0\}$ and outputs predictions to match targets $\hat{\mathbf{u}} = \{\mathbf{u}(\mathbf{x}) | \mathbf{x} \in \mathbf{X}\}$. We augment this data as

$$(\mathbf{X}_0, \mathbf{u}_0, \mathbf{X}, \hat{\mathbf{u}}) \xrightarrow{g} (g\mathbf{X}_0, g\mathbf{u}_0, g\mathbf{X}, g\hat{\mathbf{u}}) \quad (9)$$

with g sampled according to Equation 7 from the Lie point symmetry group of the KdV equation. These are namely

$$g_1(\epsilon)(x, t, u) = (x, t + \epsilon, u) \quad \text{time shift,} \quad (10)$$

$$g_2(\epsilon)(x, t, u) = (x + \epsilon, t, u) \quad \text{space shift,} \quad (11)$$

$$g_3(\epsilon)(x, t, u) = (x + \epsilon t, t, u + \epsilon) \quad \text{Galilean boost,} \quad (12)$$

$$g_4(\epsilon)(x, t, u) = (e^\epsilon x, e^{3\epsilon t}, e^{-2\epsilon} u) \quad \text{scaling,} \quad (13)$$

which can be readily found in [Olver \(1986, p. 126\)](#)—note the sign flip in g_3 and g_4 . Working with these transformations poses some technical design issues, discussed next.

Resampling. A transformation, say Equation 11, maps the grid points $(x, t) \mapsto (gx, gt) = (x + \epsilon, t)$. This poses an issue for Eulerian solvers, where the grid \mathbf{X} is always fixed. Augmented solutions (gx, gt, gu) thus have to be resampled on \mathbf{X} . In this work, we use trigonometric interpolation in the x -direction. A deeper discussion is found in Appendix B.

3.1. How complete is our symmetry list?

Theory tells us our list of symmetries is exhaustive if the PDE is ‘nondegenerate’ ([Olver, 1986, Theorem 2.71](#)), a technical condition that Δ is of *maximal rank* and *locally solvable*. Intuitively, local solvability is the condition the PDE is smooth enough that points $(\mathbf{x}, \mathbf{u}^{(n)})$ satisfying Δ can be interpolated locally by sufficiently smooth \mathbf{u} . Local solvability is implied if Δ is analytic in its arguments, a corollary of the Cauchy-Kovalevskaya theorem ([Olver, 1986, Corollary 2.74](#)). In practice though, we need not worry ourselves with demonstrating nondegeneracy.

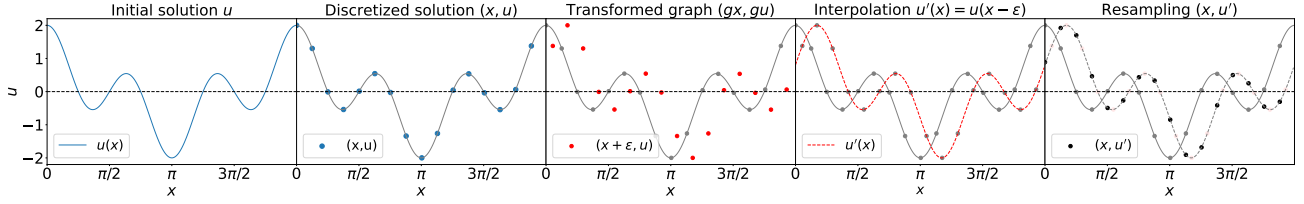


Figure 2. Schematic of the data augmentation process: From left to right we see how a solution u is drawn, discretized into its graph (x, u) , transformed as per the symmetry (translation), interpolated with a smooth interpolant, and resampled on to the original mesh \mathbf{X} .

3.2. A note on boundary conditions

The symmetries listed are defined on open neighborhoods about the origin, where ϵ is sufficiently small. In general, symmetries break at boundaries where open neighborhoods cannot be defined, requiring special care. We work with wrapped domains to avoid this issue. One could foresee, however, that we could sidestep the problem of boundaries by performing data augmentation on local patches, but we leave that for future work. Despite working in this restricted scenario, we are still left with sufficiently many interesting symmetries, which we demonstrate can improve generalization performance in our experiments.

4. Experiments

We ran experiments across different equations (Section 4.2), models (Section 4.3), training setups (Section 4.4), and choices of symmetries (Section 4.5). We tested the effectiveness of adding symmetries on raw generalization performance and survival time and demonstrate that adding LPSDA is equivalent to using larger training sets by up to a factor of $\times 16$ in size on the datasets used. We also probe the effect of LPSDA on long rollout stability (Section 4.6), and we measure the equivariance error of the learned models, comparing with classical solvers (Section 4.7).

4.1. Experimental setup

Equations. We chose to use 1D *evolution equations*:

$$u_t - F(x, u, u_x, u_{xx}, \dots) = 0, \quad (14)$$

as a test bed for LPSDA; although, our method can be applied to other classes of PDE. Unless otherwise stated (e.g., Section 4.2), we run most ablation experiments on the KdV equation. Notwithstanding, we also consider 1) the Kuramoto-Shivashinsky (KS) equation and 2) the Burgers' equation. The KS equation is known for its chaotic behavior:

$$u_t + u_{xx} + u_{xxxx} + uu_x = 0. \quad (15)$$

Above being chaotic, it is nonlinear and the biharmonic term u_{xxxx} is hard to model. The Burgers' equation is

$$u_t + uu_x - \nu u_{xx} = 0, \quad (16)$$

for viscosity $\nu \geq 0$. It exhibits shock formation, which becomes discontinuities when $\nu = 0$. In our experiments, we only work with viscid systems ($\nu > 0$) where derivatives exist everywhere. Each PDE has symmetries listed in Table 1. The Burgers' equation has an extra 'infinite dimensional subalgebra' g_α , which is not a Lie group symmetry, but for any two solutions u, α , we can return a new one as $2\nu \log \left((1 - \epsilon) e^{\frac{1}{2\nu} u(x,t)} + \epsilon e^{\frac{1}{2\nu} \alpha(x,t)} \right)$ for $\epsilon \in [0, 1]$. Training data generation is described in Appendix A.

Meshes and parameters. Unless otherwise stated, we set $\Omega = [0, L]$ discretized uniformly over $N_x = 256$ points with periodic boundaries and set $t \in [0, T]$ at $N_{t_{\text{out}}} = 100$ uniformly sampled timesteps. Equation specific settings are in Table 2. Initial conditions are sampled per Bar-Sinai et al. (2019) from a distribution over truncated Fourier series with random coefficients $\{A_k, \ell_k, \phi_k\}_k$ as

$$u_0(x) = \sum_{k=1}^K A_k \sin(2\pi \ell_k x / L + \phi_k). \quad (17)$$

Neural PDE Solvers. We experimented with two models: a litmus test that data augmentation can benefit more than one model class. These were: a baseline 1D ResNet-like model (He et al., 2016), and an adapted version of the Fourier Neural Operator (FNO) (Li et al., 2020), which is SOTA for regular meshes with periodic boundaries. The exact architectures are in Appendix C. Both models are trained in two different ways following Brandstetter et al. (2022): (i) a *neural operator method* and (ii) an *autoregressive method*.

Training methodologies. A neural operator (NO) is a mapping $\mathcal{M} : [0, T] \times \mathcal{F} \rightarrow \mathcal{F}$, where \mathcal{F} is a (possibly infinite-dimensional) function space, trained to satisfy

$$\mathcal{M}(t, \mathbf{u}_0) = \mathbf{u}(t). \quad (18)$$

It maps initial conditions \mathbf{u}_0 directly to solutions \mathbf{u} at time t . The flavor we use predicts all $N_{t_{\text{out}}}$ future times simultaneously, a technique called *temporal bundling* (Brandstetter et al., 2022). The initial conditions are also extended in time as a trajectory of the last $N_{t_{\text{in}}} = 20$ steps. This assumes our prediction task has a fixed rollout length. An autoregressive

Table 1. Lie point symmetries as function of $\epsilon \in \mathbb{R}$ for studied PDEs. $\sigma : \mathbb{R} \rightarrow [0, 1]$ is any squashing function. Sources: KdV Olver (1986, Ex. 2.44), KS (derivative form): Cvitanović et al. (2010), Burgers’: Hoarau & David (2007) and Olver (1986, Ex. 2.42).

EQUATION	g_1	g_2	g_3	g_4	g_α
KdV	$(x, t + \epsilon, u)$	$(x + \epsilon, t, u)$	$(x + \epsilon t, t, u + \epsilon)$	$(e^\epsilon x, e^{3\epsilon t}, e^{-2\epsilon} u)$	
KS	$(x, t + \epsilon, u)$	$(x + \epsilon, t, u)$	$(x + \epsilon t, t, u + \epsilon)$		
Burgers’	$(x, t + \epsilon, u)$	$(x + \epsilon, t, u)$	$(x, t, u + \epsilon)$	$(e^\epsilon x, e^{2\epsilon t}, u)$	$(u, t, 2\nu \log((1 - \sigma(\epsilon))e^{\frac{1}{2\nu}u} + \sigma(\epsilon)e^{\frac{1}{2\nu}\alpha}))$

(AR) method on the other hand solves the PDE iteratively, allowing it to run for variable length (up to multiples of the temporal bundling size). For time-dependent PDEs with time-independent coefficients, the solution is computed as

$$\mathbf{u}(t + \Delta t) = \mathcal{A}(\Delta t, \mathbf{u}(t)), \quad (19)$$

where $\mathcal{A} : \mathbb{R}_{>0} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ is the temporal update. Again, the flavor we use is a temporally bundled variant, which predicts $N_{t_{AR}}$ future timesteps simultaneously in each recurrent iteration. A schematic of these two training methodologies is shown in Figure 3. We trained in these two ways so that LPSDA is tested in different task setups.

Evaluation metrics. We report in terms of rollout averaged normalized MSE (NMSE), defined as

$$\mathcal{L} = \frac{1}{N_t} \sum_{j=1}^{N_t} \frac{\|\mathbf{u}(t_j) - \hat{\mathbf{u}}(t_j)\|_2^2}{\|\hat{\mathbf{u}}(t_j)\|_2^2}, \quad (20)$$

where \mathbf{u} is the target and $\hat{\mathbf{u}}$ is the model output. The normalized MSE is itself dimensionless and invariant to rescaling of the spatial axes. However, the average over timesteps is not invariant to time rescaling, with dimensions of error per timestep. We chose to report this error instead of error per second, because a) we can compare across different PDEs and b) we only need look at relative improvement of LPSDA on top of an arbitrary baseline without augmentation. All errors are reported with ± 2 standard deviation error bars (95% confidence intervals), computed with bootstrapping.

4.2. Different PDEs

We tested LPSDA on the KdV, KS, and Burgers’ equations. Each equation has symmetries listed in Table 1. Figure 4a shows sample complexity improvements on FNO, trained as a neural operator to predict T s into the future, as per Table 2. This shows LPSDA improves sample complexity by

Table 2. Experimental choices for studied PDEs as per Bar-Sinai et al. (2019). Ranges are sampled from uniformly.

	L	T	K	A_k	ℓ_k	ϕ_k
KdV	$[0.9, 1.1] \cdot 128$	$[0.9, 1.1] \cdot 40$	10	$[-0.5, 0.5]$	$\{1, 2, 3\}$	$[0, 2\pi]$
KS	$[0.9, 1.1] \cdot 64$	$[0.9, 1.1] \cdot 20$	10	$[-0.5, 0.5]$	$\{1, 2, 3\}$	$[0, 2\pi]$
BURGERS’	$[0.9, 1.1] \cdot 2\pi$	10	20	$[-0.5, 0.5]$	$\{3, 4, 5, 6\}$	$[0, 2\pi]$

$\times 4$ to $\times 16$. LPSDA works most strikingly on the Burgers’ equation, where it appears the infinite-dimensional subalgebra g_α is doing most of the work. This symmetry takes two old solutions and combines them into a new one, in effect squaring the size of the dataset probably accounting for the near doubling in slope we observe (Burgers’ + G line).

We plot on log-log axes and fit straight lines to the data, based on the documented phenomenon that neural network test performance follows a power law (Kaplan et al., 2020). This experiment gives a strong indication that LPSDA could be effective for a great many other PDEs. Typically roll-outs from the KS equation along with per-cell normalized squared errors is shown in Figure 5. Errors increase with time. LPSDA appears to delay the rate at which errors grow.

4.3. Different models

Figure 4b shows the effect of LPSDA for the KdV equation. We compare the FNO and ResNet trained as neural operator methods. In development the ResNet struggled to model the solution manifold, so for a fairer comparison we made the problem easier by predicting $T = 20$ s into the future instead of $T = 40$ s. Bold lines show test NMSE across training set size without LPSDA (\emptyset) and dashed lines show it with LPSDA ($g_1 g_2 g_3 g_4$). We see that LPSDA is equivalent to increasing the training set size by a factor between $\times 4$ and $\times 8$. Again, we fitted straight lines because we are in the neural operator model setting. It is interesting that LPSDA improves each model by a differing amount, especially FNO, where the sample complexity improvement appears to grow with increasing dataset size.

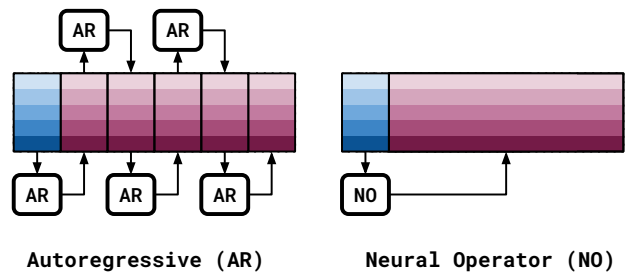


Figure 3. Autoregressive models predict the solution recurrently; whereas neural operator methods predict it directly.

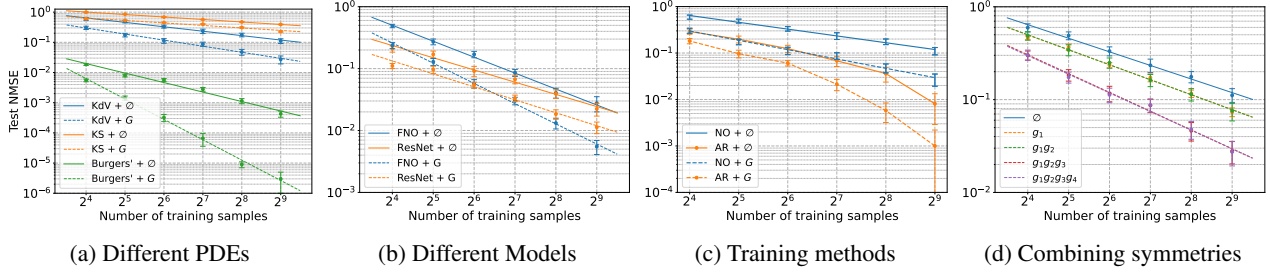


Figure 4. We compare test set normalized MSE with and without LPSDA across training set sizes. All errorbars are ± 2 std, calculated via bootstrap. A comprehensive summary of results can be found in Table 3 in the appendix. (a): LPSDA improves sample complexity on the KdV and KS equations by $\times 4$ to $\times 8$, (b) LPSDA improves sample complexity on Fourier Neural Operator and Residual Networks on the KdV equation, (c) LPSDA is also effective across neural operator (NO) and autoregressive (AR) training methodologies on the KdV equation, (d) on the KdV equation time translation (g_1) and Galilean boosts (g_3) appear to help a lot, but space translation (g_2) and scaling (g_4) do not. Code available at <https://github.com/brandstetter-johannes/LPSDA>.

4.4. Different Training Methods

Figure 4c shows the LPSDA ablation for the FNO model trained either as a neural operator (blue), predicting $N_{t_{\text{out}}} = 100$ timesteps in to the future in one shot, or as an autoregressive method (orange), predicting $N_{t_{\text{AR}}} = 20$ timesteps at a time. Again LPSDA improves both methods. This experiment is important because rollout stability and accuracy of autoregressive methods—a still poorly understood phenomenon—appears to be intimately connected with generalization capacity (Um et al., 2021; Brandstetter et al., 2022). LPSDA, in improving generalization performance, naturally improves AR methods as well as NO methods. It is interesting to note that the generalization performance of the AR method does not follow a power law.

4.5. Incremental effect of symmetries

The effect of incrementally adding symmetries on the KdV-FNO neural operator model is shown in Figure 4d. In Appendix D we show that this effect is repeated in other scenar-

ios. We see that with this setup the time translation (g_1) and Galilean boost (g_3) improve sample complexity. Interestingly, space translation (g_2) neither improves nor degrades performance. This is because the models we use are already translation equivariant. Furthermore, it appears scaling (g_4) does not help either. Our best explanation is that with our parameterization scale is ‘too simple’ a data transformation, simply multiplying solution values by a random constant. It is interesting to observe that not all symmetries improve performance equally; this corroborates what is commonly observed with classical data augmentation.

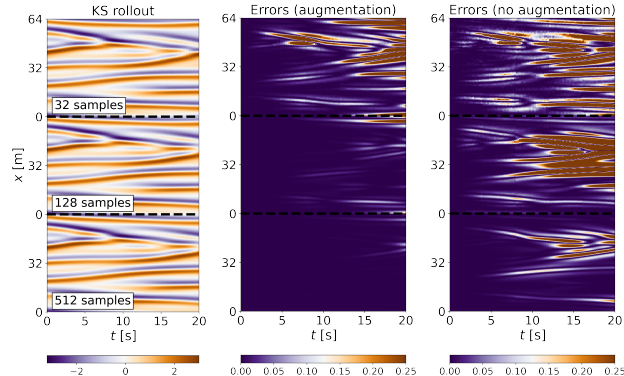


Figure 5. Visualisation of sample rollouts of the KS equation along with the corresponding per-cell normalized squared error.

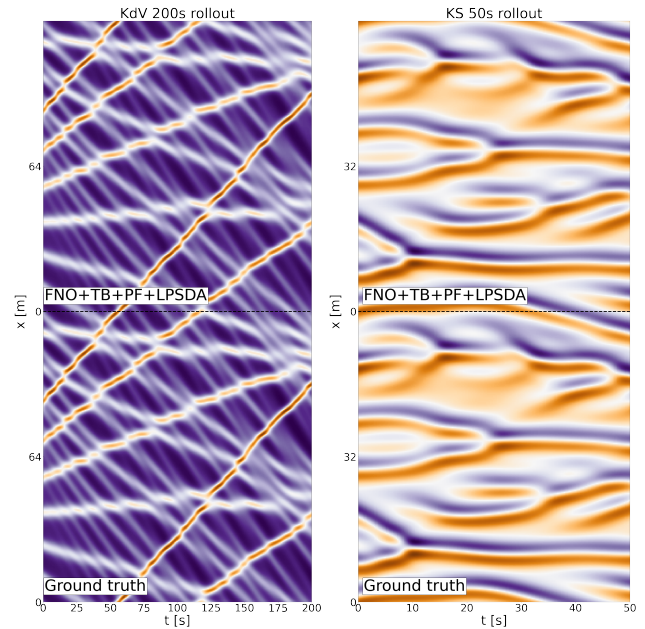


Figure 6. Examples of ‘long rollouts’ of $N_{t_{\text{out}}} = 400$ timesteps on the KdV and KS equations. Visually it is hard to spot errors.

4.6. Very long rollouts

We ran long rollouts on the KdV (KS) equation of $T = 200$ s ($T = 50$ s) with $N_{t_{\text{out}}} = 400$ timesteps to check how LPSDA can improve performance in a practical use case. We used an AR-trained FNO model. We compare a naïve model predicting $N_{t_{\text{AR}}} = 1$ timestep at a time (\emptyset), a temporally bundled model predicting $N_{t_{\text{AR}}} = 120$ timesteps at a time (TB), the $N_{t_{\text{AR}}} = 120$ model with the pushforward trick (Brandstetter et al., 2022) (TB+PF) needed for long run stability, and (TB+PF+LPSDA) where we add LPSDA. Figure 6 shows the resulting rollouts and Figure 7a shows the cumulative NMSE with respect to time. Visually it is difficult to spot errors. The error graphs, on log-lin plots, reveal interesting subexponential cumulative error growth with time. The uparrows for the \emptyset rollouts indicate error explosion. Here, again, we see that LPSDA improves rollout performance. A summary of all results is in Appendix D.

4.7. How equivariant are classical/neural solvers

Figure 7b shows the equivariance properties of classical solvers for KdV Galilean boost parameter ϵ_3 after $T = 100$ s. For model output \mathbf{u} at time T , this is $\|\mathbf{u}(g\mathbf{u}_0) - g\mathbf{u}(\mathbf{u}_0)\|_2^2 / \|\mathbf{u}(g\mathbf{u}_0)\|_2^2$. Ground truth was solved at tolerance $1e-12$. Blue dashed lines are isotherms of constant log solver tolerance (base 10). This error is mainly a function of solver tolerance. The red solid lines are isotherms of constant solution time. Thus more time is needed to maintain a constant error for larger $|\epsilon_3|$. Neural solvers, on the other hand, have constant time overhead (our AR-FNO model measured 38 ms), but without error control. Figure 7c shows equivariance error of neural models trained with and without LPSDA. LPSDA clearly encourages equivariance around the training range $\epsilon_3 \in [-0.4, 0.4]$. Classical solvers have superior equivariance error are two orders slower. We also considered autocorrelative error defined as $\|\mathbf{u}(g\mathbf{u}_0) - \mathbf{u}(\mathbf{u}_0)\|_2^2 / \|\mathbf{u}(\mathbf{u}_0)\|_2^2$. Models trained without LPSDA map $\mathbf{u}(g\mathbf{u}_0)$ close to $\mathbf{u}(\mathbf{u}_0)$ for small ϵ_3 . We suspect the \emptyset model outputs solutions close to regions of the solution manifold it has observed, where $\epsilon_3 \simeq 0$. This shows how data augmentation is important for generalization.

Speed and Difficulty. We highlight the ϵ_3 dependence of the time isotherms in Figure 7b. Adaptive timestep solvers limit the error incurred across a given unit of time by expending more operations on more ‘difficult’ trajectories. Consider a wave moving across a discrete spatial grid. To compute its position at a future timestep, we are limited by the maximum speed information can propagate across the grid. This must exceed the speed of the wavefront, otherwise we have to reduce the time increment. Faster waves require finer timesteps. For the KdV equation, we can control wavespeed by adding ϵ_3 to the initial conditions. Larger $|\epsilon_3|$ leads to faster waves, and thus longer solver times. By

contrast, data augmentation is several orders of magnitude faster (our implementation: 0.76 ms). This indicates training data generation could be sped up by solving ‘simple’ problems on a classical solver, and then transforming solutions into challenging regions of the training data space.

5. Related Work

The closest work to ours is Wang et al. (2021) who design neural PDE solvers equivariant to Lie point symmetries of the Navier-Stokes equation and Heat equation. While this method is sensible and effective, they are only able to incorporate a single symmetry (other than translation and time) at a time. With our method, we can learn equivariance to multiple symmetries at once. It is also simpler to implement a data augmentation pipeline than a equivariant network. Furthermore, the computational budget for equivariant networks is exponential in the number of symmetries, a limitation if you have 6 or 7 symmetries, a scenario unseen in the equivariance literature (Bronstein et al., 2021).

In the context of classical PDE solvers Hoarau & David (2007); Hoarau et al. (2007) analyzed the stability properties of various finite difference schemes under the action of the individual symmetries of the Burgers’ equation. Studying equivariance properties of classical solvers dates back to 1994 with Dorodnitsyn (1994). Valiquette & Winternitz (2005) discretize scalar-valued PDEs with two independent variables such that their Lie point symmetries are preserved.

Symplectic integrators (Hairer et al., 1993, § II.16), such as leapfrogging (Neal, 1993, p. 72), have a long history in machine learning for their utility in Hamiltonian Monte Carlo (Duane et al., 1987) and (Neal, 1993, § 5). These are limited to integrating symplectic systems, such as Hamiltonians, which abound in physical systems. More recent neural models learn Hamiltonians directly (Greydanus et al., 2019) or indirectly (Sanchez-Gonzalez et al., 2019; Cranmer et al., 2020; Finzi et al., 2020b). Bar-Sinai et al. (2019); Kochkov et al. (2021) explicitly conserve volume (another integral of motion) for conservation form evolution equations via finite volume projection on their solver outputs.

In the field of neural PDE solvers, there are neural operator methods of finite dimensionality (Raissi, 2018; Sirignano & Spiliopoulos, 2018; Bhatnagar et al., 2019; Guo et al., 2016; Zhu & Zabarar, 2018; Khoo et al., 2020) and infinite dimensionality (Li et al., 2020; Bhattacharya et al., 2020; Patel et al., 2021). Leading autoregressive methods are Bar-Sinai et al. (2019); Um et al. (2021); Sanchez-Gonzalez et al. (2020); Brandstetter et al. (2022). Then in so-called *neural augmentation* a neural component is added to finite elements (Hsieh et al., 2019), multigrid solvers (Greenfield et al., 2019), and eikonal solvers (Lichtenstein et al., 2019).

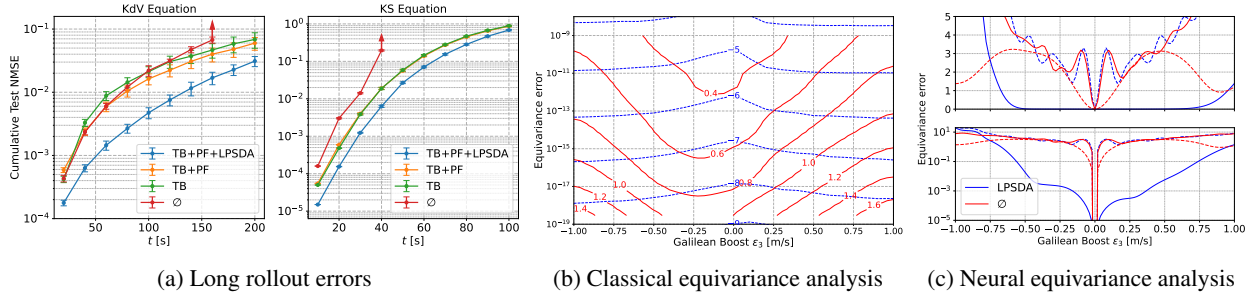


Figure 7. (a) Cumulative error growth over a long autoregressive rollout ($N_{t_{\text{out}}} = 400$ timesteps) on the KdV and KS equations. Uparrows indicate error explosion upon rollout instability. The four incrementally combined settings are: \emptyset , one timestep prediction at a time; TB, temporal bundling (Brandstetter et al., 2022); PF, the pushforward trick; LPSDA, our method. (b) Equivariance error of a single KdV rollout after $T = 100$ s against Galilean boost parameter ϵ_3 . We use $\text{tol} = 1e-12$ as groundtruth. **BLUE DASHED**: Isotherms of log solver tolerance (base 10). Equivariance error decreases with log tolerance, relatively unaffected by transformation magnitude. **RED SOLID**: Isotherms of solution time. Solution time is a function of tolerance and ϵ_3 . Neural solvers instead have constant time overhead. (c) Equivariance error of neural models trained with (LPSDA) and without LPSDA (\emptyset). Bold lines: Equivariance error $\|f(g\mathbf{u}_0) - gf(\mathbf{u}_0)\|_2^2 / \|f(g\mathbf{u}_0)\|_2^2$ (where f is model output). LPSDA is clearly equivariant around the training range $\epsilon_3 \in [-0.4, 0.4]$ and \emptyset is not. Dashed lines: Autocorrelative error $\|f(g\mathbf{u}_0) - f(\mathbf{u}_0)\|_2^2 / \|f(\mathbf{u}_0)\|_2^2$. \emptyset maps $f(g\mathbf{u}_0)$ close to $f(\mathbf{u}_0)$ for small ϵ_3 but LPSDA does not. We suspect \emptyset tries to map unseen ϵ_3 back into a region ($\epsilon_3 = 0$) often encountered in the training set.

6. Limitations and Restrictions

The space of PDEs and their symmetries is vast. We made some experimental design choices to limit the scope of our investigations. Namely, we investigated Eulerian solvers (instead of Lagrangian) on 1D evolution equations on periodic domains. These choices are a restriction of our experimental setup and not of the core idea. From a theory standpoint we expect LSPDA to generalize easily to other settings.

Methodologically, we could only implement symmetries that act globally. Purely local symmetries were not tackled, which would require data augmentation on local patches, something we are yet to develop. Furthermore, while theory provides an exhaustive list of Lie point symmetries for a given PDE, it gives no indication as to the efficacy of each symmetry. We could reason qualitatively about efficacy for, say, space translation or the infinite-dimensional subalgebra g_α , but we could not provide quantitative predictions. Lastly, given a PDE, user input is required to implement the data augmentation functions from a mathematical description of the symmetries $\{g_1, \dots, g_d\}$. Mechanizing this process would make LPSDA far more attractive. This said, we feel LPSDA is a significant step forward in terms of expanding the range of techniques available to train neural PDE solvers.

7. Conclusion and Future work

We showed it is possible to apply the commonplace practice of data augmentation to the training of neural PDE solvers. Here, the space of possible transformations, restricted to Lie point symmetries, can be exhaustively derived from the PDE definition. This places a typically intuition-based aspect of the deep learning user experience on firm mathematical

footing. We demonstrated our claims on three common fluid dynamical PDEs: the Korteweg-de Vries, Kuramoto-Shivashinsky, and Burgers’ equations. Here we showed clear decrease in sample complexity, measured by generalization error across PDEs, models, and training regime. We investigated the effect of individual symmetries and the equivariance properties of solvers, classical and learned.

There are many potential directions from here; symmetries beyond the Lie point symmetries are vast. Examples include *generalized symmetries*, such as *contact symmetries* and *Lie-Bäcklund symmetries* (c.f., Section 2). These form a broader class of transformations, depending on solution derivatives as well as value. For instance, the KdV equation has an infinite number of these symmetries. Then there are discrete symmetries, such as reflections, which can easily be integrated into our framework. If a problem can be cast into variational form, then Noether’s theorem states that every symmetry of the variational form translates into a conserved quantity of the system. This has been exploited by the ML4Physics community with Lagrangian Neural Networks (Greydanus et al., 2019), and with classical solvers this is the basis of the finite volume method and symplectic integrators.

Orthogonally, we could follow the work pioneered by Wang et al. (2021) and bake equivariance into the solver architecture itself. Beyond “simple” symmetries, such as translation or scale (which is actually very difficult), there is little work in this area, apart from Finzi et al. (2020a). In Section 4.7, we saw that classical solvers can exhibit controllable equivariance error. We speculate that merging neural PDE solvers with variable tolerance solvers may be a fruitful area for building equivariance to complicated groups.

8. Acknowledgements

The authors thank Markus Holzleitner for discussions on prolongation theory, and especially for pointing us towards the Cauchy-Kovalevskaya theorem. Johannes Brandstetter acknowledges travel support from ELISE (GA no 951847).

References

- Bar-Sinai, Y., Hoyer, S., Hickey, J., and Brenner, M. P. Learning data-driven discretizations for partial differential equations. *Proceedings of the National Academy of Sciences*, 116(31):15344–15349, Jul 2019. ISSN 1091-6490. doi: 10.1073/pnas.1814058116.
- Baumann, G. Mathlie a program of doing symmetry analysis. *Mathematics and Computers in Simulation*, 48(2):205–223, 1998. ISSN 0378-4754. doi: [https://doi.org/10.1016/S0378-4754\(98\)00143-8](https://doi.org/10.1016/S0378-4754(98)00143-8). URL <https://www.sciencedirect.com/science/article/pii/S0378475498001438>.
- Bhatnagar, S., Afshar, Y., Pan, S., Duraisamy, K., and Kaushik, S. Prediction of aerodynamic flow fields using convolutional neural networks. *Computational Mechanics*, 64(2):525–545, Jun 2019.
- Bhattacharya, K., Hosseini, B., Kovachki, N. B., and Stuart, A. M. Model reduction and neural networks for parametric pdes. *arXiv preprint arXiv:2005.03180*, 2020.
- Boussinesq, M. *Essai sur la théorie des eaux courantes*, volume XXIII. Mémoires présentés par divers savants à l’Académie des Sciences de l’Institut National de France, 1877.
- Brandstetter, J., Worrall, D. E., and Welling, M. Message passing neural pde solvers. *arXiv preprint arXiv:2202.03376*, 2022.
- Bronstein, M. M., Bruna, J., Cohen, T., and Velickovic, P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *CoRR*, abs/2104.13478, 2021. URL <https://arxiv.org/abs/2104.13478>.
- Burgers, J. A mathematical model illustrating the theory of turbulence. volume 1 of *Advances in Applied Mechanics*, pp. 171–199. Elsevier, 1948. doi: [https://doi.org/10.1016/S0065-2156\(08\)70100-5](https://doi.org/10.1016/S0065-2156(08)70100-5). URL <https://www.sciencedirect.com/science/article/pii/S0065215608701005>.
- Clevert, D.-A., Unterthiner, T., and Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). In *International Conference on Learning Representations (ICLR)*, 2016.
- Cole, J. On a quasi-linear parabolic equation occurring in aerodynamics. *Quarterly of Applied Mathematics*, 9(3): 225–236, 1951.
- Courant, R., Friedrichs, K. O., and Lewy, H. On the partial difference equations of mathematical physics. *IBM J. Res. Dev.*, 11(2):215–234, 1967. ISSN 0018-8646.
- Cranmer, M. D., Greydanus, S., Hoyer, S., Battaglia, P. W., Spergel, D. N., and Ho, S. Lagrangian neural networks. *CoRR*, abs/2003.04630, 2020. URL <https://arxiv.org/abs/2003.04630>.
- Cvitanović, P., Davidchack, R. L., and Siminos, E. On the state space geometry of the kuramoto-sivashinsky flow in a periodic domain. *SIAM J. Appl. Dyn. Syst.*, 9:1–33, 2010.
- Dorodnitsyn, V. A. Finite difference models entirely inheriting continuous symmetry of original differential equations. *International Journal of Modern Physics C*, 05: 723–734, 1994.
- Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987. ISSN 0370-2693. doi: [https://doi.org/10.1016/0370-2693\(87\)91197-X](https://doi.org/10.1016/0370-2693(87)91197-X). URL <https://www.sciencedirect.com/science/article/pii/037026938791197X>.
- Finzi, M., Stanton, S., Izmailov, P., and Wilson, A. G. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *International Conference on Machine Learning*, pp. 3165–3176. PMLR, 2020a.
- Finzi, M., Wang, K. A., and Wilson, A. G. Simplifying hamiltonian and lagrangian neural networks via explicit constraints. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b.
- Greenfield, D., Galun, M., Basri, R., Yavneh, I., and Kimmel, R. Learning to optimize multigrid PDE solvers. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 2415–2423, 2019.
- Greydanus, S., Dzamba, M., and Yosinski, J. Hamiltonian neural networks. In Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

- Guo, X., Li, W., and Iorio, F. Convolutional neural networks for steady flow approximation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 481–490. Association for Computing Machinery, 2016. doi: 10.1073/pnas.2101784118. URL <https://www.pnas.org/content/118/21/e2101784118>.
- Hairer, E., Nørsett, S., and Wanner, G. *Solving Ordinary Differential Equations I (2nd Revised. Ed.): Nonstiff Problems*. Springer-Verlag, Berlin, Heidelberg, 1993. ISBN 0387566708.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.
- Hendrycks, D. and Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.
- Hoarau, E. and David, C. Lie group computation of finite difference schemes. *Dyn. Contin. Discrete Impuls. Syst. Ser. A Math. Anal.*, 14(Advances in Dynamical Systems, suppl. S2):180–184, 2007. ISSN 1201-3390.
- Hoarau, E., David, C., Sagaut, P., and Lê, T.-H. Lie group study of finite difference schemes. *Discrete Contin. Dyn. Syst.*, (Dynamical systems and differential equations. Proceedings of the 6th AIMS International Conference, suppl.):495–505, 2007. ISSN 1078-0947.
- Hopf, E. The partial differential equation $u_t + uu_x = \mu u_{xx}$. *Communications on Pure and Applied Mathematics*, 3(3): 201–230, 1950.
- Hsieh, J., Zhao, S., Eismann, S., Mirabella, L., and Ermon, S. Learning neural PDE solvers with convergence guarantees. *arXiv preprint arXiv:1906.01200*, 2019.
- Ibragimov, N. H. *CRC Handbook of Lie Group Analysis of Differential Equations, Volume I: Symmetries, Exact Solutions, and Conservation Laws*. CRC Press, paperback edition, 12 1993. ISBN 978-0851996370.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. URL <https://arxiv.org/abs/2001.08361>.
- Khoo, Y., Lu, J., and Ying, L. Solving parametric pde problems with artificial neural networks. *European Journal of Applied Mathematics*, 32(3):421–435, Jul 2020. ISSN 1469-4425.
- Kochkov, D., Smith, J. A., Alieva, A., Wang, Q., Brenner, M. P., and Hoyer, S. Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences*, 118(21), 2021. ISSN 0027-8424.
- Korteweg, D. D. J. and de Vries, D. G. On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 39(240):422–443, 1895.
- Kuramoto, Y. Diffusion-Induced Chaos in Reaction Systems. *Progress of Theoretical Physics Supplement*, 64:346–367, 02 1978.
- Lelièvre, T. and Stoltz, G. Partial differential equations and stochastic methods in molecular dynamics. *Acta Numerica*, 25:681–880, 2016.
- Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.
- Lichtenstein, M., Pai, G., and Kimmel, R. Deep eikonal solvers. *arXiv preprint arXiv:1903.07973*, 2019.
- Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- Neal, R. M. Probabilistic inference using markov chain monte carlo methods. Technical report, 1993.
- Oliveri, F. Lie Symmetries of Differential Equations: Classical Results and Recent Contributions. *Symmetry*, 2(2):658–706, 2010. ISSN 2073-8994. doi: 10.3390/sym2020658. URL <https://www.mdpi.com/2073-8994/2/2/658>.
- Olver, P. Symmetry groups of differential equations. In *Applications of Lie Groups to Differential Equations*, pp. 77–185. Springer, 1986.
- Patel, R. G., Trask, N. A., Wood, M. A., and Cyr, E. C. A physics-informed operator regression framework for extracting data-driven continuum models. *Computer Methods in Applied Mechanics and Engineering*, 373:113500, Jan 2021. ISSN 0045-7825.
- Quarteroni, A. and Quarteroni, S. *Numerical models for differential problems*, volume 2. Springer, 2009.
- Raissi, M. Deep hidden physics models: Deep learning of nonlinear partial differential equations. *J. Mach. Learn. Res.*, 19:25:1–25:24, 2018.
- Sanchez-Gonzalez, A., Bapst, V., Cranmer, K., and Battaglia, P. W. Hamiltonian graph networks with ODE integrators. *CoRR*, abs/1909.12790, 2019. URL <http://arxiv.org/abs/1909.12790>.

- Sanchez-Gonzalez, A., Godwin, J., Pfaff, T., Ying, R., Leskovec, J., and Battaglia, P. W. Learning to simulate complex physics with graph networks. *CoRR*, abs/2002.09405, 2020. URL <https://arxiv.org/abs/2002.09405>.
- Schilders, W. *Introduction to Model Order Reduction*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-78841-6. doi: 10.1007/978-3-540-78841-6.1. URL https://doi.org/10.1007/978-3-540-78841-6_1.
- Shivashinsky, G. Nonlinear analysis of hydrodynamic instability in laminar flames—i. derivation of basic equations. *Acta Astronautica*, 4(11):1177–1206, 1977.
- Simard, P., Steinkraus, D., and Platt, J. Best practices for convolutional neural networks applied to visual document analysis. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, pp. 958–963, 2003. doi: 10.1109/ICDAR.2003.1227801.
- Sirignano, J. and Spiliopoulos, K. Dgm: A deep learning algorithm for solving partial differential equations. *Journal of Computational Physics*, 375:1339 – 1364, 2018. ISSN 0021-9991. doi: <https://doi.org/10.1016/j.jcp.2018.08.029>.
- Thuerey, N., Holl, P., Mueller, M., Schnell, P., Trost, F., and Um, K. *Physics-based Deep Learning*. WWW, 2021. URL <https://physicsbaseddeeplearning.org>.
- Um, K., Brand, R., Yun, Fei, Holl, P., and Thuerey, N. Solver-in-the-loop: Learning from differentiable physics to interact with iterative pde-solvers, 2021.
- Valiquette, F. and Winternitz, P. Discretization of partial differential equations preserving their physical symmetries. *Journal of Physics A: Mathematical and General*, 38(45):9765–9783, oct 2005. doi: 10.1088/0305-4470/38/45/004. URL <https://doi.org/10.1088/0305-4470/38/45/004>.
- Wang, R., Walters, R., and Yu, R. Incorporating symmetry into deep dynamics models for improved generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=wta_8Hx2KD.
- Zabusky, N. J. and Kruskal, M. D. Interaction of “solitons” in a collisionless plasma and the recurrence of initial states. *Phys. Rev. Lett.*, 15:240–243, Aug 1965. doi: 10.1103/PhysRevLett.15.240. URL <https://link.aps.org/doi/10.1103/PhysRevLett.15.240>.

A. Training data

Korteweg-de Vries and Kuramoto-Shivashinsky equation. Numerical groundtruth data for the Korteweg-de Vries (KdV (Boussinesq, 1877, p. 360)(Korteweg & de Vries, 1895) and the Kuramoto-Shivashinsky (KS) (Kuramoto, 1978; Shivashinsky, 1977) equation is obtained on a periodic domain using the *method of lines*, with the spatial derivatives computed using the pseudo-spectral method. The *method of lines* discretizes the domain \mathcal{X} on a (regular) mesh at a finite number of points $X = \{x_i \in \mathcal{X}\}$, where $|X| < \infty$. We can then represent the signal $\{u(t, x) | x \in X, t \in [0, T]\}$ in terms of a vector $\mathbf{u}(t)$, where the i^{th} element of the vector is the i^{th} element in X (assuming we have imposed some sensible ordering on X). We then solve

$$\partial_t \mathbf{u} = f(t, \mathbf{u}) \quad t \in [0, T], \quad (21)$$

where f is a numerical implementation of all the spatial derivatives. This is just a set of coupled ODEs. Typically we also discretize the time domain $[0, T]$ as well to yield a forward time-stepping formulation. In the pseudospectral method, the derivatives are computed in the frequency domain by first applying a fast Fourier transform (FFT) to the data, then multiplying by the appropriate values and converting back to the spatial domain with the inverse FFT. This method of differentiation is implemented by the `diff` function in the module `scipy.fftpack`. For integration in time we use an implicit Runge-Kutta method of Radau IIA family, order 5 (Hairer et al., 1993).

As a crosscheck, we use a Finite Volume method (FVM), which is a valid approach since the equations at hand are in conservation form, i.e. it can be shown via the divergence theorem that the integral of \mathbf{u} over cell i increases only by the net flux into the cell. We therefore estimate the flux at the left and right cell boundary at time t_k , which we again do via the pseudospectral method of the module `scipy.fftpack`.

Figure 8 shows visual comparisons of the described pseudospectral and FVM pseudospectral data generation. Mean numerical mean-squared errors are in the order of 10^{-5} .

Burgers’ equation. Numerical groundtruth data for the Burgers’ equation (Burgers, 1948)

$$u_t + uu_x - \nu u_{xx} = 0, \quad (22)$$

is obtained by solving the Heat equation for viscosity $\nu = 0.01$. This is possible due to the Cole-Hopf transformation (Cole, 1951; Hopf, 1950)

$$u = 2\nu \frac{\partial}{\partial x} \log(\phi), \quad (23)$$

which turns the Burgers’ equation into the Heat equation:

$$\phi_t = \nu \phi_{xx}. \quad (24)$$

The Heat equation is solved in the same manner as the KdV and the KS equations, initial conditions for the Heat equation are obtained by taking initial conditions of the Burgers’ equation and applying an inverse Cole-Hopf transformation. Finally, applying the inverse Cole-Hopf transformation on the solved trajectory yields the Burgers’ equation. Figure 9 shows exemplary Heat equation and respective Burgers’ equation trajectories. The trajectories *Burgers’ 1+2* are of special interest since they are obtained via non-linear combination of the trajectories *Burgers’ 1* and *Burgers’ 2*, which is exactly what the generator g_α in Table 1 stands for.

B. Interpolation scheme

In our method we use trigonometric interpolation to shift/rotate solutions in the x -direction. This makes use of the Fourier shift theorem: translating a function $u(x) \mapsto u(x - \epsilon)$ on periodic domain of length L linearly shifts its Fourier transform as $\hat{u}(k) \mapsto e^{-i2\pi k\epsilon/L} \hat{u}(k)$. We can thus translate a function by ϵ by first taking a Fourier transform \mathcal{F} , shifting the phase and then taking the inverse Fourier transform \mathcal{F}^{-1} :

$$u(x) \xrightarrow{\mathcal{F}} \hat{u}(k) \xrightarrow{\text{shift}} e^{-i2\pi k\epsilon/L} \hat{u}(k) \xrightarrow{\mathcal{F}^{-1}} u(x - \epsilon), \quad (25)$$

We originally chose this scheme based on standard signal processing considerations. Furthermore, it seemed a sensible choice, given that we use pseudospectral methods to compute the instantaneous spatial derivatives in the classical solver.

We benchmarked the performance of Fourier interpolation against linear interpolation after experiments as a cross-check. The results are in Figure 10. Surprisingly, the results show that whether we interpolate data using trigonometric or linear interpolation the effect on test NMSE is not statistically significant, contrary to what we primarily had suspected. This result does not invalidate our experiments. We suspect that the neural models are able to undo interpolation artifacts introduced by the linear interpolation scheme. This is potentially because the neural models take in not just an initial condition at a single time, but a trajectory of the previous N_{in} steps, across which they can average out interpolation artifacts.

C. Experimental Architectures

We experimented with two models: a baseline 1D ResNet-like model (He et al., 2016) and the state-of-the-art Fourier Neural Operator (FNO) network (Li et al., 2020). Both models take K history timesteps as input and are trained in an autoregressive way and as neural operator (see Figure 3). The input therefore has the dimension $[\text{batch}, K, c]$ where K is interpreted as channel dimension. The channel dimension is mapped to 128 and 256 dimension for the 1D ResNet and

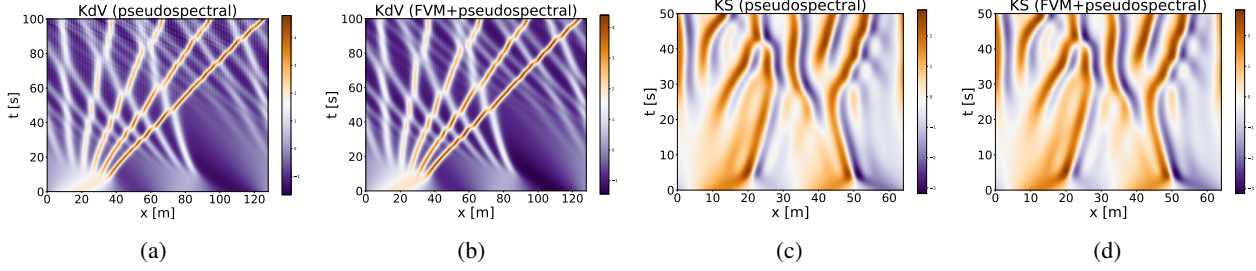


Figure 8. Comparison of pseudospectral and Finite Volume Method (FVM) pseudospectral data generation for the Korteweg de Vries (a,b) and the Kuramoto-Shivashinsky (c,d) equation. Mean numerical mean-squared errors are in the order 10^{-5} .

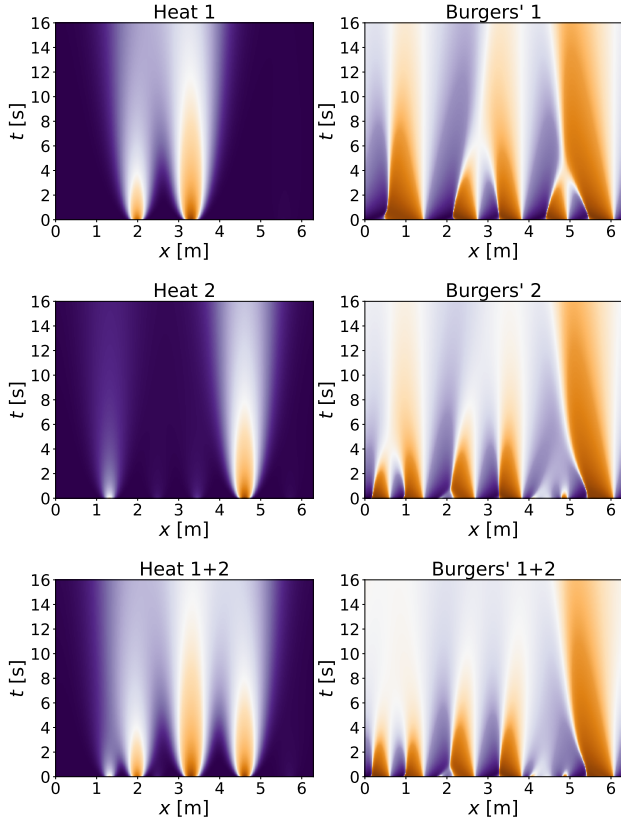


Figure 9. Comparison of exemplary trajectories of the Heat and the Burgers' equation. The Burgers' equation is obtained out of the Heat equation via the inverse Cole-Hopf transformation.

the FNO, respectively. The 1D ResNet has 8 convolutional layers using residual connections and increasing receptive fields (kernel size 3 to kernel size 15) and ELU (Clevert et al., 2016) non-linearities in between. The FNO architectures use 5 1D FNO layers as proposed in Li et al. (2020) with 32 Fourier modes. These layers have residual connections and are intertwined with GeLU (Hendrycks & Gimpel, 2016) non-linearities. We also test a 2D FNO architecture as alternatively described in Li et al. (2020), but the 1D FNO worked consistently better. For both models, the output

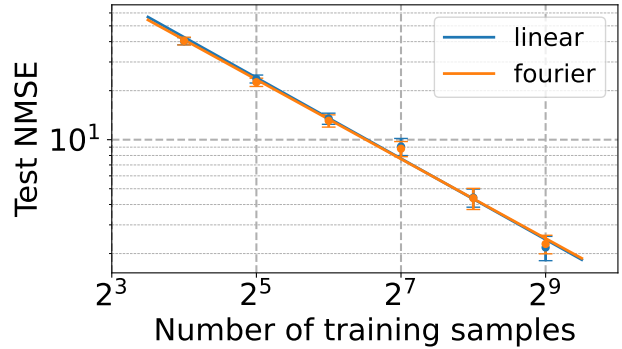


Figure 10. Comparison of the linear and Fourier interpolations schemes for different training set sizes, over the KdV equation, using the setup from Section 4.2. We see no discernable difference in performance.

is mapped to K channels in the autoregressive setting and to the number of desired timesteps in the neural operator setting.

A note on Fourier Neural Operators. The way we interpret the Fourier Neural Operator (FNO) is via its different treatment of the spatial and temporal informations. FNO layers (Li et al., 2020) consist of two parts, which are convolutions in the Fourier domain and convolutions with kernel size equals one. Whereas the former focus on processing spatial information, the latter process temporal information if temporal information is provided in the channel dimension. We therefore merge the temporal bundling idea of Brandstetter et al. (2022) into the FNO framework, where K timesteps are input to the model. Consequently, the temporal information processing is strongly enhanced. This is not only results in improved runtimes by a factor K for rollouts, but also in better stability and decreased errors, see e.g. Figure 6 and Tables 5 and 6.

Training details. We optimize models using the AdamW optimizer (Loshchilov & Hutter, 2017) with learning rate $1e-4$, weight decay $1e-8$ for 20 epochs and minimize the

normalized mean squared error (NMSE) which is outlined in Equation 20. The overall used FNO architectures consist of roughly 1 million parameters and training for the different experiments takes between 12 and 24 hours on average on a GeForceRTX 2080 Ti GPU.

D. Results

Table 3 and Table 4 contain a comprehensive summary of results presented in Figures 4a, 4b, 4c, 4d, i.e. an extensive test of Lie point symmetry data augmentation (LPSDA) on the Korteweg-de Vries (KdV) and the Kuramoto-Shivashinsky (KS) equation. Table 5 and Table 6 contain a comprehensive summary of KdV and KS long rollout results, discussed in Section 4.6 in the main paper.

Lie Point Symmetry Data Augmentation for Neural PDE Solvers

Table 3. Test of Lie point symmetry data augmentation (LPSDA) on the Korteweg-de Vries (KdV) and the Kuramoto-Shivashinsky (KS) equation. Averaged normalized MSE (NMSE) errors on the test set are reported, $\mathcal{L}_{\text{NMSE}} \frac{1}{N_t} \sum_{j=1}^{N_t} \frac{\|\mathbf{u}(t_j) - \hat{\mathbf{u}}(t_j)\|_2^2}{\|\hat{\mathbf{u}}(t_j)\|_2^2}$. Autoregressive (AR) and neural operator (NO) prediction errors are compared. A 1D ResNet-like and a Fourier Neural Operator Li et al. (2020) (FNO) solver are compared. Best performing symmetry groups are highlighted. Different solvers are tested on different combination of Lie point symmetries. Experiments show that performance gets consistently better for additional generators. This is verified for different solvers, different equations, different tasks, and different combination of generators.

Task	Solver	Sym.	Number of samples					
			512	256	128	64	32	16
KdV (20s)	FNO (AR)	-	.0030 ± .0010	.0058 ± .0010	.0225 ± .0024	.0604 ± .0036	.1328 ± .0093	.1901 ± .0105
KdV (20s)	FNO (AR)	$g_1 g_2 g_3 g_4$.0010 ± .0001	.0037 ± .0006	.0119 ± .0013	.0387 ± .0025	.0959 ± .0082	.1399 ± .0065
KdV (20s)	FNO (NO)	-	.0276 ± .0038	.0407 ± .0037	.0858 ± .0056	.1699 ± .0100	.2717 ± .0146	.4887 ± .0156
KdV (20s)	FNO (NO)	$g_1 g_2 g_3 g_4$.0055 ± .0007	.0132 ± .0013	.0302 ± .0021	.0574 ± .0045	.1289 ± .0084	.2376 ± .0136
KdV (20s)	ResNet (NO)	-	.0160 ± .0014	.0273 ± .0014	.0443 ± .0026	.0574 ± .0030	.1021 ± .0056	.1178 ± .0064
KdV (20s)	ResNet (NO)	$g_1 g_2$.0161 ± .0014	.0231 ± .0023	.0291 ± .0031	.0434 ± .0025	.0896 ± .0072	.1239 ± .0095
KdV (20s)	ResNet (NO)	$g_1 g_2 g_3$.0168 ± .0014	.0209 ± .0011	.0310 ± .0024	.0426 ± .0021	.0824 ± .0049	.0990 ± .0037
KdV (20s)	ResNet (NO)	$g_1 g_2 g_4$.0121 ± .0013	.0204 ± .0016	.0276 ± .0021	.0474 ± .0021	.1314 ± .0077	.1188 ± .0054
KdV (20s)	ResNet (NO)	$g_1 g_2 g_3 g_4$.0118 ± .0010	.0205 ± .0016	.0277 ± .0020	.0407 ± .0026	.0774 ± .0048	.0746 ± .0036
KdV (20s)	ResNet (AR)	-	.0223 ± .0026	.0392 ± .0031	.0669 ± .0050	.0914 ± .0087	.1701 ± .0110	.2087 ± .0121
KdV (20s)	ResNet (AR)	$g_1 g_2$.0200 ± .0021	.0284 ± .0022	.0470 ± .0031	.0782 ± .0049	.1463 ± .0099	.1911 ± .0110
KdV (20s)	ResNet (AR)	$g_1 g_2 g_3$.0111 ± .0010	.0185 ± .0013	.0318 ± .0022	.0517 ± .0029	.0883 ± .0068	.1145 ± .0069
KdV (20s)	ResNet (AR)	$g_1 g_2 g_4$.0155 ± .0019	.0269 ± .0026	.0445 ± .0028	.0763 ± .0045	.1602 ± .0104	.1887 ± .0088
KdV (20s)	ResNet (AR)	$g_1 g_2 g_3 g_4$.0113 ± .0012	.0184 ± .0016	.0333 ± .0020	.0544 ± .0030	.0965 ± .0064	.1106 ± .0052
KdV (40s)	FNO (AR)	-	.0081 ± .0026	.0361 ± .0062	.0681 ± .0084	.1248 ± .0108	.1981 ± .0167	.2904 ± .0184
KdV (40s)	FNO (AR)	g_1	.0023 ± .0004	.0082 ± .0010	.0293 ± .0024	.0674 ± .0055	.1294 ± .0081	.2224 ± .0123
KdV (40s)	FNO (AR)	$g_1 g_2$.0016 ± .0008	.0088 ± .0017	.0278 ± .0045	.0673 ± .0058	.1251 ± .0106	.2160 ± .0135
KdV (40s)	FNO (AR)	$g_1 g_2 g_3$.0011 ± .0005	.0062 ± .0012	.0229 ± .0035	.0550 ± .0058	.0940 ± .0079	.1818 ± .0153
KdV (40s)	FNO (AR)	$g_1 g_2 g_4$.0017 ± .0009	.0102 ± .0016	.0287 ± .0048	.0795 ± .0077	.1271 ± .0096	.2092 ± .0138
KdV (40s)	FNO (AR)	$g_1 g_2 g_3 g_4$.0010 ± .0006	.0058 ± .0013	.0213 ± .0029	.0606 ± .0040	.0968 ± .0091	.1817 ± .0118
KdV (40s)	FNO (NO)	-	.1113 ± .0100	.1761 ± .0122	.2343 ± .0194	.3331 ± .0185	.4892 ± .0237	.5954 ± .0301
KdV (40s)	FNO (NO)	g_1	.0735 ± .0040	.1162 ± .0053	.1723 ± .0083	.2350 ± .0076	.3674 ± .0149	.4627 ± .0112
KdV (40s)	FNO (NO)	$g_1 g_2$.0761 ± .0086	.1140 ± .0087	.1622 ± .0121	.2494 ± .0158	.3432 ± .0232	.4884 ± .0254
KdV (40s)	FNO (NO)	$g_1 g_2 g_3$.0279 ± .0038	.0462 ± .0052	.0868 ± .0072	.1170 ± .0111	.1842 ± .0130	.3001 ± .0169
KdV (40s)	FNO (NO)	$g_1 g_2 g_4$.0779 ± .0086	.1140 ± .0098	.1757 ± .0130	.2534 ± .0139	.3731 ± .0207	.4764 ± .0204
KdV (40s)	FNO (NO)	$g_1 g_2 g_3 g_4$.0273 ± .0040	.0478 ± .0052	.0873 ± .0070	.1127 ± .0101	.1743 ± .0118	.3037 ± .0190
KS (20s)	FNO (AR)	-	.0279 ± .0013	.1084 ± .0048	.1983 ± .0073	.2898 ± .0080	.4234 ± .0106	.5279 ± .0110
KS (20s)	FNO (AR)	g_1	.0045 ± .0004	.0421 ± .0022	.1347 ± .0045	.2241 ± .0067	.3690 ± .0086	.4984 ± .0143
KS (20s)	FNO (AR)	$g_1 g_2$.0057 ± .0004	.0581 ± .0020	.1321 ± .0056	.2316 ± .0085	.3605 ± .0109	.4945 ± .0122
KS (20s)	FNO (AR)	$g_1 g_2 g_3$.0028 ± .0002	.0261 ± .0012	.0755 ± .0035	.1504 ± .0064	.2656 ± .0074	.4398 ± .0133
KS (20s)	FNO (NO)	-	.3905 ± .0049	.4733 ± .0051	.5662 ± .0051	.6855 ± .0097	.8502 ± .0064	1.0005 ± .0076
KS (20s)	FNO (NO)	g_1	.2936 ± .0044	.3863 ± .0050	.4813 ± .0050	.5884 ± .0057	.7438 ± .0054	.8426 ± .0060
KS (20s)	FNO (NO)	$g_1 g_2$.2948 ± .0047	.3864 ± .0043	.4854 ± .0043	.5810 ± .0061	.7388 ± .0048	.8359 ± .0070
KS (20s)	FNO (NO)	$g_1 g_2 g_3$.2268 ± .0049	.3089 ± .0066	.3981 ± .0075	.4501 ± .0069	.5321 ± .0038	.6045 ± .0057

Table 4. Test of Lie point symmetry data augmentation (LPSDA) on the Burgers' equation. Averaged normalized MSE (NMSE) errors on the test set are reported, $\mathcal{L}_{\text{NMSE}} \frac{\|\mathbf{u}(t_j) - \hat{\mathbf{u}}(t_j)\|_2^2}{\|\hat{\mathbf{u}}(t_j)\|_2^2}$ is omitted due to the small values. Autoregressive (AR) and neural operator (NO) prediction errors are compared. Fourier Neural Operator Li et al. (2020) (FNO) solver are tested on different combination of Lie point symmetries. Best performing symmetry groups are highlighted. Experiments show that performance gets consistently better for additional generators.

Task	Solver	Sym.	Number of samples					
			512	256	128	64	32	16
Burgers' (10s)	FNO (AR)	-	.0020 ± .0002	.0091 ± .0008	.0696 ± .0057	.3442 ± .0226	.7962 ± .0358	1.6493 ± .0581
Burgers' (10s)	FNO (AR)	g_1	.0010 ± .0001	.0061 ± .0006	.0455 ± .0034	.2767 ± .0189	.8489 ± .0461	2.5599 ± .0954
Burgers' (10s)	FNO (AR)	$g_1 g_5$.0003 ± .0001	.0010 ± .0002	.0063 ± .0014	.0228 ± .0031	.0656 ± .0104	.3729 ± .0303
Burgers' (10s)	FNO (NO)	-	.0428 ± .0049	.1145 ± .0110	.2792 ± .0199	.5692 ± .0306	.7959 ± .0409	1.8415 ± .0645
Burgers' (10s)	FNO (NO)	g_1	.0092 ± .0014	.0566 ± .0050	.2723 ± .0156	.6092 ± .0290	.9784 ± .0445	2.3453 ± .0663
Burgers' (10s)	FNO (NO)	$g_1 g_5$.0003 ± .0001	.0009 ± .0001	.0066 ± .0015	.0322 ± .0043	.1436 ± .0112	.5511 ± .0281

Table 5. Long rollouts of $T = 200$ s with $N_{t_{out}} = 400$ timesteps on the Korteweg-de Vries (KdV) equation to check how LPSDA can improve performance in a practical use case. The cumulative error normalized over seconds is reported. We use an autoregressively-trained FNO model. Additionally, the training procedure of the solvers is augmented with temporal bundling (TB) and the pushforward trick (PF) as introduced in Brandstetter et al. (2022). For temporal bundling variants we input and output 20 timesteps at a time, which we compare to one timestep predictions (1-step). All training tricks are ablated. Best performing symmetry groups are highlighted.

Timesteps	Training methods			
	1-step	TB	TB+PF	TB+PF+LPSDA
20s	.0008 ± .0005	.0008 ± .0005	.0011 ± .0005	.0003 ± .0001
40s	.0046 ± .0011	.0065 ± .0021	.0048 ± .0016	.0012 ± .0003
60s	.0120 ± .0021	.0176 ± .0046	.0119 ± .0034	.0028 ± .0009
80s	.0242 ± .0037	.0287 ± .0065	.0209 ± .0052	.0053 ± .0013
100s	.0441 ± .0084	.0423 ± .0083	.0326 ± .0064	.0094 ± .0018
120s	.0628 ± .0032	.0610 ± .0110	.0446 ± .0080	.0151 ± .0024
140s	.0946 ± .0074	.0744 ± .0125	.0616 ± .0102	.0231 ± .0040
160s	.1345 ± .0093	.0932 ± .0150	.0801 ± .0123	.0338 ± .0047
180s	∞	.1167 ± .0181	.0959 ± .0137	.0454 ± .0048
200s	∞	.1369 ± .0188	.1198 ± .0136	.0621 ± .0056

Table 6. Long rollouts of $T = 100$ s with $N_{t_{out}} = 400$ timesteps on the Kuramoto-Shivashinsky (KS) equation to check how LPSDA can improve performance in a practical use case. The cumulative error normalized over seconds is reported. We use an autoregressively-trained FNO model. Additionally, the training procedure of the solvers is augmented with temporal bundling (TB) and the pushforward trick (PF) as introduced in Brandstetter et al. (2022). For temporal bundling variants we input and output 20 timesteps at a time, which we compare to one timestep predictions (1-step). All training tricks are ablated. Best performing symmetry groups are highlighted.

Timesteps	Training methods				
	1-step	TB	TB+PF	TB+PF+LPSDA	TB+LPSDA
10s	.0003 ± .0001	.0001 ± .0000	.0001 ± .0000	.0000 ± .0000	.0000 ± .0000
20s	.0060 ± .0010	.0009 ± .0001	.0011 ± .0001	.0003 ± .0000	.0002 ± .0000
30s	.0283 ± .0040	.0077 ± .0013	.0080 ± .0013	.0024 ± .0002	.0017 ± .0002
40s	.0391 ± .0359	.0377 ± .0054	.0381 ± .0060	.0126 ± .0021	.0101 ± .0019
50s	∞	.1166 ± .0128	.1126 ± .0138	.0536 ± .0051	.0419 ± .0049
60s	∞	.2914 ± .0222	.2793 ± .0224	.1415 ± .0127	.1278 ± .0110
70s	∞	.5604 ± .0273	.5459 ± .0283	.3088 ± .0188	.2853 ± .0190
80s	∞	.9387 ± .4007	.9011 ± .0393	.5668 ± .0256	.5241 ± .0264
90s	∞	1.3422 ± .0523	1.3049 ± .0598	.9369 ± .0392	.8736 ± .0390
100s	∞	1.7821 ± .0564	1.7431 ± .0587	1.3692 ± .0572	1.2741 ± .0570