# Neural Wave Machines: Learning Spatiotemporally Structured Representations with Locally Coupled Oscillatory Recurrent Neural Networks

**T. Anderson Keller** [1] **Max Welling** [1]

## Abstract

Traveling waves have been measured at a diversity of regions and scales in the brain, however a consensus as to their computational purpose has yet to be reached. An intriguing hypothesis is that traveling waves serve to structure neural representations both in space and time, thereby acting as an inductive bias towards natural data. In this work, we investigate this hypothesis by introducing the Neural Wave Machine (NWM) – a locally coupled oscillatory recurrent neural network capable of exhibiting traveling waves in its hidden state. After training on simple dynamic sequences, we show that this model indeed learns static spatial structure such as topographic organization, and further uses complex spatiotemporal structure such as traveling waves to encode observed transformations. To measure the computational implications of this structure, we use a suite of sequence classification and physical dynamics modeling tasks to show that the NWM is both more parameter efficient, and is able to forecast future trajectories of simple physical dynamical systems more accurately than existing state of the art counterparts. We conclude with a discussion of how this model may allow for novel investigations of the computational hypotheses surrounding traveling waves which were previously challenging or impossible.

## 1. Introduction

In machine learning, inductive biases can be understood as limiting the search space of possible hypotheses a priori, and indeed, it is known that without any inductive bias, learning generalizations beyond the training data is theoretically impossible (Wolpert, 1996). Modern machine learning researchers have adopted many task-specific inductive biases almost by default, such as convolution for spatially structured data. Similarly, natural intelligence as implemented by biological systems also has many inductive biases by virtue of the diversity of constraints that it must simultaneously satisfy such as metabolic efficiency. The fields of psychology, cognitive science, and neuroscience have all studied these biases and their observed signatures, often hypothesizing about their computational implications.

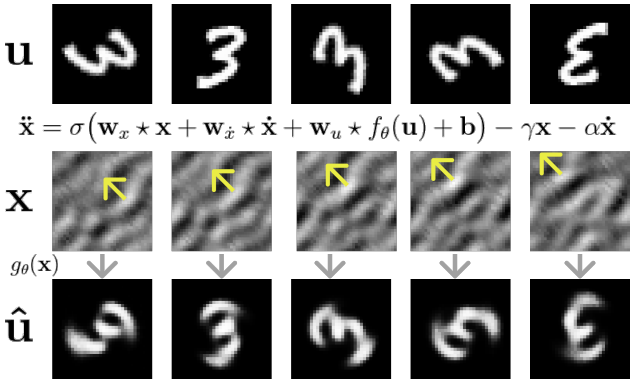


Figure 1: Overview of the Neural Wave Machine. The input sequence $\mathbf{u}$ is encoded with $f_\theta$ to act as a driving term in the hidden state $\mathbf{x}$ which is modeled temporally ($\ddot{\mathbf{x}}$) as a network of locally coupled oscillators. The network is then trained to reconstruct the input sequence: $\hat{\mathbf{u}} = g_\theta(\mathbf{x})$. The yellow arrows track a traveling wavefront over time.

One such observation which has recently gained increasing interest in the neuroscience community is that of traveling waves of neural activity. Such waves have been measured at both local (Davis et al., 2020) and global (Muller et al., 2016b) scales, and have been shown to be strongly related to alpha, theta, and gamma oscillations in a variety of brain regions (Zhang et al., 2018; Besserve et al., 2015). Prompted by these observations, a large number of theoretical hypotheses have been developed which attempt to explain the computational purposes of traveling waves (Muller et al., 2018), and the inductive biases which they may mediate.

Of particular relevance to the machine learning community, one hypothesis is that traveling waves serve to bene-

[1]UvA Bosch Delta Lab, University of Amsterdam, Amsterdam, Netherlands. Correspondence to: T. Anderson Keller <t.anderson.keller@gmail.com>.

*Proceedings of the $40^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. 

ficially structure neural representations in both space and time (Lubenov & Siapas, 2009; Jancke et al., 2004), acting as an inductive bias towards similarly structured natural data. Structured representations have been previously demonstrated in the machine learning community to be extremely valuable, making learning models both more efficient and robust (Worrall et al., 2017). A prime example of this is group equivariance (Cohen & Welling, 2016); in the case of translation, this resulted in the convolutional neural network which reduced the sensitivity of existing fully connected artificial neural networks to small image shifts and deformations (Fukushima, 1980; Lecun et al., 1998), thereby facilitating the rapid growth of the field of deep learning (Krizhevsky et al., 2012). In the case of traveling waves, it is thus suggested that they may facilitate a similar kind of spatiotemporal structure in neural representations, thereby granting the observed robustness and efficiency of natural intelligence which is still lacking in modern deep neural networks (Lake et al., 2017).

To date, however, testing ideas related to the computational purposes of traveling waves has been challenging due to a lack of neural network architectures which have a notion of spatial locality necessary for modeling such spatio-temporal dynamics. Further, existing networks which do have such spatial structure often do not have temporal structure (Keller et al., 2021; Lee et al., 2020), or are not sufficiently flexibly parameterized to allow them to be trained on standard machine learning benchmarks (Davis et al., 2021).

In this work, we propose to investigate the computational hypotheses surrounding traveling waves through a bottom-up approach; we build a flexibly parameterized computational model known to be capable of producing traveling waves, and show that it indeed learns to exhibit complex spatiotemporal dynamics when modeling real data. We then show, relevant to the computational neuroscience community, how such a network indeed learns spatial and temporal structure reminiscent of that found in the brain. Specifically, we observe that our network learns topographically organized selectivity, similar to the observed orientation columns and hypercolumns of the primary visual cortex (Wiesel & Hubel, 1974). Further, we show that our network learns to use complex spatiotemporal organization such as traveling waves to encode transformations by artificially inducing waves in the hidden state and observing that this allows us to further progress or reverse the transformations of generated images.

As it relates to inductive biases, we asses the computational implications of the observed representational structure by training the model on the physical dynamics forecasting suite introduced in the paper 'Which Priors Matter?' (Botev et al., 2021). We see that our model is more accurate at predicting future trajectories of simple physical dynamics when compared with existing state of the art models, providing evidence that the structure mediated by traveling waves is indeed a beneficial inductive bias for modeling such smooth natural transformations. Further, due to our model's local connectivity, we see that it is more efficient both in terms of parameters, and in terms of biological concerns such as wiring length, suggesting a connection between locality of connections, waves, and an inductive transformation bias in biological systems.

Overall we believe our work offers the concrete contribution of a new powerful model at the interface of computational neuroscience and modern machine learning. We show that this model allows for the investigation of the computational hypotheses surrounding complex synchrony in the brain in a new way, and further provides preliminary evidence for the existing hypothesis that traveling waves serve to induce spatiotemporal structure in neural representations.

## 2. Background

**Structured Representations** In machine learning, an increasingly popular way to incorporate structure into neural networks is through equivariant architectures such as group equivariant convolutional neural networks (Cohen & Welling, 2016). Formally, a map $f$ is equivariant if it commutes with the transformation: $f(\tau_\rho[\mathbf{u}]) = \Gamma_\rho[f(\mathbf{u})]$, where $\tau$ and $\Gamma$ are the representations of the action of the group element $\rho$ on the input and output spaces respectively. At a high level, this can be understood to mean that for a given set of input transformations of interest, there is a corresponding known and well-behaved transformation of the representations in output space. One of the simplest and most well known examples of an equivariant map is the convolutional layer; a translation of the input results in a corresponding translation of the output feature maps. Such models have been observed to improve sample efficiency, generalization, and robustness both empirically (Fukushima, 1980; Cohen & Welling, 2016; Worrall et al., 2017; Veeling et al., 2018; van der Pol et al., 2020) and theoretically (Elesedy & Zaidi, 2021; Bordelon & Pehlevan, 2022) by serving as an inductive bias towards representations with naturally realistic symmetry. Despite their efficacy, however, their application to more complex non-group transformations has been limited by the restrictions of the underlying group theory. A recently developing research goal has thus been to build equivariant maps for a broader range of transformations, including models which aim to learn symmetries from the data itself (van der Wilk et al., 2018; Bouchacourt et al., 2021; Keller & Welling, 2021).

**Traveling Waves in Neuroscience** Neural oscillations and traveling waves have long been a subject of study in neuroscience and neurophysiology (Hughes, 1995; Muller et al., 2018). Although such waves were originally measured

primarily in anesthetized subjects, improved multi-channel recording and analysis techniques have recently demonstrated propagating wave activity in awake functioning subjects as well, originating from both external stimuli and internal 'spontaneous' recurrent connections (Sato et al., 2012; Muller et al., 2014; 2018). While many hypotheses have been put forth for their precise computational role, a consensus has yet to be reached. Example hypotheses include that traveling waves may: influence visual perception (Zanos et al., 2015); modulate information transfer (Besserve et al., 2015); correlate with conscious awareness (Bhattacharya et al., 2022b); facilitate predictive coding (Friston, 2019; Alamia & VanRullen, 2019); lower the threshold for detection of weak stimuli (Davis et al., 2020); serve as a short term memory (King & Wyart, 2021; Bhattacharya et al., 2022a); or as a mechanism for the formation of long-term memories during sleep (Muller et al., 2016a). Relevant to this work, traveling waves have directly been implicated in the encoding of motion (Heitmann & Ermentrout, 2020), and have been measured to correlate strongly with perceived perceptual illusions of motion (Jancke et al., 2004). Further, it has been suggested that they form the basis of alpha and theta oscillations (Zhang et al., 2018; Lubenov & Siapas, 2009) and may serve to both structure and integrate information across space and time (Sato et al., 2012; Sato, 2022). Due to the fundamental relationship between neural synchrony and the coordination of spike timing (Bragin et al., 1995), it is natural to wonder if more complex forms of spatiotemporal synchrony such as traveling waves may play a similarly more complex structural role.

**Computational Models of Traveling Waves** In the fields of computational and theoretical neuroscience, multiple models have been developed to help explain the observed complex synchronous dynamics of neural systems. One classical model is that of a network of locally coupled oscillators (Diamant & Bortoff, 1969; Ermentrout & Kopell, 1984). However, to date, such models have been limited to those which either are built for the primary purpose of analysis (Kuramoto, 1981; Ermentrout & Kleinfeld, 2001; Davis et al., 2021), or those which perform very simple binary operations (Gong & van Leeuwen, 2009; Izhikevich & Hoppensteadt, 2008), with neither set leveraging the flexible computational capabilities of modern deep neural networks. One line of work has aimed to integrate classical Kuramoto models into deep neural networks by directly parameterizing activations in terms of phase values (Ricci et al., 2021), however such models lack a notion of spatial locality, making the existence of spatio-temporal dynamics less concrete. Most recently, Davis et al. (2021) studied a large scale locally connected spiking neural network model, quantifying the conditions necessary for the emergence of traveling waves, and showed such waves appeared to uniquely agree with human cortical traveling waves in a variety of dimensions. However, similar to most existing models in this category, the model is formulated as a spiking neural network thus requiring more sophisticated training mechanisms which are yet to scale to the same performance as deep neural networks (Neftci et al., 2019).

## 3. Neural Wave Machines

In the following section we introduce the Neural Wave Machine (NWM), a deep neural network architecture which exhibits traveling waves and other complex spatiotemporal dynamics in the service of flexible differentiable computation. To achieve this, we take inspiration from the seminal models of traveling waves built as networks of locally coupled oscillators (Ermentrout & Kleinfeld, 2001), and propose to integrate them into a modern deep learning framework by taking advantage of the recently developed coupled oscillatory Recurrent Neural Network (coRNN) of Rusch & Mishra (2021).

### 3.1. Coupled Oscillatory Recurrent Neural Networks

In (Rusch & Mishra, 2021) the authors propose to solve the Exploding and Vanishing Gradient Problem (EVGP) in recurrent neural networks by defining a new recurrent neural network with hidden state dynamics given by the parameterized equations of a system of coupled, damped, and driven oscillators. Explicitly, the hidden state of the recurrent neural network $\mathbf{x}$ is updated by solving the following second order partial differential equation:

$$\ddot{\mathbf{x}} = \sigma\left(\mathbf{W}_x\mathbf{x} + \mathbf{W}_{\dot{x}}\dot{\mathbf{x}} + \mathbf{V}\mathbf{u} + \mathbf{b}\right) - \gamma\mathbf{x} - \alpha\dot{\mathbf{x}} \quad (1)$$

Where $\frac{\partial \mathbf{x}}{\partial t} = \dot{\mathbf{x}}$, $\frac{\partial^2 \mathbf{x}}{\partial t^2} = \ddot{\mathbf{x}}$ are the first and second derivatives of the hidden state with respect to time, and $\mathbf{u}$ denotes the input at each time step. The terms $\mathbf{W}_x\mathbf{x}$, $\mathbf{W}_{\dot{x}}\dot{\mathbf{x}}$, and $\mathbf{V}\mathbf{u}$ can then be interpreted as the coupling, damping, and driving terms respectively. Finally, $\sigma$ is a nonlinear activation function such as the hyperbolic tangent, and $\gamma$ & $\alpha$ are scalar variables which can be fixed or learned in combination with the above matrices. In practice, the above differential equation can be discretized and integrated numerically using an IMEX (implicit-explicit) discretization scheme shown to preserve the desirable bounds of the continuous system. Such a discretization can be achieved by first introducing a 'velocity' variable $\mathbf{v} = \dot{\mathbf{x}}$, turning the second order system into a set of two coupled first order equations:

$$\dot{\mathbf{x}} = \mathbf{v}, \quad \dot{\mathbf{v}} = \sigma\left(\mathbf{W}_x\mathbf{x} + \mathbf{W}_{\dot{x}}\mathbf{v} + \mathbf{V}\mathbf{u} + \mathbf{b}\right) - \gamma\mathbf{x} - \alpha\mathbf{v} \quad (2)$$

Then, for a fixed time step $0 < \Delta t < 1$, the hidden state $\mathbf{x}$ and velocity $\mathbf{v}$ of the RNN at time $t+1$ can be updated as:

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \Delta t(\mathbf{v}_{t+1}) \qquad \mathbf{v}_{t+1} = \mathbf{v}_t + \Delta t(\mathbf{v}'_t) \quad (3)$$

$$\mathbf{v}'_t = \sigma\big(\mathbf{W}_x\mathbf{x}_t + \mathbf{W}_{\dot{x}}\mathbf{v}_t + \mathbf{V}\mathbf{u}_{t+1} + \mathbf{b}\big) - \gamma\mathbf{x}_t - \alpha\mathbf{v}_t \quad (4)$$

This model was theoretically demonstrated to have a bounded gradient and hidden state magnitude under assumptions on the time-step $\Delta t$ and the infinity norm of the coupling parameters. Empirically, such stable gradient dynamics were shown to yield better performance than existing RNNs on tasks with very long time-dependencies.

In relation to our goals, the oscillatory dynamics of the coRNN make it amenable to synchronous activity, unlike most existing deep neural network models, and the stable gradient dynamics make it a powerful and flexibly parameterizable sequence model, unlike existing models of traveling waves based on spiking neural networks. However, given that the hidden state $\mathbf{x}$ is not endowed with any notion of spatial layout, it is still not meaningful to study spatiotemporal dynamics in such a model. In the following subsection we describe how such a spatial layout may be implemented efficiently by replacing the fully connected recurrent coupling matrices $\mathbf{W}_x$ and $\mathbf{W}_{\dot{x}}$ with convolution operations.

### 3.2. Local Connectivity

In (Davis et al., 2021), the authors study a large scale spiking neural network model, quantifying the emergence of traveling waves, and comparing them with waves observed in the human cortex. At a high level, as it is relevant to this work, the study concludes that locally restricted connectivity and distance dependant conduction delays are both necessary and sufficient to produce traveling waves. Further they observe that such waves are fairly robust to the synaptic strengths of their model when given a sufficiently large number of neurons. Given these findings, we hypothesize that the Coupled Oscillitory Recurrent Neural Network may yield traveling waves if similarly constrained.

To impose such constraints we begin by defining an arbitrary topographic layout for the $N$-dimensional hidden state $\mathbf{x}$ in the model. For computational simplicity, we propose to use a regular 1 or 2 dimensional grid, $\mathbf{x}_{1D} \in \mathbb{R}^{C_h \times N}$ or $\mathbf{x}_{2D} \in \mathbb{R}^{C_h \times \sqrt{N} \times \sqrt{N}}$ respectively, where $C_h$ is the number of simultaneous 'channels' in our hidden state. We then see that specifically, if the recurrent connections $\mathbf{W}_x$ and $\mathbf{W}_{\dot{x}}$ are made local over our spatial dimensions rather than global, and a distance-dependant time-delay introduced, the aforementioned constraints will be satisfied and the remainder of the properties such as synaptic strength and the precise local distribution of connections will be left up to the model to learn. In practice, we simplify the model by restricting the topographic connectivity of each neuron to its immediately adjacent neighbors in the grid, and define all distances (and thus time-delays) to these neurons to be equal to 1. Such a simplification allows us to efficiently implement the local time-delayed connections with a simple size 3 or $3 \times 3$ convolutional kernel for 1 and 2 dimensional grids respectively. In summary, our model is then given identically as in Equations 3 & 4 but with convolutional layers in place of the dense recurrent matrices. Explicitly, in the 2-dimensional setting, for convolutional kernels $\mathbf{w}_x, \mathbf{w}_{\dot{x}} \in \mathbb{R}^{C_h \times C_h \times 3 \times 3}$, we get:

$$\mathbf{v}'_t = \sigma\left(\mathbf{w}_x \star \mathbf{x}_t + \mathbf{w}_{\dot{x}} \star \mathbf{v}_t + f_\theta(\mathbf{u}_{t+1}) + \mathbf{b}\right) - \gamma \mathbf{x}_t - \alpha \mathbf{v}_t \quad (5)$$

We see we have additionally replaced the linear encoder $\mathbf{V}$ with a function $f_\theta$ which can be a convolutional or 'deconvolutional' neural network, or any other mapping from the input to a spatially organized driving force. Importantly, we see that our imposed local connectivity does not immediately invalidate any of the assumptions required for the theorems of Rusch & Mishra (2021) about mitigating the EVGP since the infinity norm of the weights is unlikely to significantly increase when simply switching from fully to locally-connected matrices. We include the updated bounds and corresponding proofs in Appendix B. In the end, we denote this model the Neural Wave Machine due to its emergent wave-like dynamics, facilitated by both the oscillatory update equations of the coRNN, and the local connectivity constraints of biological models. In the next section we measure these desired spatiotemporal dynamics of the NWM and further study their impact as an inductive bias on computation.

## 4. Experiments

In the following two subsections we provide experiments which demonstrate: first, that our model learns spatiotemporal structure reminiscent of natural observations from neuroscience; and second, that such structure is beneficial to both efficiency and accuracy. We outline our methods briefly below, and more thoroughly in Appendix A.

**Methods** All datasets used in this paper will be considered as unsupervised unless otherwise noted, and thus we will train the model from Section 3 as an autoregressive model. To do this, we add a learned decoder from the hidden state $\mathbf{x}_t$ back to the input at the next timestep $\mathbf{u}_{t+1}$, and train the model with a mean-squared error loss. Explicitly, $\hat{\mathbf{u}}_{t+2} = g_\theta(\mathbf{x}_{t+1})$, and $\mathcal{L} = ||\hat{\mathbf{u}}_{t+2} - \mathbf{u}_{t+2}||_2^2$, where $g_\theta$ is the decoder which can again be a convolutional neural network, or any network which maps from the spatial hidden state back to the input space. For the simple tasks in Section 4.1, and the sequence classification tasks of Section 4.2 we use minimal encoders and decoders corresponding to single linear layers or small MLPs. For the more complex physical forecasting tasks of Section 4.2 we use the baseline deep convolutional encoders and decoders defined in the benchmark. As a second minor addition which we observe improves performance on long-term trajectory modeling tasks, we introduce an additional encoder network which learns to predict the initial conditions $\mathbf{x}_0$ and $\mathbf{v}_0$ of the network given a partial 'inference' sequence. Explicitly, we can write this as: $\mathbf{x}_0, \mathbf{v}_0 = f_\theta^{IC}(\{\mathbf{u_t}\}_{t=0}^{T_{inf}})$. Such an initial-condition network is common in the Neural-ODE literature (Chen et al.,

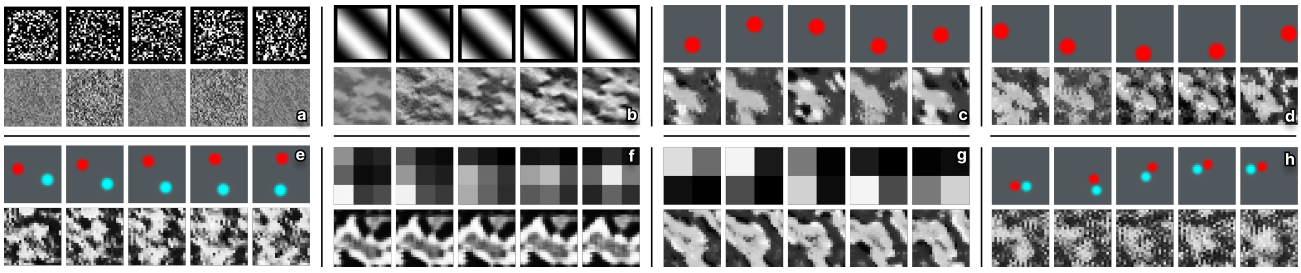


Figure 2: Plot of different datasets used in this work (top) and the associated learned hidden state dynamics (bottom). We see the NWM learns different spatiotemporal structure for each dataset, and no structure when trained on random noise (a). Additional videos of dynamics, and code for experiments, can be found at: github.com/akandykeller/NeuralWaveMachines.

2018), and in this setting it is beneficial to initialize the latent dynamics which would otherwise take a significant number of iterations to reach their final magnitude.

**Datasets** To investigate how the NWM's representations change when modeling different datasets, we focus on three training sets in this study. Most simply, we first use a dataset of oriented sine functions (depicted in Figure 2 b) with a slowly progressing phase over time steps. This dataset is meant to be a very rough approximation to the spontaneously generated retinal waves observed during development (Ackman et al., 2012). For this dataset, the wavelength and magnitude of the sine waves are fixed, and sequences are generated by randomly sampling an orientation between 0 to $\pi$ and then sequentially progressing the phase by $\frac{1}{9}\pi$ for each timestep until two periods are complete. As a second dataset, we borrow the rotating MNIST dataset from the equivariance literature (Keller & Welling, 2021), consisting of sequences of MNIST digits with each timestep rotated by an additional $\frac{1}{9}\pi$ radians. This dataset serves to allow us to investigate the existence of generalizable spatio-temporal structure in a limited setting. Finally, for more realistic dynamics, we make use of the recent hamiltonian dynamics suite (Botev et al., 2021). At a high level, the benchmark consists of a diversity of tasks governed by known equations of motion, including toy physics examples such as idealized springs, pendulums, orbits, and double-pendulums (Fig 2 c, d, e & h), as well as cyclic games (f & g). Models are evaluated based on their ability to accurately forecast dynamics into the future from a limited number of inference frames.

### 4.1. Measuring Spatiotemporal Structure

To measure the spatiotemporal representational structure that the NWM learns, and its alignment with natural structure, we start with the two simplest tasks: modeling simple sine waves, and modeling rotating MNIST digits. We use three separate methods for analyzing the representations learned on these tasks: Cohen's d selectivity metric (Cohen, 1988) to depict spatial organization, the Hilbert transform to measure the instantaneous phase and velocity of putative waves (Davis et al., 2020), and *artificially induced* traveling waves combined with visualized reconstructions to measure the approximate equivalence of latent traveling waves with observed transformations.

**Topographic Orientation Selectivity** One of the most common methods to demonstrate spatial organization of neural representations is by measuring their selectivity with respect to different features and plotting this with respect to each neuron's position (Hubel & Wiesel, 1974). As an initial test of a basic form of selectivity, namely orientation selectivity, we consider a hypothesis from the literature about how such structure might arise initially in animals (Ackman et al., 2012). Specifically, we investigate whether simple periodic inputs, such as the spontaneous retinal waves observed during early development, are sufficient to encourage smooth topographic organization of orientation selectivity when modeled by a minimal NWM. To test this, we train our model on the simple sine waves dataset, and measure the orientation selectivity of each hidden neuron's time-averaged response to a static 36-element sequences of oriented gratings using Cohen's $d$ metric (Cohen, 1988). In Figure 3 we plot the resulting color/angle of maximal $d$ value for each of the $72 \times 72$ neurons (or a black $\mathbf{x}$ if all $d < 0.65$). We see that the simulated retinal waves do appear to induce topographic organization of orientation selectivity with superficial similarity to the orientation columns of primary visual cortex (Hubel et al., 1978). Outlined in white, we show a manually identified 'pinwheel' where selectivity for all orientations meet, a hallmark of early visual system organization in many species. In relation to prior models of orientation columns (Swindale, 1982), our work does not presuppose the existence of orientation selectivity, but rather it is absent at initialization and it is instead learned in conjunction with topographic organization. We note that the exact statistics of our learned orientation maps have not been measured, and therefore may differ in their current form from those measured in animal studies (Kaschube et al., 2010). In Appendix C.5 we include additional results studying formation mechanism of this orientation selectivity as well as the model parameters which affect the typical length scale of the columns. We leave further precise investigation of the biological similarity to future work.

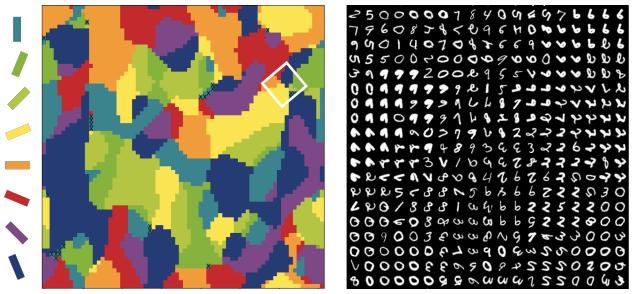

Figure 3: (Left) Plot of orientation selectivity of each NWM hidden neuron $\mathbf{x}$ after training on simple sine waves. (Right) Plot of the maximum activating image for a subset of NWM hidden neurons after training on the rotating MNIST dataset (See Sec. C.6 for full). We see the NWM learns smooth spatial topographic structure tailored to the input dataset.

**General Topographic Organization** On the right of Figure 3, we show the spatial structure of feature selectivity for a network trained on rotating MNIST digits instead. Specifically, we plot the image from the MNIST dataset which maximally activates each neuron in our 2-dimensional hidden state (at the final timestep). We see that neurons are organized with respect to digit class and style, but also orientation, implying that activity is likely to travel over these paths as a traveling wave for observed rotation transformations. Such structure is reminiscent of the higher level category selectivity of the ventral temporal cortex (Kanwisher et al., 1997; Khosla et al., 2022), and also the temporal structure observed to be related to theta oscillations and waves in the hippocampus (Lubenov & Siapas, 2009).

**Instantaneous Phase and Velocity** Next, we demonstrate that the proposed model indeed exhibits full spatiotemporal structure beyond static spatial structure. Compared with biological neural networks, it is easy for us to directly visualize the spatio-temporal activity of our network and qualitatively validate the existence of structure. Figures 1, 2, and 4 provide such examples, while additional samples can be found in Appendix C.7 and the github repository. For additional rigor, however, we borrow state of the art methods from neuroscience to directly compute the instantaneous phase and velocity of putative waves from noisy real-valued signals. Specifically, we follow the work of (Davis et al., 2020) and compute the 'generalized phase' of a real valued signal $\mathbf{x}(t)$ by first transforming the signal to a complex-valued analytic signal $\mathbf{x}_a(t)$ through the Hilbert transform $\mathcal{H}$ and then taking the complex argument of this signal as the phase $\phi(t)$ at each point in space and time. Formally: $\mathbf{x}_a(t) = \mathbf{x}(t) + i\mathcal{H}[\mathbf{x}(t)]$, and $\phi(t) = Arg[\mathbf{x}_a(t)]$. Finally, wave velocities can then straightforwardly be computed using the spatial gradient of this phase: $\boldsymbol{\nu} = -\nabla\phi$. In Figure 4 we depict such phases and velocities for the NWM trained on the rotating MNIST task. We see that, in alignment with expectation, the estimated phases have a spatially periodic

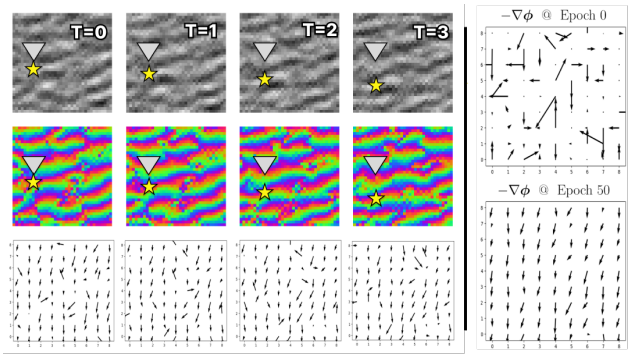


Figure 4: (Left) Plot of hidden state $\mathbf{x}$ (top), generalized phase $\phi$ (mid), and estimated wave velocity $-\nabla\phi$ (bot) over the course of a transformation sequence $T = 0$ to 3. A small gold star moves along with a wave front, relative to a stationary grey triangle, both added to help track the approximate peak of a traveling wave in the hidden state. (Right) Estimated wave velocity before and after training.

pattern which oscillates with sequence length, while the estimated velocities similarly align to point in the downward direction after training (but not before training, as outlined by the disjoint velocity vectors in Figure 4 top right).

**Controlled Generation with Induced Traveling Waves** One of the benefits of structured representations in generative models is that they allow for controlled generation of new observations by taking advantage of the known latent operator for a desired input transformation. In this section we demonstrate that such controlled generation is indeed similarly possible by artificially inducing traveling waves in the NWM hidden state, thereby evidencing the spatiotemporal structure of its representations. Given the high degree of flexibility of the potentially emergent wave dynamics of the 2-D system presented in Figure 4, we concede that two restrictions must be placed on the model in order for us to be able to accurately induce waves which match those the model has learned. Explicitly, we first define the latent space to be a set of disjoint 1-dimensional tori such that learned wave propagation will be restricted to a single axis. Secondly, we restrict our topographic coupling to be 1-directional by masking out all weights except for one (non-central weight) in our convolutional kernel which is shared over all tori. In combination, these restrictions ensure that *if* traveling waves are learned by the model, they will likely be able to be approximately modeled by solutions to the 1-dimensional 1-way wave equation: $y(x,t) = f(x - vt)$.

In Figure 5 we depict the results of this experiment. In detail, we train the 1D NWM described above on a dataset of length $T = 18$ sequences of rotating MNIST digits. At test time, we encode a full sequence (left) and take the final hidden state $\mathbf{x}_T$ as the initial state for our system. We then *induce a traveling wave* in the hidden state in the re-

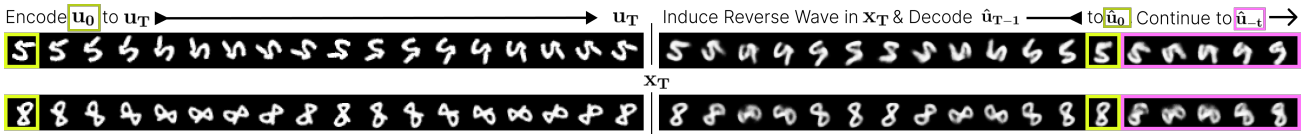


Figure 5: Visualization of controlled generation with induced traveling waves. An input sequence from $\mathbf{u}_0$ to $\mathbf{u}_T$ (left) gets encoded to a hidden state $\mathbf{x}_T$. We then induce a traveling wave in the opposite direction of the estimated instantaneous velocity and observe we can decode back to the original input $\hat{\mathbf{u}}_0$ (highlighted yellow, right). Furthermore, we see by continuing the wave, we can continue the transformation past the bounds of the input sequence (highlighted pink, right).

verse direction of the instantaneous velocity. In practice, since we have limited our system to 1-dimensional tori, this corresponds to sequentially cyclically shifting (or linearly interpolating) activations across the spatial dimension of each circular subspace according to the inverse of our assumed velocity. The result in Figure 5 (right) shows that indeed by inducing such reverse traveling waves we can then decode the original input sequence, and even predict elements before the start of the sequence (highlighted in pink). Such sensible decodings highlight the generalization power of the representational structure learned by the NWM. In this example we propagate waves with assumed velocity $v = 1$ and observe that this is slightly faster than the ground truth transformation, resulting in a return to the start state in 14 steps rather than 18. Additional transformations can be found in Appendix Figure 11.

### 4.2. Computational Implications of Structure

Given the structure measured in Section 4.1 is known to be related to beneficial inductive biases (Fukushima, 1980; Keller & Welling, 2021), in this section we perform preliminary experiments to measure such potential benefits in the context of sequence modeling.

**An Inductive Bias for Simple Physical Dynamics** First, inspired by the literature relating traveling waves to visual motion perception (Jancke et al., 2004) and spatiotemporal structure in the hippocampus (Lubenov & Siapas, 2009), we hypothesize that the spatiotemporal structure of the NWM demonstrated in Section 4.1 may serve as an inductive bias towards simple physical dynamics. To measure this, we train NWM models on a representative subset of the Hamiltonian dynamics suite, and measure their error when attempting to forecast long test trajectories into the future. Specifically, we consider six distinct dynamic modeling tasks: three simple physical dynamics including the pendulum, spring, and two body gravitational tasks; one less physical but still temporally smooth task, namely the matching pennies task; and the last, the double pendulum, a complex chaotic physical dynamics task. We compare performance of the NWM with the state of the art baselines using optimal hyperparameters directly given in prior work (Botev et al., 2021; Higgins et al., 2021). These include the HGN++ (Higgins et al., 2021), a standard autoregressive model (AR) (Hochreiter & Schmidhuber, 1997), and a Neural ODE (Chen et al., 2018) trained both forwards and backwards in time (ODE [TR]). We additionally include a final globally coupled coRNN baseline with equivalent parameters to our NWM to study the direct impact of the imposed structure on model performance. In Table 1 we see that, in alignment with our intuition, the NWM models achieve the lowest forecasting error on the simple physical dynamics tasks, providing evidence in support of the hypothesis that the observed spatiotemporal structure of Section 4.1 is beneficial for modeling such systems. Further, we see that the coRNN baseline performs the best on the less physical but predictable matching pennies task, while the maximally flexible Neural ODE performs the best on the chaotic double pendulum task. Despite these promising results, we note that accurately measuring forecasting performance in image space is notoriously hard (Botev et al., 2021; Higgins et al., 2021), and therefore recommend future work pursue the development of alternative benchmarks and metrics for evaluating the beneficial inductive biases present in the NWM and other forecasting models. In Appendix C.3 and the limitations section below we include additional discussion of these considerations.

**Efficiency** As a second potential benefit related to the NWM's demonstrated spatiotemporal structure, our neural wave machines are highly parameter efficient by design when compared to the globally coupled coRNN. As explained in Section 3, the recurrent connections of our model are restricted to be entirely local as implemented by the convolution operation, thereby allowing for arbitrarily large hidden state sizes with a constant number of recurrent parameters, significantly improving over the quadratically increasing number of parameters in the coRNN. In Table 4 we see that on the canonical long sequence classification tasks of sequential MNIST (sMNIST) and permuted sequential MNIST (psMNIST) (Rusch & Mishra, 2021), our model achieves comparable performance with the coRNN (and thus existing state of the art) while requiring a fraction of the parameters. In Appendix C.2 we include additional results on other sequence modeling tasks such as IMDB sentiment classification and long sequence addition showing the same benefits. Interestingly, efficiency in terms of wiring length is also implicated in the formation of orientation columns in natural systems (Koulakov & Chklovskii,

Table 1: Forward extrapolation mean squared reconstruction error on the Hamiltonian Dynamics Benchmark held-out test set (displayed in units of $1 \times 10^{-8}$). We see, in alignment with intuition, the 1 and 2-dimensional Neural Wave Machines (NWM 1D & 2D) perform best on simple physically realistic dynamics such as the spring, pendulum, and two body problem. The globally coupled coRNN performs best on the smooth, but non-physical, matching pennies task, while the maximally flexible Neural ODE performs best on the highly complex and chaotic double pendulum task.

| | AR | HGN++ | ODE | coRNN | NWM 2D | NWM 1D |
|---|---|---|---|---|---|---|
| Spring | 20.97 | 1.58 | 1.58 | 2.52 | 5.46 | **1.45** |
| Pendulum | 4,208.0 | 166.5 | 166.0 | 548.0 | **110.9** | 237.2 |
| Two Body | 91.4 | 5.0 | 4.2 | 2.0 | 1.9 | **0.9** |
| Pennies | 126.3 | 190.0 | 119.3 | **28.2** | 47.2 | 43.1 |
| Double Pendulum | 3,905.0 | 1,531.0 | **1,296.0** | 1,666.0 | 2,512.0 | 2,821.0 |

2001). We believe that our work reinforces this relationship from another perspective by showing that when a recurrent oscillatory computational system is constrained to be wiring length efficient by design, it naturally learns topographic organization (e.g. Figure 3) in order to optimally function.

## 5. Discussion

In this work we introduce the Neural Wave Machine, a recurrent neural network model shown to learn spatiotemporally structured representations through local connectivity and oscillatory dynamics. We propose this model as a rich testing ground for the diversity of computational hypotheses surrounding traveling waves in the neuroscience literature, and demonstrate its potential value in this regard by providing evidence for a variety of hypotheses, including one relating to the origin of orientation columns, and one relating to a simple physical inductive bias. Further, we show that this model is competitive with state of the art on sequence modeling tasks, hoping to encourage future use of such models to study the computational purpose of spatiotemporal dynamics in natural systems.

**Related Work** In recent years, multiple works have attempted to integrate topographic organization in deep neural networks for various purposes including learning generalized invariance (Kavukcuoglu et al., 2009), learning generalized equivariance (Keller & Welling, 2021) or for developing more accurate models of the development and structure of natural systems (Lee et al., 2020; Doshi & Konkle, 2022; Blauch et al., 2022). Other work has studied the temporal aspects of neural activations and attempted to integrate such structure into deep neural networks. For example, researchers have studied the integration of recurrence into feed forward classification networks (Kietzmann et al., 2019), or the integration spike-time coding through complex activations (Löwe et al., 2022). Separately, others have aimed to directly integrate natural architectural biases by fixing early layers of a convolutional neural network to mimic the early stages of the natural visual stream, ultimately resulting in improved robustness (Dapello et al., 2020). Our work is highly related to these efforts in motivation, but largely unique in terms of methodology and its focus on complex spatiotemporal dynamics such as traveling waves. One class of models which shares some relation intuitively is reservoir computing (Lukoševičius & Jaeger, 2009). A primary difference between the NWM and reservoir computing frameworks is that our network has a significant number of learned parameters within its recurrence that mediate complex hidden dynamics, while prior work typically relies on a reservoir of fixed dynamics.

Table 2: Test accuracy on supervised sequence benchmarks. All results are mean $\pm$ std. over 3 random initalizations.

| | sMNIST | | psMNIST | |
|---|---|---|---|---|
| | Acc. | $\#\theta$ | Acc. | $\#\theta$ |
| coRNN | $99.1 \pm 0.1$ | 134k | $95.0 \pm 2.4$ | 134k |
| NWM | $98.6 \pm 0.3$ | 50k | $94.8 \pm 1.1$ | 50k |

**Limitations** In this work we have put significant effort into quantifying the existence of complex spatiotemporal structure and its impact on the NWMs computational performance. However, due to the inherent flexibility of the possible dynamics which may emerge, there remain limitations in our ability to due so. In future work, we would hope to be able to get a more concrete metric corresponding to spatiotemporal structure to better correlate the structure of our models with their performance. Furthermore, on tasks such as forecasting dynamics, it is still an open question how to best compare the performance of such models in the most comprehensive and fair manner (Higgins et al., 2021). In Appendix C.3 we include additional metrics evaluating model performances on the Hamiltonian Dynamics Suite, highlighting this challenge. Finally, our explorations of parameter efficiency are inherently preliminary and use fully connected encoders and decoders in the NWM, ultimately contributing 45k of the 50k parameters noted for the NWM in Table 4. If we were able to replace these components with similarly locally connected functions, such as

convolutional networks, the parameter efficiency would further dramatically increase.

**Conclusion** As a flexible computational model of traveling waves, we believe the NWM framework offers significant potential to the computational neuroscience community as a method for testing other computational hypotheses relating to traveling waves and synchronous neural dynamics broadly. Similar to convolutional neural networks for modeling the visual system (Yamins et al., 2014; Cadieu et al., 2014; Kanwisher et al., 2023), neural wave machines do not match all biologically relevant details of neural dynamics, but we believe they may capture sufficient abstract properties to be useful for performing investigations that otherwise wouldn't be possible. Examples of initial hypotheses which we believe would be primarily suited for future study would be the use of traveling waves as a short term memory mechanism (Bhattacharya et al., 2022a), or as a mechanism for sequencing actions (Sato, 2022). Ultimately, we believe this work suggests that complex spatiotemporal dynamics and structure should be investigated further in the future to develop the next set of inductive biases necessary to bring deep neural networks to the same levels of efficiency and robustness that we see in natural intelligence.

**Acknowledgements** We would like to thank the creators of Weight & Biases (Biewald, 2020) and PyTorch (Paszke et al., 2019). Without these tools our work would not have been possible. We thank the Bosch Center for Artificial Intelligence for funding. Finally, we thank the reviewers for their proposed additions and constructive comments.

## References

Ackman, J. B., Burbridge, T. J., and Crair, M. C. Retinal waves coordinate patterned activity throughout the developing visual system. *Nature*, 490(7419):219–225, October 2012. doi: 10.1038/nature11529. URL https://doi.org/10.1038/nature11529.

Alamia, A. and VanRullen, R. Alpha oscillations and traveling waves: Signatures of predictive coding? *PLOS Biology*, 17(10):e3000487, October 2019. doi: 10.1371/journal.pbio.3000487. URL https://doi.org/10.1371/journal.pbio.3000487.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization, 2016. URL https://arxiv.org/abs/1607.06450.

Besserve, M., Logothetis, N., and Schölkopf, B. Shifts of gamma phase across primary visual cortical sites reflect dynamic stimulus-modulated information transfer. 01 2015. doi: 10.15496/publikation-10582.

Bhattacharya, S., Brincat, S. L., Lundqvist, M., and Miller, E. K. Traveling waves in the prefrontal cortex during working memory. *PLOS Computational Biology*, 18(1):1–22, 01 2022a. doi: 10.1371/journal.pcbi.1009827. URL https://doi.org/10.1371/journal.pcbi.1009827.

Bhattacharya, S., Donoghue, J. A., Mahnke, M., Brincat, S. L., Brown, E. N., and Miller, E. K. Propofol anesthesia alters cortical traveling waves. *Journal of Cognitive Neuroscience*, 34(7):1274–1286, 2022b.

Biewald, L. Experiment tracking with weights and biases, 2020. URL https://www.wandb.com/. Software available from wandb.com.

Blauch, N. M., Behrmann, M., and Plaut, D. C. A connectivity-constrained computational account of topographic organization in primate high-level visual cortex. *Proceedings of the National Academy of Sciences*, 119(3):e2112566119, 2022. doi: 10.1073/pnas.2112566119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2112566119.

Bordelon, B. and Pehlevan, C. Population codes enable learning from few examples by shaping inductive bias. *eLife*, 11:e78606, dec 2022. ISSN 2050-084X. doi: 10.7554/eLife.78606. URL https://doi.org/10.7554/eLife.78606.

Botev, A., Jaegle, A., Wirnsberger, P., Hennes, D., and Higgins, I. Which priors matter? benchmarking models for learning latent dynamics, 2021. URL https://arxiv.org/abs/2111.05458.

Bouchacourt, D., Ibrahim, M., and Deny, S. Addressing the topological defects of disentanglement via distributed operators. *ArXiv*, abs/2102.05623, 2021.

Bragin, A., Jandó, G., Nádasdy, Z., Hetke, J., Wise, K., and Buzsáki, G. Gamma (40-100 hz) oscillation in the hippocampus of the behaving rat. *Journal of neuroscience*, 15(1):47–60, 1995.

Breakspear, M., Heitmann, S., and Daffertshofer, A. Generative models of cortical oscillations: Neurobiological implications of the kuramoto model. *Frontiers in Human Neuroscience*, 4, 2010. ISSN 1662-5161. doi: 10.3389/fnhum.2010.00190. URL https://www.frontiersin.org/articles/10.3389/fnhum.2010.00190.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., and DiCarlo, J. J. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLOS Computational Biology*, 10(12):1–18, 12 2014. doi: 10.1371/

journal.pcbi.1003963. URL https://doi.org/10.1371/journal.pcbi.1003963.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. Neural ordinary differential equations, 2018. URL https://arxiv.org/abs/1806.07366.

Cohen, J. *Statistical Power Analysis for the behavioral sciences*. L. Erlbaum Associates, 1988.

Cohen, T. and Welling, M. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999, 2016.

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D. D., and DiCarlo, J. J. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *bioRxiv*, 2020. doi: 10.1101/2020.06.16.154542. URL https://www.biorxiv.org/content/early/2020/06/17/2020.06.16.154542.

Davis, Z. W., Muller, L., Martinez-Trujillo, J., Sejnowski, T., and Reynolds, J. H. Spontaneous travelling cortical waves gate perception in behaving primates. *Nature*, 587(7834):432–436, October 2020. doi: 10.1038/s41586-020-2802-y. URL https://doi.org/10.1038/s41586-020-2802-y.

Davis, Z. W., Benigno, G. B., Fletterman, C., Desbordes, T., Steward, C., Sejnowski, T. J., Reynolds, J. H., and Muller, L. Spontaneous traveling waves naturally emerge from horizontal fiber time delays and travel through locally asynchronous-irregular states. *Nature Communications*, 12(1), October 2021. doi: 10.1038/s41467-021-26175-1. URL https://doi.org/10.1038/s41467-021-26175-1.

Diamant, N. and Bortoff, A. Nature of the intestinal low-wave frequency gradient. *American Journal of Physiology-Legacy Content*, 216(2):301–307, February 1969. doi: 10.1152/ajplegacy.1969.216.2.301. URL https://doi.org/10.1152/ajplegacy.1969.216.2.301.

Doshi, F. R. and Konkle, T. Visual object topographic motifs emerge from self-organization of a unified representational space. *bioRxiv*, 2022. doi: 10.1101/2022.09.06.506403. URL https://www.biorxiv.org/content/early/2022/09/08/2022.09.06.506403.

Elesedy, B. and Zaidi, S. Provably strict generalisation benefit for equivariant models. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2959–2969. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/elesedy21a.html.

Ermentrout, B., Borisyuk, A., Friedman, A., and Terman, D. *Neural Oscillators*, volume 1860, pp. 69–106. 01 1970. ISBN 978-3-540-23858-4. doi: 10.1007/978-3-540-31544-5_3.

Ermentrout, G. B. and Kleinfeld, D. Traveling electrical waves in cortex: insights from phase dynamics and speculation on a computational role. *Neuron*, 29(1):33–44, 2001.

Ermentrout, G. B. and Kopell, N. Frequency plateaus in a chain of weakly coupled oscillators, i. *SIAM journal on Mathematical Analysis*, 15(2):215–237, 1984.

Friston, K. J. Waves of prediction. *PLOS Biology*, 17 (10):e3000426, October 2019. doi: 10.1371/journal.pbio.3000426. URL https://doi.org/10.1371/journal.pbio.3000426.

Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36 (4):193–202, April 1980. doi: 10.1007/bf00344251. URL https://doi.org/10.1007/bf00344251.

Gong, P. and van Leeuwen, C. Distributed dynamical computation in neural circuits with propagating coherent activity patterns. *PLOS Computational Biology*, 5(12):1–11, 12 2009. doi: 10.1371/journal.pcbi.1000611. URL https://doi.org/10.1371/journal.pcbi.1000611.

Heitmann, S. and Ermentrout, G. B. Direction-selective motion discrimination by traveling waves in visual cortex. *PLOS Computational Biology*, 16(9):1–20, 09 2020. doi: 10.1371/journal.pcbi.1008164. URL https://doi.org/10.1371/journal.pcbi.1008164.

Higgins, I., Wirnsberger, P., Jaegle, A., and Botev, A. Symetric: Measuring the quality of learnt hamiltonian dynamics inferred from vision. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 25591–25605. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/d6ef5f7fa914c19931a55bb262ec879c-Paper.pdf.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Hubel, D. H. and Wiesel, T. N. Sequence regularity and geometry of orientation columns in the monkey striate cortex. *Journal of Comparative Neurology*, 158(3):267–293, 1974.

Hubel, D. H., Wiesel, T. N., and Stryker, M. P. Anatomical demonstration of orientation columns in macaque monkey. *Journal of Comparative Neurology*, 177(3):361–379, 1978.

Hughes, D. J. R. The phenomenon of travelling waves: A review. *Clinical Electroencephalography*, 26(1):1–6, 1995. doi: 10.1177/155005949502600103. URL https://doi.org/10.1177/155005949502600103.

Izhikevich, E. M. and Hoppensteadt, F. C. Polychronous wavefront computations. *International Journal of Bifurcation and Chaos*, 19(5):1733–1739, 01 2008.

Jancke, D., Chavane, F., Naaman, S., and Grinvald, A. Imaging cortical correlates of illusion in early visual cortex. *Nature*, 428(6981):423–426, 2004.

Jeong, S.-O., Ko, T.-W., and Moon, H.-T. Time-delayed spatial patterns in a two-dimensional array of coupled oscillators. *Phys. Rev. Lett.*, 89:154104, Sep 2002. doi: 10.1103/PhysRevLett.89.154104. URL https://link.aps.org/doi/10.1103/PhysRevLett.89.154104.

Kanwisher, N., McDermott, J., and Chun, M. M. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11):4302–4311, 1997.

Kanwisher, N., Khosla, M., and Dobs, K. Using artificial neural networks to ask 'why' questions of minds and brains. *Trends in Neurosciences*, 2023. ISSN 0166-2236. doi: https://doi.org/10.1016/j.tins.2022.12.008. URL https://www.sciencedirect.com/science/article/pii/S0166223622002624.

Kaschube, M., Schnabel, M., Löwel, S., Coppola, D. M., White, L. E., and Wolf, F. Universality in the evolution of orientation columns in the visual cortex. *Science*, 330(6007):1113–1116, 2010. doi: 10.1126/science.1194869. URL https://www.science.org/doi/abs/10.1126/science.1194869.

Kavukcuoglu, K., Ranzato, M., Fergus, R., and LeCun, Y. Learning invariant features through topographic filter maps. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1605–1612, 2009. doi: 10.1109/CVPR.2009.5206545.

Keller, T. A. and Welling, M. Topographic vaes learn equivariant capsules. In *Advances in Neural Information Processing Systems*, volume 34, pp. 28585–28597. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/f03704cb51f02f80b09bffba15751691-Paper.pdf.

Keller, T. A., Gao, Q., and Welling, M. Modeling category-selective cortical regions with topographic variational autoencoders. In *SVRHM 2021 Workshop @ NeurIPS*, 2021. URL https://openreview.net/forum?id=yGRq_lW54bI.

Khosla, M., Ratan Murty, N. A., and Kanwisher, N. A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition. *Current Biology*, 32(19):4159–4171.e9, 2022. ISSN 0960-9822. doi: https://doi.org/10.1016/j.cub.2022.08.009. URL https://www.sciencedirect.com/science/article/pii/S0960982222012866.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., and Kriegeskorte, N. Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43):21854–21863, 2019. doi: 10.1073/pnas.1905544116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1905544116.

King, J.-R. and Wyart, V. The human brain encodes a chronicle of visual events at each instant of time through the multiplexing of traveling waves. *The Journal of Neuroscience*, 41(34):7224–7233, April 2021. doi: 10.1523/jneurosci.2098-20.2021. URL https://doi.org/10.1523/jneurosci.2098-20.2021.

Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Koulakov, A. A. and Chklovskii, D. B. Orientation preference patterns in mammalian visual cortex: a wire length minimization approach. *Neuron*, 29(2):519–527, 2001.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Kuramoto, Y. Rhythms and turbulence in populations of chemical oscillators. *Physica A: Statistical Mechanics and its Applications*, 106(1-2):128–143, 1981.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017. doi: 10.1017/S0140525X16001837.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Lee, H., Margalit, E., Jozwik, K. M., Cohen, M. A., Kanwisher, N., Yamins, D. L. K., and DiCarlo, J. J. Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. *bioRxiv*, 2020. doi: 10.1101/2020.07.09.185116. URL https://www.biorxiv.org/content/early/2020/07/10/2020.07.09.185116.

Lubenov, E. V. and Siapas, A. G. Hippocampal theta oscillations are travelling waves. *Nature*, 459(7246):534–539, May 2009. doi: 10.1038/nature08010. URL https://doi.org/10.1038/nature08010.

Lukoševičius, M. and Jaeger, H. Reservoir computing approaches to recurrent neural network training. *Computer science review*, 3(3):127–149, 2009.

Löwe, S., Lippe, P., Rudolph, M., and Welling, M. Complex-valued autoencoders for object discovery, 2022. URL https://arxiv.org/abs/2204.02075.

Muller, L., Reynaud, A., Chavane, F., and Destexhe, A. The stimulus-evoked population response in visual cortex of awake monkey is a propagating wave. *Nature Communications*, 5(1), April 2014. doi: 10.1038/ncomms4675. URL https://doi.org/10.1038/ncomms4675.

Muller, L., Piantoni, G., Koller, D., Cash, S. S., Halgren, E., and Sejnowski, T. J. Rotating waves during human sleep spindles organize global patterns of activity that repeat precisely through the night. *eLife*, 5, November 2016a. doi: 10.7554/elife.17267. URL https://doi.org/10.7554/elife.17267.

Muller, L., Piantoni, G., Koller, D., Cash, S. S., Halgren, E., and Sejnowski, T. J. Rotating waves during human sleep spindles organize global patterns of activity that repeat precisely through the night. *eLife*, 5:e17267, nov 2016b. ISSN 2050-084X. doi: 10.7554/eLife.17267. URL https://doi.org/10.7554/eLife.17267.

Muller, L., Chavane, F., Reynolds, J., and Sejnowski, T. J. Cortical travelling waves: mechanisms and computational principles. *Nature Reviews Neuroscience*, 19(5):255–268, March 2018. doi: 10.1038/nrn.2018.20. URL https://doi.org/10.1038/nrn.2018.20.

Neftci, E. O., Mostafa, H., and Zenke, F. Surrogate gradient learning in spiking neural networks, 2019. URL https://arxiv.org/abs/1901.09948.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. 2019.

Ricci, M., Jung, M., Zhang, Y., Chalvidal, M., Soni, A., and Serre, T. Kuranet: Systems of coupled oscillators that learn to synchronize, 2021. URL https://arxiv.org/abs/2105.02838.

Rusch, T. K. and Mishra, S. Coupled oscillatory recurrent neural network (cornn): An accurate and (gradient) stable architecture for learning long time dependencies. In *International Conference on Learning Representations*, 2021.

Sato, N. Cortical traveling waves reflect state-dependent hierarchical sequencing of local regions in the human connectome network. *Scientific Reports*, 12(1), January 2022. doi: 10.1038/s41598-021-04169-9. URL https://doi.org/10.1038/s41598-021-04169-9.

Sato, T. K., Nauhaus, I., and Carandini, M. Traveling waves in visual cortex. *Neuron*, 75(2):218–229, July 2012. doi: 10.1016/j.neuron.2012.06.029. URL https://doi.org/10.1016/j.neuron.2012.06.029.

Swindale, N. V. A model for the formation of orientation columns. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 215(1199):211–230, 1982. ISSN 00804649. URL http://www.jstor.org/stable/35596.

van der Pol, E., Worrall, D. E., van Hoof, H., Oliehoek, F. A., and Welling, M. MDP homomorphic networks: Group symmetries in reinforcement learning. *CoRR*, abs/2006.16908, 2020. URL https://arxiv.org/abs/2006.16908.

van der Wilk, M., Bauer, M., John, S., and Hensman, J. Learning invariances using the marginal likelihood. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/d465f14a648b3d0a1faa6f447e526c60-Paper.pdf.

Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., and Welling, M. Rotation equivariant cnns for digital pathology. *CoRR*, abs/1806.03962, 2018. URL http://arxiv.org/abs/1806.03962.

Wiesel, T. N. and Hubel, D. H. Ordered arrangement of orientation columns in monkeys lacking visual experience. *Journal of Comparative Neurology*, 158(3):307–318, 1974. doi: https://doi.org/10.1002/cne.901580306. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cne.901580306.

Wolpert, D. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8, 03 1996. doi: 10.1162/neco.1996.8.7.1341.

Worrall, D. E., Garbin, S. J., Turmukhambetov, D., and Brostow, G. J. Harmonic networks: Deep translation and rotation equivariance. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7168–7177, 2017.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. doi: 10.1073/pnas.1403112111. URL https://www.pnas.org/doi/abs/10.1073/pnas.1403112111.

Zanos, T. P., Mineault, P. J., Nasiotis, K. T., Guitton, D., and Pack, C. C. A sensorimotor role for traveling waves in primate visual cortex. *Neuron*, 85(3):615–627, 2015.

Zhang, H., Watrous, A. J., Patel, A., and Jacobs, J. Theta and alpha oscillations are traveling waves in the human neocortex. *Neuron*, 98(6):1269–1281.e4, June 2018. doi: 10.1016/j.neuron.2018.05.019. URL https://doi.org/10.1016/j.neuron.2018.05.019.

# A. Experiment Details

Videos of traveling waves and code to reproduce all experiments in the paper can be found at the following github repository: `https://github.com/akandykeller/NeuralWaveMachines`.

The code is built as extensions of three existing public repositories, allowing us to reproduce all baseline results from the original authors' code. Specifically, we make use: (I) The coRNN repository (`https://github.com/tk-rusch/coRNN`) for the supervised sequence experiments, (II) The Topographic VAE repository (`https://github.com/akandykeller/TopographicVAE/`) for the rotating MNIST experiments, and (III) The DeepMind Physics Inspired Models repository (`https://github.com/deepmind/deepmind-research/tree/master/physics_inspired_models`) for the Hamiltonian Dynamics Suite Experiments.

## A.1. Sequence Classification

The efficiency experiments from Section 4.2 were performed by modifying the published code for the original coRNN (Rusch & Mishra, 2021) to incorporate the local connectivity constraints outlined in the main text. All hyperparameters were thus set to the defaults in the published code which matched the optimal hyperparameters stated by the authors to be found from a grid search on each dataset independently. The baseline coRNN values in Table 4 are thus simply from re-running the original authors code, and we observe similar values to those published in (Rusch & Mishra, 2021). We acknowledge that running a separate grid search for the NWM models may be beneficial to their performance but we were unable to do so due to time and computational constraints and thus leave this to future work. In practice, we found the original coRNN parameters worked well enough to give an initial intuition for the relative performance of the NWM.

For the NWM, the topology of the hidden state was defined to be a regular square 2D grid with side lengths equal to square root of the default hidden state size (or the integer floor of the square root for non-perfect-square values). Each neuron was defined to be connected to its immediate surrounding 8 cells in the grid, in addition to a self-connection. The boundary conditions of the topology were defined to be periodic (implemented through circular padding) such that the global topology was that of a 2-dimensional torus. The recurrent local coupling parameters were shared over all spatial locations of the grid, allowing the above local connectivity to be implemented as a periodic convolution with a kernel of size $3 \times 3$. We noted that increasing the number of channels in the convolutional layers dramatically improved performance, and thus for the NWM models in Table 4 we use 16 channels in the hidden state. This yeilded a parameter count computation of: $\#\theta = 1 \times 256 \times 16 + 16 \times 16 \times 3 \times 3 \times 2 + 256 \times 16 \times 10 = 49,664$.

## A.2. Rotating MNIST and Sine Waves

The experiments on measuring spatiotemporal structure using the MNIST and simple sine waves datasets were performed by modifying the published code for the Topographic VAE (Keller & Welling, 2021) to introduce our proposed NWM in place of the 'shifting temporal coherence' construction of the topographic Student's-T variable in the original paper. To achieve this, the encoder and decoder ($f_\theta$ & $g_\theta$) were implemented as a variational autoencoder (Kingma & Welling, 2014) with a standard Gaussian prior and Bernoulli distribution for the likelihood of the data. Practically, this was achieved by setting the output dimensionality of the encoder $f_\theta$ to twice the hidden state dimensionality, defining half of the outputs as the posterior mean $\mu_\theta$, and the second half as the log of the posterior variance $\sigma_\theta$. We additionally found that applying Layer Normalization (Ba et al., 2016) (denoted LN) to the output of the encoder helped increase convergence speed. Explicitly, the model can thus be described as:

$$\mathbf{z}_{t+1} \sim q_\theta(\mathbf{z}_{t+1}|\mathbf{u}_{t+1}) = \mathcal{N}\big(\mathbf{z}_{t+1}; \mu_\theta(\mathbf{u}_{t+1}), \sigma_\theta(\mathbf{u}_{t+1})\mathbf{I}\big), \qquad \bar{\mathbf{z}}_{t+1} = \mathrm{LN}(\mathbf{z}_{t+1}) \tag{6}$$

$$\mathbf{v}_{t+1} = \mathbf{v}_t + \Delta t\big(\sigma\left(\mathbf{w}_x \star \mathbf{x}_t + \mathbf{w}_{\dot{x}} \star \mathbf{v}_t + \mathbf{V}\bar{\mathbf{z}}_{t+1} + \mathbf{b}\right) - \gamma\mathbf{x}_t - \alpha\mathbf{v}_t\big) \tag{7}$$

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \Delta t\left(\mathbf{v}_{t+1}\right) \tag{8}$$

$$p_\theta(\mathbf{u}_{t+2}|g_\theta(\mathbf{x}_{t+1})) = \mathrm{Bernoilli}(\mathbf{u}_{t+2}; g_\theta(\mathbf{x}_{t+1})) \tag{9}$$

Where the objective is then computed by averaging the evidence lower bound (ELBO) over the length of the sequence:

$$\mathcal{L}(\mathbf{u}_{1:T}; \boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathbf{z}_t \sim q_\theta(\mathbf{z}_t|\mathbf{u}_t)}\big(\log p_\theta(\mathbf{u}_{t+1}|g_\theta(\mathbf{x}_t)) - D_{KL}[q_\theta(\mathbf{z}_t|\mathbf{u}_t)||p_{\mathbf{Z}}(\mathbf{z}_t)]\big) \tag{10}$$

The initial conditions for the NWM were then given by simply setting the initial position equal to the first encoder output, and the initial velocity to zero, i.e. $\mathbf{x}_0 = \bar{\mathbf{z}}_0$ & $\mathbf{v}_0 = \mathbf{0}$. Although we did not test the MNIST experiments with a deterministic

autoencoder, we note that traveling waves can also clearly be seen in the hidden states of the deterministic models presented in Sections 3 and 4.2 (as visualized in Figure 2 and the supplementary material), implying that the variational formulation is not necessary for the emergence of traveling waves.

For the experiment depicted in Figure 4 of Section 4, we used a simple linear encoder and decoder, and a hidden state dimensionality of 1296 reshaped into a 2D grid of shape $36 \times 36$. As in the rest of the paper, our topographic connectivity was implemented using a convolutional kernel of shape $3 \times 3$ shared over all elements of the grid, with circular padding to enforce periodic boundary conditions on the grid. For training, we presented the model with length 18 sequences of MNIST digits rotating at 20 degrees per step (thus completing a full period per training sequence). At test time, to create the visualization in Figure 4, we increased the sequence length to 72 elements (or four periods) and visualize a portion of the final period, allowing the system to reach a steady state of wave activity for better visualization. We see that despite not being trained on such long sequences, the NWM is able to generalize and maintain wave activity. For computing the generalized phase, we set use a 4-th order butterworth bandpass filter with bounds set at 0.2 and 0.4 of the Nyquist frequency. As hyperparamters for training, we used standard SGD with momentum of $0.9$, a learning rate of $2.5 \times 10^{-4}$, and a batch size of $128$ for 50 epochs. Following the suggestion outlined in (Rusch & Mishra, 2021), we allowed the parameters $\gamma$, $\alpha$, & $\Delta t$ to be learned during training by initializing them to $\Delta t = \sigma^{-1}(0.125) = -1.95$, $\gamma = 1.0$, & $\alpha = 0.5$ and then applying appropriate activation functions to keep them within the desired bounds (e.g. sigmoid, ReLU, & ReLU respectively). These hyperparameters and initalization values were determined by implementing a simple toy version of the model with random data and random weights and manually altering parameters to determine the ranges for which coherent wave dynamics were likely to emerge. We note that the properties of the emergent waves appear qualitatively different for different random initalizations of the model. Specifically the wavelength and velocity of the waves appears to vary greatly from run-to-run. We show a few of these different learned dynamics in the additional results section below.

For the experiment depicted in Figure 5 of Section 4, we used a 3-layer Multi-Layer Perceptron (MLP) for the both encoder and decoder, and a hidden state of dimensionality 1296 reshaped into a set of 24 disjoint 1-D tori (circles) each composed of 54 neurons. We implemented topographic coupling between the immediate neighbors on each circle via a 1-dimensional convolutional kernel of size 3 with circular padding. We then implemented the uni-directionality constraint outlined in the main text be masking the first two elements of the kernel to 0, yielding a kernel with a single trainable parameter explicitly connecting each neuron with its neighbor directly to one side. For training, the dataset and hyperparameters all remained the same as in Figure 4 described above, however the batch size was reduced to $8$ for quicker evaluation. We found that additionally adding another layer normalization layer between recurrent steps improved the consistency of the learned waves and thus allowed us to simulate them more accurately at test time. Explicitly this amounted to modifying Equation 8 to: $\mathbf{x}_{t+1} = \mathrm{LN}\big(\mathbf{x}_t + \Delta t\,(\mathbf{v}_{t+1})\big)$. Furthermore, to ensure consistency of waves across each circular subspace separately, we shared the bias vector $\mathbf{b}$ across each subspace. To induce a traveling wave in the hidden state of the network and thereby generate the transformation sequence shown in the bottom row of the figure, we first encode the input sequence (shown in the top row), using the equations outlined in this section. We take the final hidden state of the network ($\mathbf{x}_T$) as the initial state from which we begin the wave propagation. Then, across each 1-D circular subspace of the hidden state, we update the values of the hidden state based on the 1-D 1-way wave equation $y(x,t) = f(x - vt)$ for a velocity $v = 1$ for time $t = 1$ to $18$. Written in terms of the hidden state $\mathbf{x}_t$, we can effectively propagate waves backwards through the hidden state by moving activation from one spatial location $l$ to a location shifted by $v\Delta t$: $\mathbf{x}_T(l) \rightarrow \mathbf{x}_T(l - v\Delta t)$. Practically, this amounts to sequentially circularly shifting the hidden state activation across each circular subspace as depicted in Figure 5.

### A.3. Hamiltonian Dynamics Suite

The experiments in Section 4.2 were performed using the DeepMind Physics Inspired Models and Hamiltonian Dynamics Suite, implemented in JAX, as a starting point. All values reported for the baselines (HGN++, AR, and ODE [TR]) were thus obtained by re-running the original code with the hyperparameters stated in (Botev et al., 2021). Specifically, for the HGN++, we trained the model both forwards and backwards in time, including over the inference steps, with a final beta value of $0.1$ in the ELBO. For the AR model, we used an LSTM with all other paramters default. For the ODE, we used the default parameters with forwards and backwards training, again including inference steps. The only change to the default hyperparamters for all three models was to reduce the batch size to 8 per GPU (thus 32 total per iteration) to fit on our GPUs.

The coRNN and NWM architectures were added as extensions to the auto-regressive model already implemented in library. They thus made use of all the same default hyperparameters, with the only changed values being the aforementioned reduced batch size, an increased number of inference steps (31), an increased number of target steps (60), and an increased hidden state size ($23 \times 23$). The increased number of inference and target steps was found useful to improve performance on more

chaotic tasks such as the pendulum where the accuracy of the initial state is hugely important to the model forecasting performance. Additionally, we note that these values are within the values searched by the grid search of the authors in (Botev et al., 2021) making their use here for comparison relatively fair. The size of the hidden state was picked as the largest which fit in our GPU memory across all devices. The values of $\alpha$, $\gamma$, and $\Delta t$ were initialized to the same values as the MNIST experiments described above, and were again allowed to be updated during training simultaneously with the other model parameters. For the 2D NWM, the hidden state topology was again defined to be a 2D torus of size (23×23) implemented through periodic convolution with a $3 \times 3$ kernel. The 1D NWM topology was similarly composed of 23 disjoint 1D circles each with 23 neurons, again implemented with periodic convolution with a $1 \times 3$ kernel. The coRNN and NWM models additionally used a separate initial condition network to initialize $\mathbf{x}_0$ and $\mathbf{v}_0$. This network was implemented as a GRU with a hidden state of size $2 \times 23 \times 23$ which ran backwards over the inference sequence (length 31) first embedded with the model encoder $f_\theta$. The final hidden state of the model was then split in half and taken to initialize the inital positions and velocities of the coRNN & NWMs.

All models make use of the same deep convolutional encoder with ReLU activations and a similarly deep convolutional spatial broadcast decoder as in the original work. They were similarly all trained for 500,000 iterations to match the original work.

### A.4. Hardware Details

All models were run on a cluster across roughly 8 NVIDIA GeForce 1080Ti GPUs, 8 NVIDIA GeForce 980Ti GPUs, and 8 NVIDIA Titan X Gpus. Each model in Table 1 thus required roughly 6-8 GPU days to train to the final number of iterations.

## B. Analytical Treatment of Neural Wave Machines

In this section we extend the analytical treatment of Neural Wave Machines, verifying that the model does indeed inherit many of the same beneficial bounds on hidden state and gradient magnitudes as the original coRNN, as stated in the main text. Specifically, by carefully reviewing the proofs for Proposition 3.1 (bounded hidden state energy) and Proposition 3.2 (bounded hidden state gradients) of Rusch & Mishra (2021), it can be shown that the Neural Wave Machine satisfies the conditions necessary for these bounds to similarly hold with minor modifications. At a high level, the intuition for why these bounds hold is that our convolutional parameterization of the coupling matrices does not change the theoretical bounds on the infinity norm of the weights, the crucial element necessary for bounding these quantities (e.g. see equation (13) of Rusch & Mishra (2021)). In the following, we detail each of these bounds more precisely.

### B.1. Bounds on Hidden State Energy

Identically following the proof of Proposition 3.1, from Section E.1 of Rusch & Mishra (2021), defining the total energy of our model's hidden state as $\mathbf{x}_n^T\mathbf{x}_n + \mathbf{v}_n^T\mathbf{v}_n$, it can be seen this value is bounded at time-step $n$, and with hidden state size $m$, as:

$$\mathbf{x}_n^T\mathbf{x}_n + \mathbf{v}_n^T\mathbf{v}_n \leq \mathbf{x}_0^T\mathbf{x}_0 + \mathbf{v}_0^T\mathbf{v}_0 + nm\Delta t$$

We see that this bound does not change from the original work as the derivation is not dependent on the parameterization of the coupling matrices $\mathbf{W}, \mathcal{W}$. Furthermore, this bound applies equally in the case when we have non-zero initial conditions (as through our initial condition network).

### B.2. Sensitivity to Inputs

From Section E.2, Proposition E.1, of (Rusch & Mishra, 2021), it can be seen that the NWM also inherits a bound on how much differences in inputs are able to change the hidden state. Specifically, since the activation function we use is $\tanh$, our bound is identical. This is the theoretical justification for our comment regarding the NWM's apparent inability to model chaotic dynamics (which we expand on in Appendix C.4).

### B.3. Bounds on Hidden State Gradient

From Section E.3, following the proof of Proposition 3.2, of Rusch & Mishra (2021), we see that, assuming $\alpha = \gamma = 1$, we can again derive bounds on the gradient of the loss with respect to the model parameters. Specifically, the outline of the proof is nearly identical, with only equation (28) being modified to reflect the fact that our parameters are now shared over all spatial locations (due to the convolution). In detail, the matrix $\mathbf{Z}_{m,\bar{m}}^{i,j}$ no longer only has a single non-zero value, but instead $m$ non-zero values equal to $\sigma'(\mathbf{A}_{k-1})_i$ (for an $m$ sized hidden state). We see that when this matrix is then multiplied

with each vector $(\mathbf{x}_{k-1}, \mathbf{v}_{k-1}, \mathbf{u}_k)$, using the bound $||\mathbf{Z}^{i,j}_{m,\bar{m}}(\mathbf{A}_{k-1})||_\infty \leq 1$, the upper bounds in equation (29) change from $||\mathbf{x}_{k-1}||_\infty,\ ||\mathbf{v}_{k-1}||_\infty,\ ||\mathbf{u}_k||_\infty$ to $m||\mathbf{x}_{k-1}||_\infty,\ m||\mathbf{v}_{k-1}||_\infty,\ m||\mathbf{u}_k||_\infty$. Carrying these extra factors of $m$ through the rest of the proof, we arrive at the following final bound on the gradient of the loss function $\xi$ with respect to any parameter $\theta$:

$$\left|\frac{\partial \xi}{\partial \theta}\right| \leq \frac{3}{2}(m + \bar{X}m^{3/2})$$

where $\bar{X} = \max_n ||\bar{x}_n||_\infty$.

### B.4. Assumptions

As with the proofs for the coRNN, the same assumptions are necessary for the bounds to hold. Specifically, it is assumed that $\Delta t$ is chosen such that:

$$\max\Big(\frac{\Delta t(1 + ||\mathbf{W}||_\infty)}{1 + \Delta t}, \frac{\Delta t||\mathcal{W}||_\infty}{1 + \Delta t}\Big) \leq \Delta t^r, \quad \frac{1}{2} \leq r \leq 1 \tag{11}$$

Since this assumption is indeed satisfied throughout training for the original coRNN, we assume that it is likely satisfied with the NWM as well. Intuitively, we find no reason to believe that changing the fully connected matrices $\mathbf{W}$ & $\mathcal{W}$ to convolutional matrices will have the necessary order-of-magnitude impact on the infinity norm of the weight matrices necessary to invalidate this assumption. In preliminary experiments on sMNIST we also find this intuition to hold. Specifically, for the optimal value of $\Delta t = 0.042$, and $r = \frac{1}{2}$, we see that the maximum over training of the quantity of interest (Equation 11) is actually lower for the NWM than the coRNN ($0.157$ vs. $0.188$) with both being lower than the limit ($0.205$).

## C. Extended Results

### C.1. Impact of $\Delta t$ parameter

In this section we include an additional preliminary analysis to measure the impact of changing the $\Delta t$ parameter. In practice, we see that the parameter has an impact not only on the numerical integration, but also on the speed at which the network's hidden state is able to update. Therefore, similar to prior work with the coRNN, we find it best to treat this parameter as a hyperparameter and tune it in addition to the other hypterparameters. In the table below, we show the results of our model on sMNIST for a range of $\Delta t$ values:

Table 3: Test accuracy on the sMNIST dataset for a range of $\Delta t$ values.

| $\Delta t$ | 0.001 | 0.1 | 0.042 | 0.15 | 0.30 | 0.45 |
|---|---|---|---|---|---|---|
| Test Accuracy | 87.7 | 90.6 | 98.4 | 97.5 | 89.8 | NaN |

We see that a moderate value of $\Delta t$ is optimal, while too large causes divergence (perhaps due to excessive discretization errors) and too small disrupts information processing in the RNN.

### C.2. Additional Efficient Sequence Modeling Results

In this section we include additional results comparing the coRNN and NWM on different sequence modeling tasks. Specifically, we show model performance on the long-sequence addition task initially introduced by Hochreiter & Schmidhuber (1997), and the IMDB sentiment classification task (Rusch & Mishra, 2021). On both datasets we see that the NWM achieves comparable performance to the coRNN while requiring significantly fewer parameters, in line with results on the sMNIIST and psMNIST datasets.

### C.3. Additional Hamiltonian Dynamics Results

In this section we include an alternative metric for measuring model forecasting performance on the Hamiltonian Dynamics Suite. Specifically in Table 5, we report the 'Valid Prediction Time' as reported in prior work (Botev et al., 2021), defined as the number of time steps into the future the models are able to accurately predict the dynamics of the system with reconstruction error under a predefined threshold (MSE $< 0.025$). Given the high variance of the VPT value from batch-to-batch, the values reported in Table 5 are computed as the mean and standard deviation of the VPT over the final 5

Table 4: Test accuracy on additional sequence modeling benchmarks including the long-sequence Addition task from Hochreiter & Schmidhuber (1997), and the IMDB sentiment classification task. All results are mean ± std. over 3 random initalizations. We see similar results to those shown in Table 4, the NWM achieves comparable performance while requiring significantly fewer parameters.

| | Adding Task | | IMDB | |
|---|---|---|---|---|
| | Accuracy | $\#\theta$ | Accuracy | $\#\theta$ |
| coRNN | $0.0035 \pm 0.01$ | 131k | $86.4 \pm 0.2$ | 46k |
| NWM | $0.0046 \pm 0.0016$ | <1k | $86.1 \pm 0.3$ | 13k |

evaluation iterations. We see that the values roughly agree with those reported in (Botev et al., 2021), however certain discrepancies may still appear due to the fact that the authors of (Botev et al., 2021) only report the range of the grid search they performed but not the actual hyperparameter values of their best performing models. Further, we see that the ranking of model performance under this metric is quite noisy due to the high variance of the metric. We therefore urge future work to consider alternative benchmarks and metrics for evaluating the forecasting performance of such models.

Table 5: Valid Prediction Time 'VPT' (± std.) on the Hamiltonian Dynamics Benchmark. We highlight in bold results which fall within one standard deviation of the best performing model. We see that the VPT metric has large standard deviation owing to the reliance on an arbitrary threshold of image-space similarity, however the NWM models still perform favorably compared with existing state of the art.

| | AR | HGN++ | ODE [TR] | coRNN | NWM 2D | NWM 1D |
|---|---|---|---|---|---|---|
| Spring | 302 (63) | **447 (0)** | **430 (26)** | 375 (14) | 311.8 (27) | **431 (24)** |
| Pendulum | 3 (4) | 105 (21) | **212 (65)** | **179 (91)** | 155.1 (24) | **174 (65)** |
| Two Body | 263 (92) | **444 (3)** | **439 (11)** | **431 (40)** | **413 (53)** | **420 (27)** |
| Pennies | 118 (25) | 79 (6) | **164 (14)** | **165 (23)** | **141 (37)** | **163 (9)** |
| Double Pendulum | 0 (0) | 11 (5) | **22 (7)** | 3 (1) | 9 (9) | **10 (8)** |

### C.4. On Modeling Chaotic Dynamics

In this section, we include an extended evaluation to investigate the apparent inability of the NWM to model more chaotic dynamics such as the double pendulum task. To do this, we perform an analogous experiment to that reported in Appendix A of the original coRNN work (Rusch & Mishra, 2021). Specifically, we measure the ability of our model to predict the state of a system at a fixed 25-time steps ahead for a Lorentz '96 attractor ($x'_j = (x_{i+1} - x_{i-2})x_{i-1} - x_i + F$). Here, $F$ is an external force which controls how chaotic the trajectories are, where $F = 8$ corresponds to a highly caotic trajectory and $F < 1$ is significantly less chaotic. Ultimately, we see that, similar to the original coRNN work, the LSTM performs significantly better than the NWM in the chaotic regime, providing empirical evidence for the theoretical claim that the coupled oscillator networks are unable to model chaotic dynamics.

Table 6: Test Mean Squared Error of an LSTM and NWM when forecasting the Lotentz '96 attractor. We see that the NWM performs better in the non-chaotic regime ($F = 0.9$), while in chaotic regime ($F = 8$) the LSTM performs significantly better.

| Model | $F = 0.9$ | $F = 8.0$ |
|---|---|---|
| LSTM | $5.2 \times 10^{-3}$ | $1.9 \times 10^{-2}$ |
| NWM | $2.4 \times 10^{-3}$ | $4.8 \times 10^{-2}$ |

### C.5. On the Formation of Orientation Maps

Although there is significant prior work which can give intuition as to why the smooth orientation selectivity maps of Figure 3 may arise from our model, we believe we are the first to demonstrate a system which actually learns these types of maps from data in the service of sequence modeling. At the highest level, the intuition for the mechanism behind these maps can be seen to come from the combination of phase-synchrony of coupled oscillator systems, and the necessity to model temporally correlated transformations. Extensive prior work on so-called 'phase-reduced' Kuramoto models demonstrates the emergence of complex spatiotemporal patterns such as plane waves, spirals, and pinwheel lattices. Examples include early work from

Ermentrout et al. (1970) (Figure 6), showing various steady state phase relationships in the solutions of the locally coupled oscillator dynamics. Similarly, more recent works, (Jeong et al., 2002) (Figs. 3 & 4) and (Breakspear et al., 2010) (Figs. 5 & 6) have studied how this phase-locking can vary for different types of chosen couplings. Given that these phase-reduced systems are theoretical approximations to the more flexible (non-reduced) oscillator dynamics implemented in the NWM, it makes sense that we also see these types of phase relationships (e.g. Fig. 4 of the main text). When such complex phase-synchrony is combined with the task of sequence modeling, the synchrony can be seen to essentially be inducing local correlations between neurons for each time-step. Thus, when the training set contains input at a variety of different angles, and the model is required to represent these over time, the intuition follows that there will be spatially-smooth orientation selectivity corresponding to these induced correlations. In Figure 6 we provide some quantitative measurements which align with this intuition. Specifically, the figure shows the instantaneous phase measurement of each neuron (right) next to the orientation selectivity of the same neurons (left). As can be seen, there is a rough correlation between phase values and orientation selectivity, with unexplained variance likely arising due to computing the depicted instantaneous phase values from a single training example, while selectivity measurements are computed over an entire dataset. Furthermore, in Figure 7 we show how different hyperparameters affect the size of the resulting learned orientation columns. We see that both the wavelength of the training dataset ($\lambda^{train}$ of sine waves) and the kernel size ($\text{size}(\mathbf{w}_z)$) have a direct increasing relationship with the size of the learned orientation columns, suggesting these parameters could be tuned to better fit observations from neuroscience.

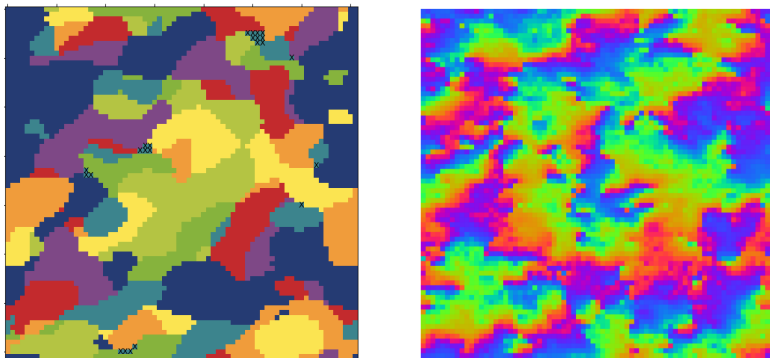

Figure 6: Orientation selectivity (left) and instantaneous phase at a random sequence element (right) for a model trained on the sine waves dataset. We see that the phase synchrony across the neurons is roughly in alignment with the orientation selectivity, supporting the hypothesis that this is one of the primary mechanisms for topographic organization in the NWM.

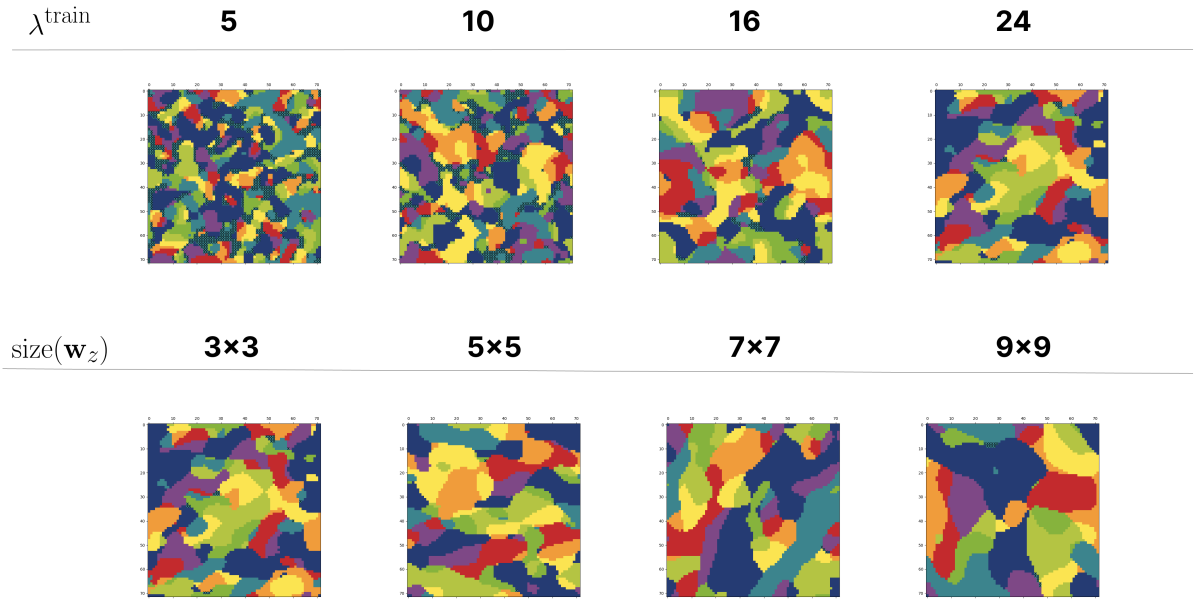


Figure 7: Orientation selectivity maps as a function of training dataset wavelength ($\lambda^{train}$), and kernel size ($\text{size}(\mathbf{w}_z)$).

### C.6. Full Rotating MNIST Topographic Organization

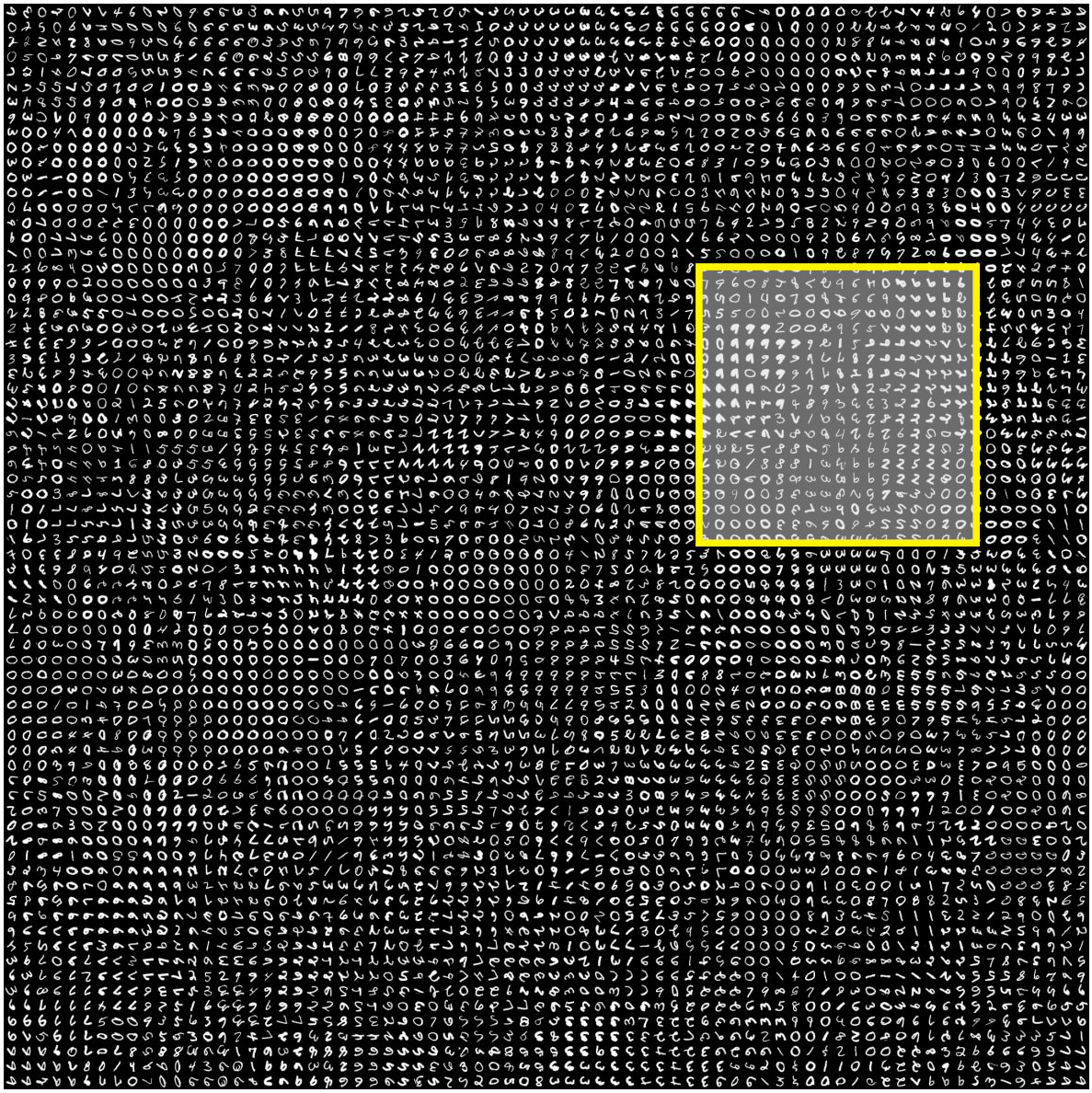

Figure 8: Depiction of the maximum activating image for the full set of neurons in the NWM when training on Rotated MNIST. The subset depicted in Figure 3 is highlighted in yellow. We see that topographic organization is widespread and roughly continuous throughout the hidden state.

### C.7. Visualizing Traveling Waves on MNIST

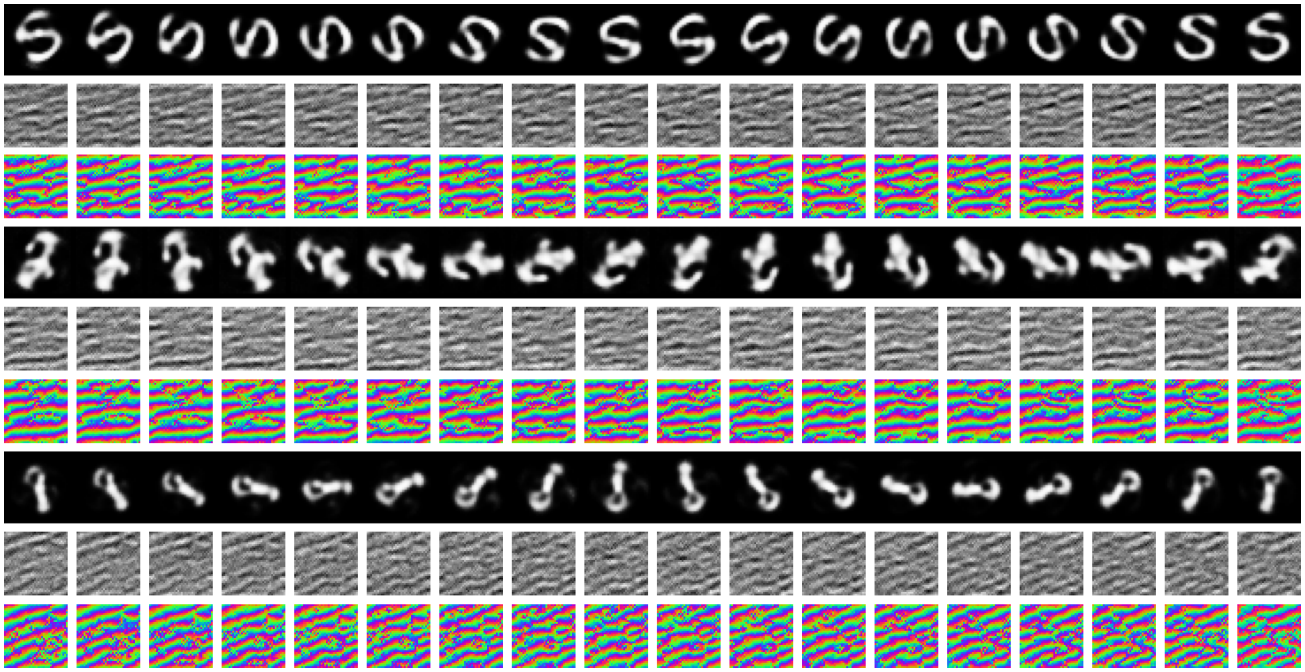

Figure 9: Additional hidden state visualizations for the model in Figure 4. Reconstructions (Top), Hidden state (middle) and generalized phase (bottom), for the final 18 timesteps of the test sequence.

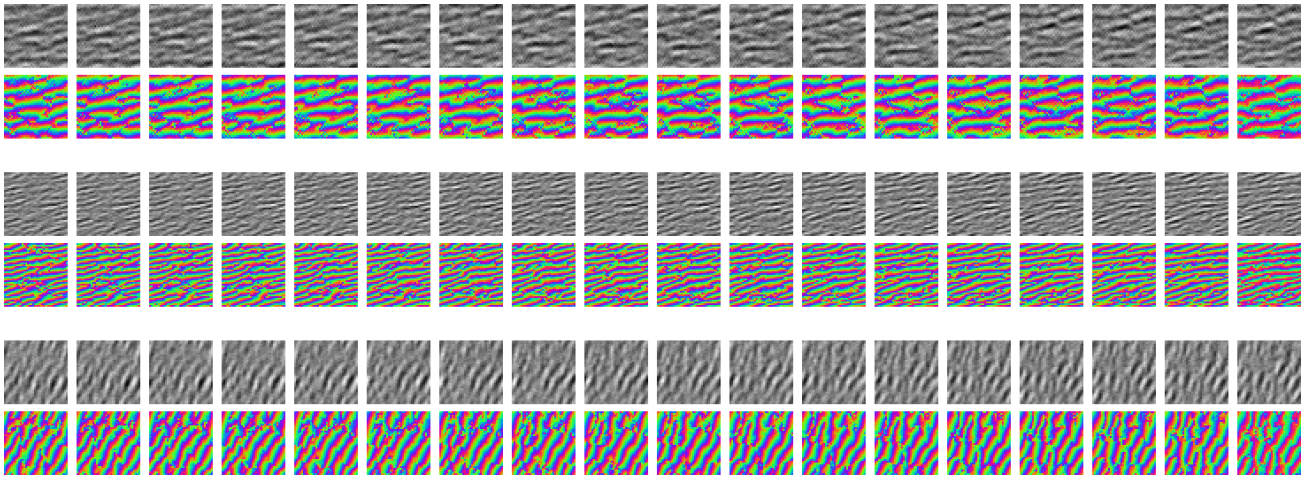

Figure 10: Visualization of the hidden state and phase for three models identical to those in Figure 4, but with different random initalizations. We see that the models learn different wavelengths and velocities depending on their initialization.

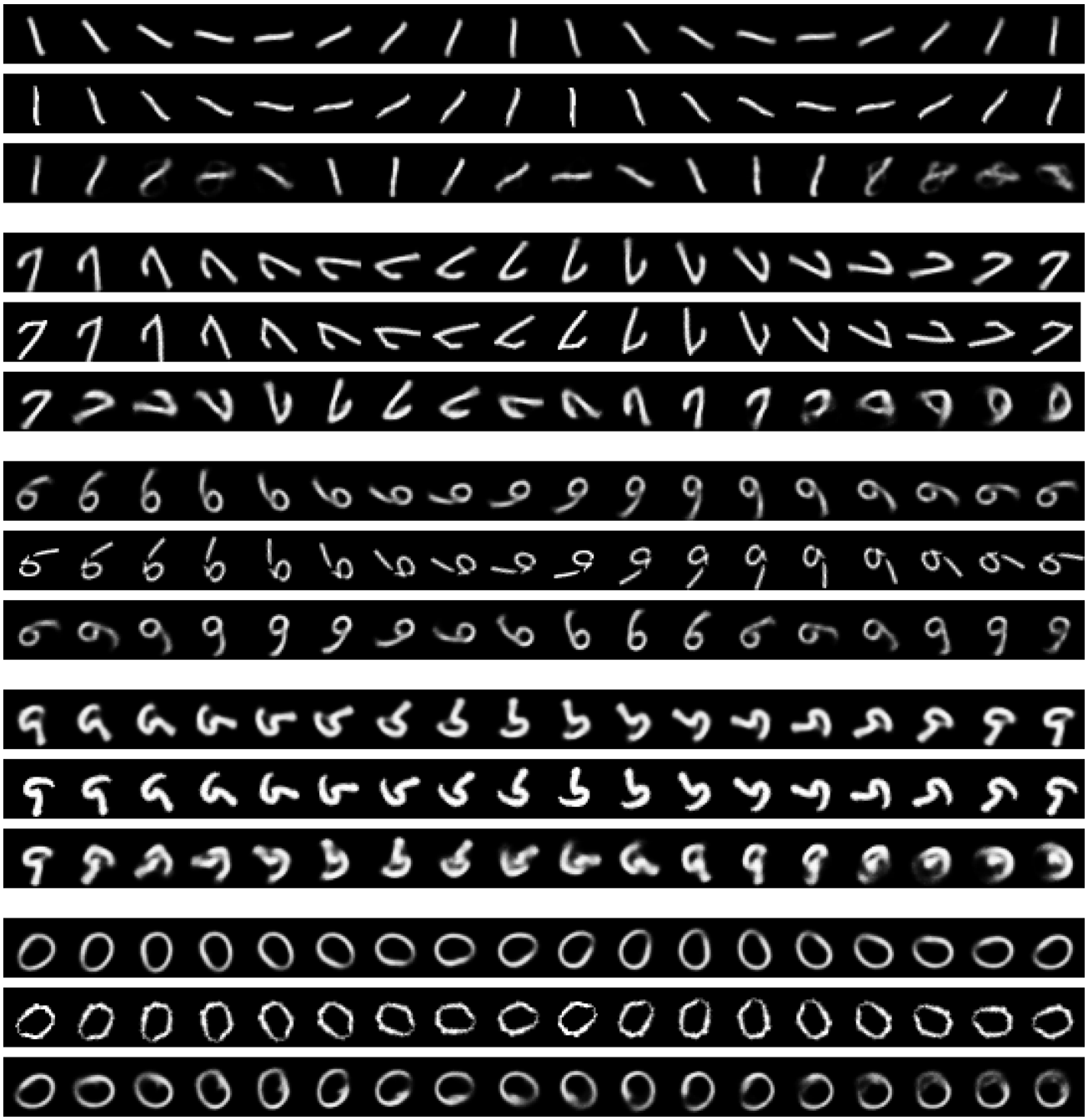

Figure 11: Additional visualizations of reconstructions from induced wave activity in the hidden state of the 1D NWM as depicted in Figure 5. We show a set of random input sequences (top), the original model reconstruction (middle), and images generated by sequentially propagating the initial state backwards by an induced wave and decoding at each step (bottom). We see that, as in the main text, the assumed wave velocity of $v = 1$ is slightly faster than the actual velocity, and thus the reconstructed transformations are slightly faster than the input transformations. Because of this, we also observe that for certain examples, the induced wave reconstructions lose consistency with the input after the first period. This appears to imply that both the initial location of the wave activity matters in addition to its wave properties, and thus our model has learned to only propagate waves over parts of the feature space to optimize the capacity of the hidden state for this dataset. Finally, we observe that the induced transformations occur in reverse order due to the fact that our induced waves propagate in the reverse direction to those naturally exhibited for training examples, effectively propagating backwards in time.