Weight-Space Linear Recurrent Neural Networks

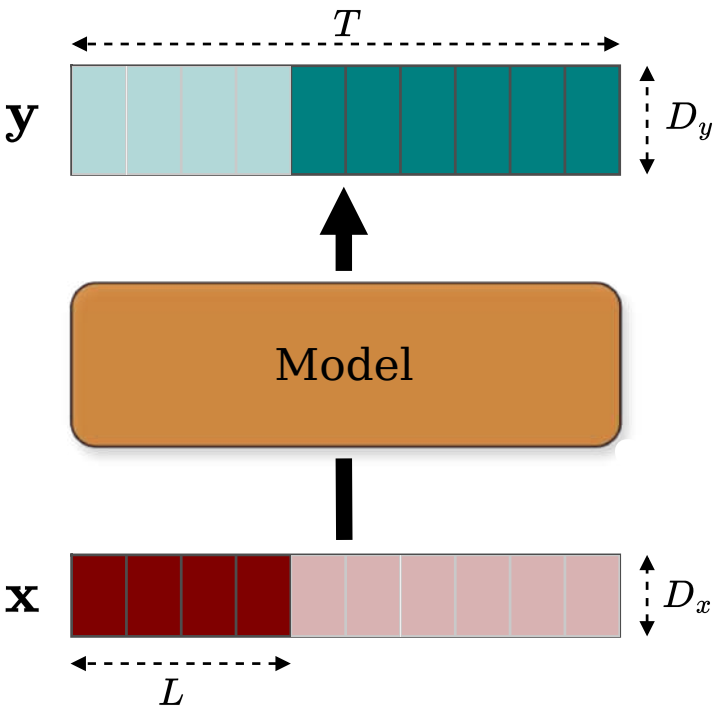
Roussel Desmond Nzoyem & Enrique Crespo-Fernandez

(Co-authors: Nawid Keshtmand, Idriss Tsayem, Raul Santos-Rodriguez, David AW Barton, and Tom Deakin)

Problem setting and challenges

Problem setting

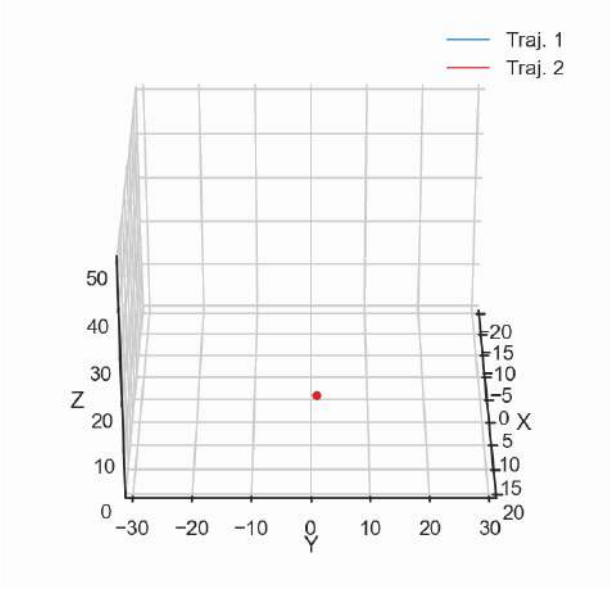
Sequence-to-sequence modelling seeks to capture sequential **dynamics**.



e.g., initial value problem, classification, forecasting

Challenges :

- ✗ Non-linear sequential dynamics
- ✗ Unreliable for long-term forecasts
- ✗ Computationally challenging

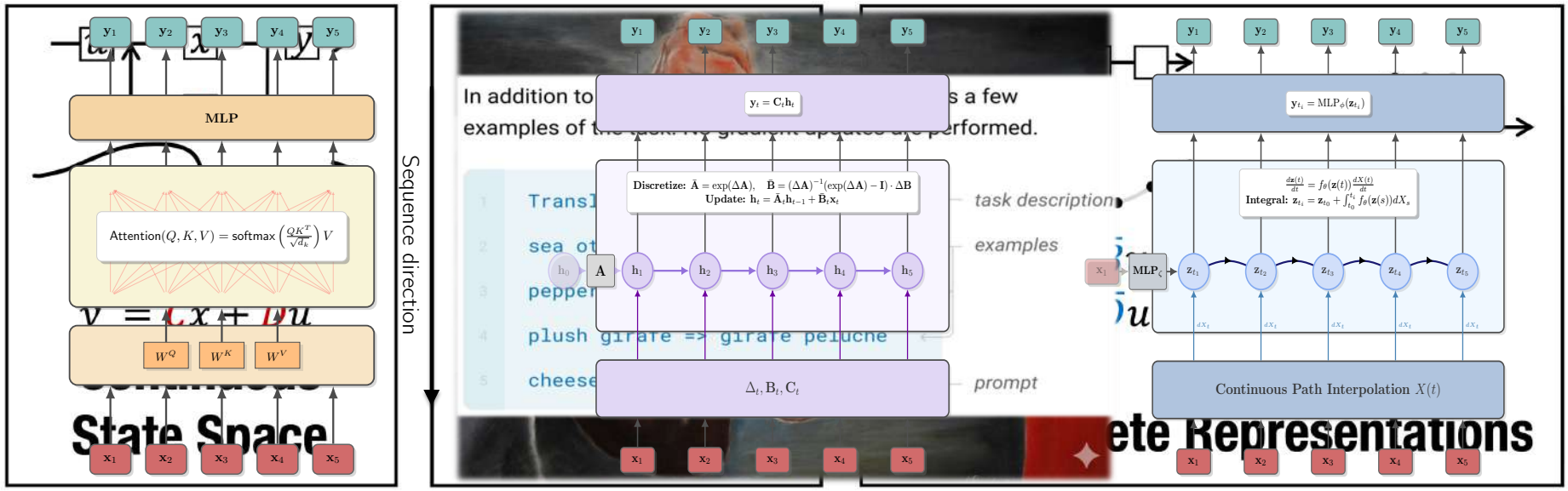


Example: Lorenz chaotic behaviour

Desiderata for next-generation deep sequence modelling

- **Non-linearities** for expressiveness
- **Linearity** for efficiency and interpretability
- **Gradient-free** online adaptation
- **Physics-informed** machine learning

Softmax-Attention | Mamba | Neural Controlled Diff. Eq.

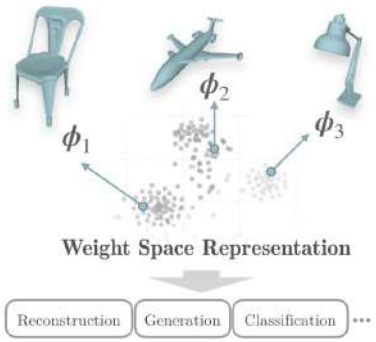


Vaswani et al, "Attention Is All You Need", NeurIPS 2017
Gu and Dao., "Mamba: Linear-Time Sequence Modeling with Selective State Spaces", COLM 2024
Kidger et al. "Neural Controlled Differential Equations for Irregular Time Series", NeurIPS 2020
Beck et al. 2025, "xLSTM: Extended Long Short-Term Memory", NeurIPS 2024

Weight-Space Learning (WSL) offers a path forward !

WSL views the **weights and biases** of a neural network (NN) as **data points** for another learning system.

(e.g., INRs, hypernetworks, optimisers, etc.)



Why can WSL help?

- NNs are universal approximators
- Linearly updating NN weights can result in expressive non-linear functions in the observation space
- NN weights can be adapted in-context, approximating GD
- We can choose the model family based on physics priors

Schürholt, "Towards Scalable and Versatile Weight Space Learning", ICML 2024
Toderici, "Full Resolution Image Compression with Recurrent Neural Networks", CVPR 2017

Weight-space linear RNNs for sequence-to-sequence modelling (WARP)

WARP blends **linear** recurrence with **non-linear** decoding.

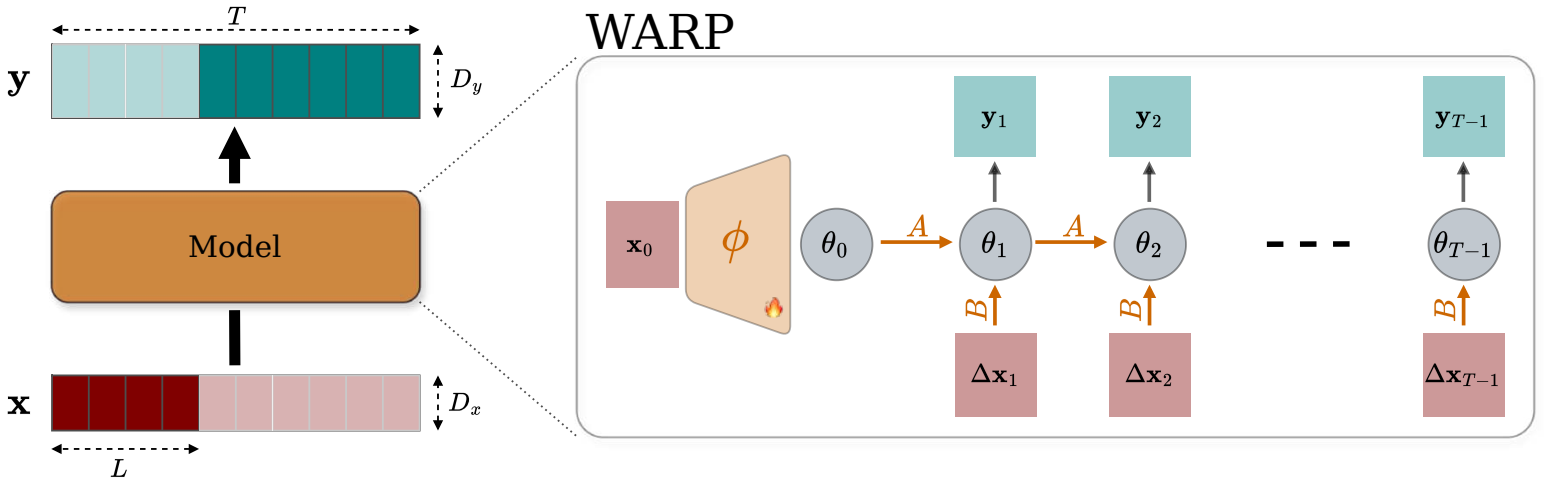
Standard RNNs

$$\mathbf{h}_t = f_\Phi(\mathbf{h}_{t-1}, \mathbf{x}_t)$$
$$\mathbf{y}_t = g_\Psi(\mathbf{h}_t)$$

Weight-Space Linear RNNs

$$\theta_t = A\theta_{t-1} + B(\mathbf{x}_t - \mathbf{x}_{t-1})$$
$$\mathbf{y}_t = \mathrm{MLP}_{\theta_t}(\tau)$$

τ is positional encoding:
- Normalised time t
- Pixel coordinates
- Sine/Cosine waves



Nzoyem et al., "Weight-Space Linear Recurrent Neural Networks", ICLR 2026

SoTA classification on time series benchmarks

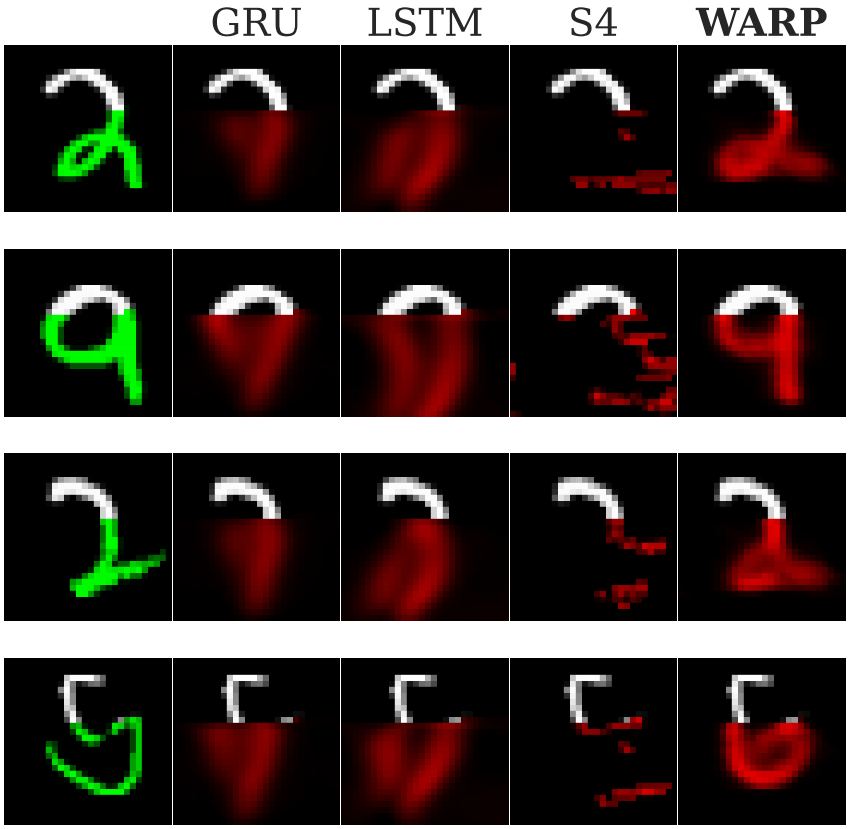
Results on UEA classification

	Worms	**SCP1**	**SCP2**	**Ethanol**	**Heartbeat**	**Motor**
Seq. length	17,984	896	1,152	1,751	405	3,000
# Classes	5	2	2	4	2	2
NRDE	77.2 ± 7.1	76.7 ± 5.6	48.1 ± 11.4	31.4 ± 4.5	73.9 ± 2.6	54.0 ± 7.8
NCDE	62.2 ± 3.3	80.0 ± 2.0	53.6 ± 6.2	22.0 ± 1.0	68.1 ± 5.8	51.6 ± 6.7
LRU	85.0 ± 6.2	84.5 ± 4.6	47.4 ± 4.0	29.8 ± 2.8	78.1 ± 7.6	51.9 ± 8.6
S5	83.9 ± 4.1	87.1 ± 2.1	55.1 ± 3.3	25.6 ± 3.5	73.9 ± 3.1	53.0 ± 3.9
Mamba	70.9 ± 15.8	80.7 ± 1.4	48.2 ± 3.9	27.9 ± 4.5	76.2 ± 3.8	47.7 ± 4.5
S6	85.0 ± 1.2	82.8 ± 2.7	49.9 ± 9.4	26.4 ± 6.4	76.5 ± 8.3	51.3 ± 4.2
Log-NCDE	82.8 ± 2.7	82.1 ± 1.4	54.0 ± 2.6	35.9 ± 6.1	74.2 ± 2.0	57.2 ± 5.6
LinOSS	$\mathbf{95.0 \pm 4.4}$	$\mathbf{87.8 \pm 2.6}$	58.2 ± 6.9	29.9 ± 0.6	75.8 ± 3.7	$\mathbf{60.0 \pm 7.5}$
FACTS	86.7 ± 3.0	73.3 ± 2.8	$\mathbf{70.3 \pm 8.8}$	28.2 ± 3.3	70.3 ± 8.8	49.8 ± 3.8
Griffin	79.5 ± 5.1	80.0 ± 1.5	43.1 ± 5.3	24.0 ± 3.5	77.7 ± 2.9	43.8 ± 3.3
WARP	70.93 ± 2.7	83.53 ± 2.0	57.89 ± 1.4	$\mathbf{36.49 \pm 2.8}$	$\mathbf{80.65 \pm 1.9}$	56.14 ± 5.1

Nzoyem et al., "Weight-Space Linear Recurrent Neural Networks", ICLR 2026
Bagnall et al. "The UEA multivariate time series classification archive, 2018", 2018

Image completion as sequence modelling

MNIST with (300 initial context steps)

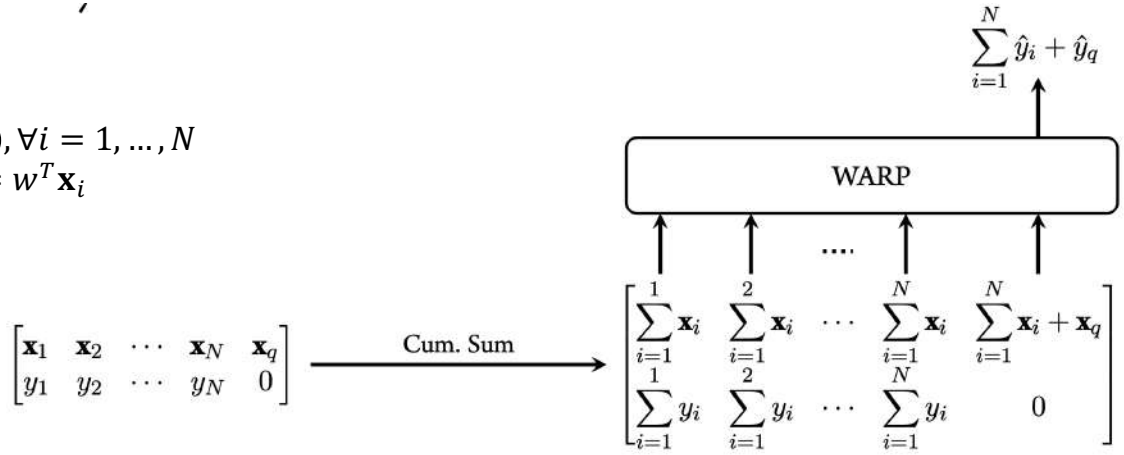


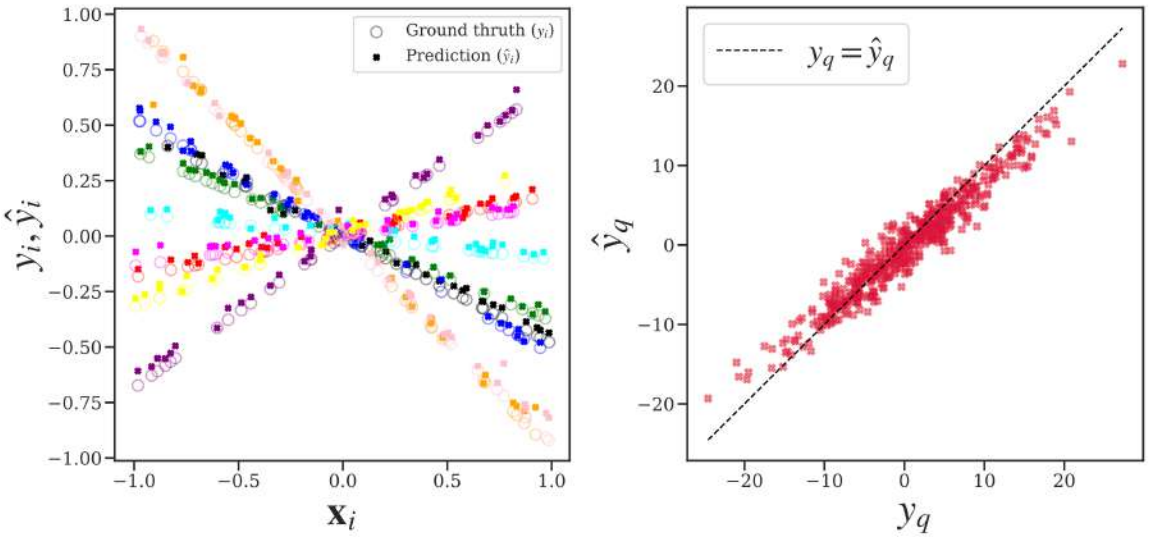
Nzoyem et al., "Weight-Space Linear Recurrent Neural Networks", ICLR 2026

In-context learning for linear regression

Each input sequence:

- is a matrix of shape $(D_{\mathbf{x}} + 1, N + 1)$
- keys randomly sampled $\mathbf{x}_i \sim \mathcal{U}(-1, 1), \forall i = 1, \ldots, N$
- scalar values via linear mapping $y_i = w^T \mathbf{x}_i$
- defined by one vector $w \sim \mathcal{U}(-1, 1)$





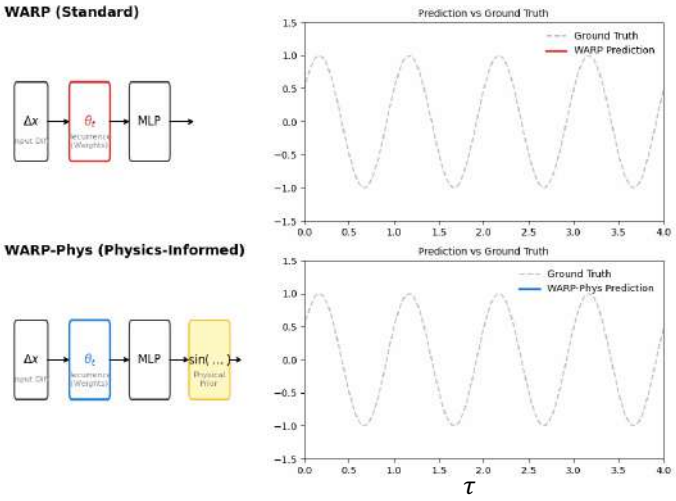
Nzoyem et al., "Weight-Space Linear Recurrent Neural Networks", ICLR 2026

Physics-informed ML: leveraging continuous dynamics

Useful for:

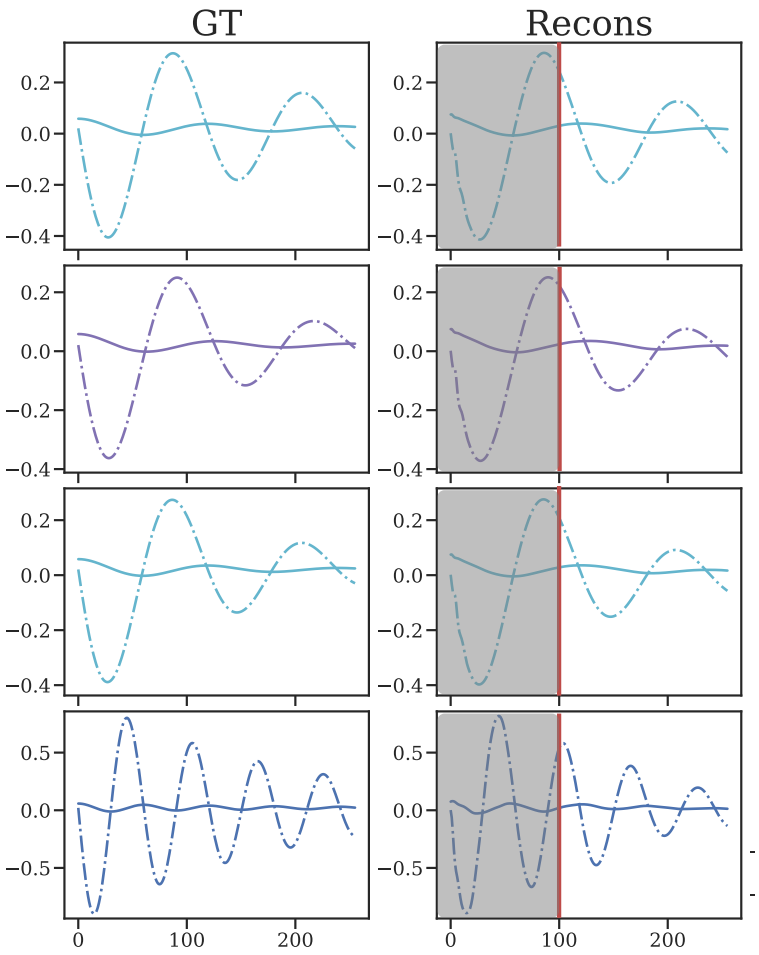
- improved (generalisation) accuracy
- better sample efficiency

$$\tau \mapsto \sin(2\pi\tau + \hat{\varphi})$$



	MSD		MSD-Zero		LV		SINE*	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
GRU	1.43 ± 0.09	5.05	0.55± 0.75	3.27± 0.13	5.83± 0.37	13.1± 0.42	4.90± 0.45	179± 9.23
LSTM	1.46 ± 0.14	5.43 ± 0.28	0.57± 0.05	3.46±0.08	6.18 ± 0.19	13.6 ± 0.61	9.48± 0.12	248± 3.45
Transformer	0.34± 0.12	2.25 ± 0.42	0.48± 0.24	2.90± 0.32	11.27± 0.62	18.6± 0.49	1728± 10.8	2204± 27.0
WARP	0.94 ± 0.09	3.04 ± 0.11	0.32± 0.02	2.59± 0.07	**4.72** ± 0.25	**10.9**± 0.45	2.77± 0.09	125± 8.46
WARP-Phys	**0.03**±0.04	**0.66**± 0.02	**0.04**± 0.01	**0.75**± 0.03	X	X	**0.62**± 0.01	**6.47**± 0.51

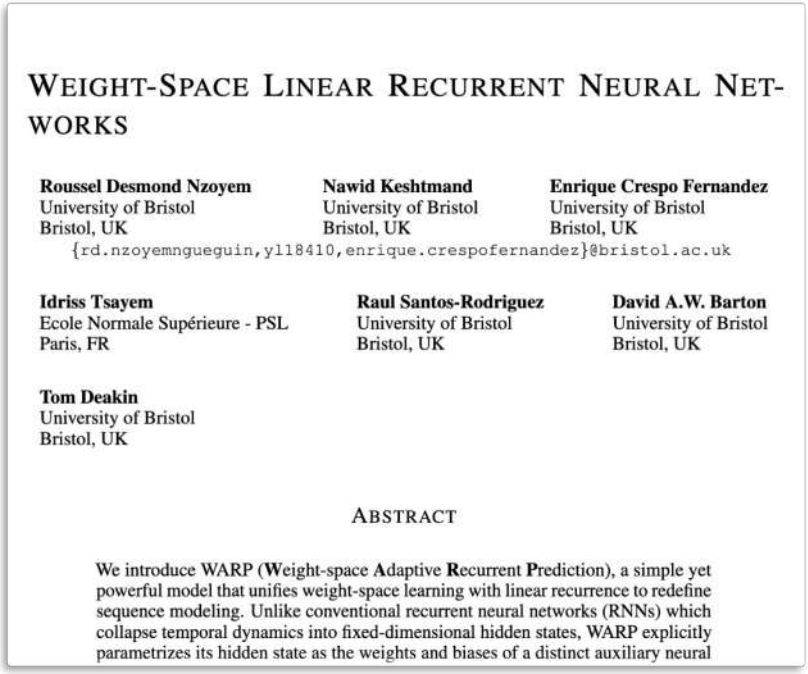
Mass-Spring-Damper (100 initial steps)



Summary & Future Work

- Weight-Space Linear RNNs meet requirements:
  - Trains like a **linear** model, performs like a **non-linear** one
  - Driven by an innovative **input difference**
  - Capable of in-context learning
  - Capable of physics-informed machine learning

- Future Work:
  - Addressing **theory** and **neuron equivariance**
  - Eigen-decomposition and parameter efficiency



WEIGHT-SPACE LINEAR RECURRENT NEURAL NETWORKS

Roussel Desmond Nzoyem
University of Bristol
Bristol, UK

Nawid Keshtmand
University of Bristol
Bristol, UK

Enrique Crespo Fernandez
University of Bristol
Bristol, UK

{rd.nzoyemngueguin,yl18410,enrique.crespofernandez}@bristol.ac.uk

Idriss Tsayem
Ecole Normale Supérieure - PSL
Paris, FR

Raul Santos-Rodriguez
University of Bristol
Bristol, UK

David A.W. Barton
University of Bristol
Bristol, UK

Tom Deakin
University of Bristol
Bristol, UK

ABSTRACT

We introduce WARP (**W**eight-space **A**daptive **R**ecurrent **P**rediction), a simple yet powerful model that unifies weight-space learning with linear recurrence to redefine sequence modeling. Unlike conventional recurrent neural networks (RNNs) which collapse temporal dynamics into fixed-dimensional hidden states, WARP explicitly parametrizes its hidden state as the weights and biases of a distinct auxiliary neural

Paper: https://arxiv.org/abs/2506.01153

Code: https://github.com/ddrous/warp

Reach out:

–rd.nzoyemngueguin@bristol.ac.uk

–enrique.crespofernandez@bristol.ac.uk

Thank You 😀