Adaptive Programmable Networks for In Materia Neuromorphic Computing

Kilian D. Stenning1,2*, Jack C. Gartside1,†, Luca Manneschi3,†, Christopher T. S. Cheung1, Tony Chen1, Alex Vanstone1, Jake Love4,5, Holly H. Holder1, Francesco Caravelli6, Karin Everschor-Sitte4, Eleni Vasilaki3, and Will R. Branford1

1Blackett Laboratory, Imperial College London, London SW7 2AZ, United Kingdom
2London Centre for Nanotechnology, University College London, London WC1H 0AH, United Kingdom
3University of Sheffield, Sheffield S10 2TN, United Kingdom
4Faculty of Physics, University of Duisburg-Essen, 47057 Duisburg, Germany
5Institute of Physics, Johannes Gutenberg University Mainz, 55128 Mainz, Germany
6Theoretical Division (T4), Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA
†These authors contributed equally
*Corresponding author e-mail: k.stenning18@imperial.ac.uk

ABSTRACT

Modern AI and machine-learning provide striking performance. However, this comes with rapidly-spiralling energy costs1,2 arising from growing network size and inefficiencies of the von Neumann architecture. ‘Reservoir computing’3–10 offers an energy-efficient11 alternative to large networks, fixing randomised weights for energetically-cheap training. The massively-parallel processing underpinning machine-learning is poorly catered for by CMOS, with in materia neuromorphic computing an attractive solution3,4,12–16.

Nanomagnetic artificial spin-systems are ideal candidates for neuromorphic hardware. Their passive memory, state-dependent dynamics and nonlinear GHz spin-wave response provide powerful computation8. However, any single physical reservoir must trade-off between performance metrics including nonlinearity and memory-capacity10,17–20, with the compromise typically hard-coded during nanofabrication. Here, we present three artificial spin-systems: square artificial spin ice, square artificial spin-vortex ice and a disordered pinwheel artificial spin-vortex ice. We show how tuning system geometry and dynamics defines computing performance.

We engineer networks where each node is a high-dimensional physical reservoir20–27, implementing parallel, deep and multilayer physical neural network architectures. This solves the issue of physical reservoir performance compromise, allowing a small suite of synergistic physical systems to address diverse tasks and provide a broad range of reprogrammable computationally-distinct configurations. These networks outperform any single reservoir across a broad taskset. Crucially, we move beyond reservoir computing to present a method for reconfigurably programming inter-layer network connections, enabling on-demand task optimised performance.

Physical computing has been implemented across a number of physical systems including memristor5, 28–33, optical34 and nanomagnetic systems7–10, 35–43 demonstrating the performance, scalability and energy savings of a physical hardware approach11,44. However, memristors must overcome memory volatility and physical degradation45 before technological implementation is feasible. Optical neural networks offer speed-of-light computation but memory must be engineered via optical delay-lines, limiting reconfigurability and occupying large on-chip footprints.

Nanomagnetic systems are attractive physical computing candidates, offering solutions to memory volatility and physical degradation. The magnetic configuration (microstate) possesses passive long-term memory. Nanomagnets interact wirelessly through dipolar coupling, allowing spin-wave information transfer with no electron movement and waste heat14, 15, 46–51. Long-range coupling provides high-dimensional, massively-parallel nonlinear processing which is sensitive to the local microstate14, 52, a feature which must be engineered-in at-cost in other approaches. Nanomagnets switch at ns timescales and spin-waves (magnons) offer GHz computing15, 46, 49, 50, 53–60.

Until recently, nanomagnetic computation networks were limited to small spin-torque oscillator arrays7, 36, largely due to the engineering difficulties of spatial-domain readout or fabricating individual current-address lines for each element. We recently engineered a readout solution, ferromagnetic resonance (FMR) providing frequency-domain multi-dimensional output3, allowing state-sensitive measurements52, 59 of large networks (108 nanomagnets). We demonstrated reservoir computing (RC) in an ‘artificial spin-vortex ice’, exploiting the nonlinear spin-wave dynamics52, 59, 61–66 and long-term ‘fading memory’67.
List of Abbreviations

| Abbreviation | Description |
|--------------|-------------|
| ASR          | Artificial Spin Reservoir |
| FMR          | Ferromagnetic Resonance |
| MC           | Memory-Capacity |
| sMC          | Signal Memory-Capacity |
| MS           | Macrospin Artificial Spin Ice |
| MSE          | Mean Squared Error |
| NARMA        | Nonlinear Auto-Regressive Moving Average |
| NL           | Nonlinearity |
| sNL          | Signal Nonlinearity |
| PNN          | Physical Neural Network |
| PW           | Pinwheel Artificial Spin-Vortex Ice |
| RC           | Reservoir Computing |
| SEM          | Scanning Electron Micrograph |
| WM           | Width-Modified Artificial Spin-Vortex Ice |

provided by magnetic vortex-formation. The effect of array-geometry and magnetisation dynamics in determining computation performance is still unknown. The parameter space is huge, with variables including nanoisland dimensions, array geometry and inter-island coupling strength strongly influencing RC performance.

The efficacy of any single physical reservoir is typically limited by the fundamental compromise between mutually-exclusive metrics defining RC performance: memory-capacity (MC) and nonlinearity (NL)\textsuperscript{10,17–20}. Theoretical work has shown combining multiple reservoirs with differing metrics in parallel and deep networks\textsuperscript{20–27} can significantly improve performance. Parallel networks have been physically implemented\textsuperscript{6,68,69}, however, they lack connectivity for inter-reservoir information transfer - limiting performance. Translating deep networks to physical systems is nontrivial and so-far unrealised due to the large number of possible inter-layer configurations and long nanofabrication and data-acquisition times between iterations.

Here, we present solutions to key outstanding problems in the physical RC field: how physical system design governs computation performance, how combining physical reservoirs in networks can improve performance, and how to efficiently program inter-layer network connections.

We fabricate and evaluate performance and metrics of three distinct artificial spin reservoirs (ASRs) (Figure 1). These ASRs show broadly differing MC and NL metrics and signal-transformation and chaotic time-series prediction task performance, linked to their underlying diverse physical properties and dynamics. Gradual vortex nucleation enhances memory and prediction performance, the spectral richness of disordered arrays generates high NL and strong signal-transformation performance, improving mean squared error (MSE) values up to 31.6 × versus conventional artificial spin ice (Figure 2 f).

We then construct parallel and deep networks where each network node is a high-dimensional ASR. Parallel networks excel at complex transformation tasks by combining unique memory and non-linear responses into one network - achieving up to 4.4 × MSE improvement (Figure 3 h). Parallel networks do not exhibit significantly enhanced MC as no information flows between reservoirs. Deep networks solve this, providing enhanced memory and future-prediction performance with MSEs up to 4 × lower versus the best single reservoir (Figure 4 f).

Crucially, we describe a method for intelligently programming network interconnections. Specific FMR output channels are selected as input for the next network layer. The MC and NL of chosen interconnection channels are strongly correlated with the overall network metric enhancement, enabling on-demand reprogrammable performance with no iterative refinement. This approach is powerful and can be implemented across any physical system. We demonstrate the strength of this programming by implementing a multilayer physical neural network where each neuron is a highly-dimensional recurrent and nonlinear physical reservoir (Figure 5). The multilayer physical neural network substantially outperforms all other network designs for time-series prediction with MSE improvement up to 5.1 × and demonstrates strong performance across signal-transformation tasks. The innovations described here move physical RC beyond the limits of single physical systems towards a next-generation of powerful, versatile computational networks combining the synergistic strengths of multiple physical systems.

### Artificial spin reservoirs

The ASRs in this work are based on square and pinwheel artificial spin ice\textsuperscript{70}. Three samples were fabricated via electron beam lithography lift-off process: MS (Figure 1 a-f), WM (Figure 1 g-l), and PW (Figure 1 m-r). Samples are 25 nm thick Ni\textsubscript{80}Fe\textsubscript{20}. Figure 1 shows scanning electron micrographs (SEM, a,g,m), field-swept ferromagnetic resonance (FMR) spectra when the sample is prepared in an AC-demagnetised state (b,h,n), and spectral evolution when subject to a sinusoidal field-input (c,i,o). Each input time-step corresponds to a minor field loop, i.e. +H\textsubscript{app} then -H\textsubscript{app}, with spectra recorded at -H\textsubscript{app}. Samples are designed to give a varying degree of magnetic and structural complexity, producing a diverse set of spectral responses and computational performance.

MS is a square artificial spin ice. Bars are high-aspect-ratio (530 nm × 120 nm) and only support macrospin states\textsuperscript{8} (Figure 1 d,e). MS FMR spectra comprises two dominant modes 10-11 GHz with opposite linear gradients corresponding to macrospins aligned parallel (positive gradient) and anti-parallel (negative) to H\textsubscript{app} (Figure 1 b,c)). When applying a sine input, two modes appear and vanish out-of-phase with each other as the microstate transitions from saturated (higher field-values, 39-42 mT) to
Figure 1. Artificial spin reservoirs. a-f) Square artificial spin ice with only macrospin states (MS). g-l) Width-modified square artificial spin-vortex ice (WM) comprising two subsets of nanomagnets (one wide, one thin). Wide-bars host macrospins and vortices, thin bars just macrospins. m-r) Pinwheel artificial spin-vortex ice (PW). Bar dimensions vary from fully disconnected to partially connected, giving both magnetic and structural complexity. Bars host macrospins and vortices. SEM images are shown in a), g) and m). b,h,n) Ferromagnetic resonance (FMR) heatmaps measured after AC-demagnetisation. MS (b) shows linear macrospin modes. WM (h) shows linear macrospin modes (6-7 GHz wide-bar, 8-9 GHz thin bar) and non-linear vortex modes. PW (n) exhibits rich linear and non-linear modes. c,i,o) FMR spectral evolution from a sinusoidal field-series input. Each time-step is a minor field loop. MS (c) shows two modes corresponding to bars aligned with and against the applied field. WM (i) shows two dominant modes corresponding to the macrospin resonances. Modes shift linearly. Wide-bar mode amplitude changes non-linearly and lags w.r.t the applied field due to macrospin/vortex conversion. PW (o) displays a range of modes with non-linear frequency-shifts and amplitude changes. MFM images taken at remanence after field-saturation (d,j,p) and AC-demagnetisation (e,k,q). When saturated, all samples contain only macrospins. When demagnetised, WM and PW (j,o) show vortices. MS (f), WM (l) and PW (r) FMR amplitudes at maximum (orange) and minimum (blue) fields of a chaotic time-series (Mackey-Glass) input. Signal memory (sMC$_n$) evaluated on each output frequency-channel across previous time-steps (n) 0-7. sMC (black) and nonlinearity (sNL) (red) across all output frequencies is shown. sMC is a sum of sMC$_n$ from n = 0-7. For MS (d) and WM (j), sMC is highest at FMR modes and subharmonics. sNL is highest at maximum and minimum points of modes and zero-amplitude regions. For PW p), sMC is highest away from the main FMR mode (9-10 GHz). High sNL is observed across the majority of the main resonant peak.
demagnetised (lower fields 35-38 mT, Figure 1 c,f)).

WM is a square artificial spin-vortex ice with a subset of wider, lower-coercivity bars (Figure 1 g)). Bars are 600 nm × 200 nm (wide-bar)/125 nm (thin-bar). Wide bars host both macrospin and vortex states (Figure 1 k)), thin bars host just macrospins. The FMR spectra comprises four dominant modes with rich responses: wide and thin bar linear macrospin modes (7 GHz and 9 GHz respectively), χ-shaped vortex mode (2-6 GHz) and ‘whispering-gallery’-like high-frequency vortex mode (10 GHz) (Figure 1 h)). At low input fields (18-20 mT), macrospins gradually convert to vortices resulting in non-linear wide-bar mode amplitude changes which lag w.r.t. the input-field (Figure 1 i), t = 20-30). Combining linear (macrospin) and non-linear (vortex) responses is termed a ‘mixture reservoir’ which can outperform purely linear or nonlinear systems.

PW is a pinwheel-lattice artificial spin-vortex ice (Figure 1 m)) with greater density and inter-island coupling. A gradient of bar dimensions is patterned across the sample (extremities shown in Figure 1 m), ranging from fully-disconnected (m, top) to partially-connected islands (m, bottom) giving a range of spectral-responses (panel n) and vortex-nucleation rates. This spatial texturing of the physical response has been shown to enhance reservoir richness and complexity in simulated Skyrmion systems. Additional SEM and description is provided in supporting note 1. Bar dimensions are constant across 100 × 100 µm² (length 450 nm, width 240 nm / 265 nm (lower / upper panel) in m)). Bars host macrospins and vortices (Figure 1 p,q). Low bar aspect ratio gives fast vortex nucleation - seen as rapid macrospin-mode amplitude reduction at lower field cycles (Figure 1 o), t = 15-20).

MC is a measure of how the current state can recall previous inputs. For each previous input (n), MC can range from 0 (no memory) to 1 (perfect memory) (see Methods for details of MC calculation). MC for each frequency channel is $\sum_{n=0}^{k-1} MC_n$, where k is the number of past inputs that MC is evaluated on, hence MC ranges from 0-k with k here set to 8. Higher values of k falsely enhance performance due to the periodic nature of our inputs (supporting note 2 evaluates MC vs. k). In ASRs, MC physically arises from history-dependent microstate evolution and vortex nucleation.

NL measures how well past inputs can be linearly mapped to current outputs. NL ranges from 0 (perfectly linear) to 1 (entirely uncorrelated). In ASRs, NL arises from four effects: non-linear mode-frequency shifts with field, the shape of FMR peaks, complex microstate dynamics, and misleadingly from outputs which do not vary with applied field (i.e. noise). The last effect is avoided by discarding low range (max. amplitude - min. amplitude) outputs. Metric calculations are described in the methods. Reservoir metrics allow mapping of physical to computational properties and relatively comparing different physical reservoirs, shedding light on how the underlying system dynamics define computational properties.

In this work we evaluate metrics on our input data (Mackey-Glass chaotic time-series). This gives metrics which are correlated to conventional MC and NL scores, with some small convolution of the input signal - negligible for our purposes of relatively assessing ASRs and designing network interconnections. For clarity, we denote our signal-evaluated metrics ‘signal memory-capacity’ (sMC) and ‘signal nonlinearity’ (sNL).

Figures 1f,l,r) show FMR spectra at max (orange) and min (blue) input field-amplitude (top), sMC for 0-7 previous time-steps (n), total sMC ($sMC_n = \sum_{n=0}^{k-1} sMC_n$) where sMC_n is the memory of the nth previous input and sNL of each output frequency-channel (bottom) for MS (f), WM (l) and PW (r). For MS, high sMC is limited to n < 3 for the majority of outputs. In contrast, WM and PW have some outputs which remember 0-3 prior time-steps and others remembering 4-7 steps (e.g. WM, 7.2 - 7.5 GHz). For MS and WM, high sMC is focused at the peaks and sub-harmonics of that resonance (e.g. 8 GHz and 5.75 GHz for WM (l)). For PW (r), sMC is highest at ~ 9 GHz corresponding to the longest time-delay between vortex-formation and saturated macrospin states.

In all ASRs, noise-dominated/mode-free channels have high NL as they are uncorrelated to the input. In PW, the main FMR mode has high NL due to complex microstate dynamics. The richer PW spectra provides more meaningful nonlinear channels and hence the highest sNL score.
Individual reservoir performance

a) Single reservoir schematic

b) SNL

MSE vs sMC

SINE TRANSFORMATIONS

f) $Y = \cos(2x)\sin(3x)$
   NL only
   Best reservoir: PW

MSE = 0.0221

MS

WM

PW

g) $Y = |\sin(x/2)|$
   MC + NL
   Best reservoir: WM

MSE = 0.0589

h) $Y = \sin(3x) + \cos(x)$
   MC + NL
   Best reservoir: WM

MSE = 0.0542

MACKEY GLASS

i) Future Prediction t+7
   Best reservoir: WM

MSE = 0.068

j) NARMA 7
   Best reservoir: WM

MSE = 0.447
We now evaluate the computational performance of each individual ASR, how this relates to the physical system dynamics and with $A = 0.3$, $B = 0.01$, $C = 2$, $D = 0.1$. $X$ varies between 1-15, defined as ‘NARMAX’. This task adds additional nonlinear (or vice-versa) - a highly challenging task. Any meaningful computational system requires strong prediction performance (Figure 2d,e). This is clear when predicting Mackey-Glass $t+7$ and NARMA7-transforming Mackey-Glass (Figure 2 i,j)) where (Figure 2 f).

Figure 2. Individual ASR performance. a) Schematic of the single-reservoir computation scheme. Scaled input data is applied via field loops (+$H_{app}$ then -$H_{app}$). FMR output spectra measured at -$H_{app}$ is used as computational output. b) sMC and sNL of the reservoirs. Vortices produce high sMC (WM, PW). Richer FMR spectra produces high sNL (PW). c) MSE when transforming a sine input to a variety of targets. Tasks are chosen to be symmetric (NL only) or asymmetric (MC + NL) w.r.t the input. Vortices enhance performance up to $31.6 \times$ for NL only. Higher sMC enhances MC + NL transformations. d) MSE when predicting future values of the Mackey-Glass time-series. High sMC and low sNL (WM) gives best performance. e) MSE for NARMA-transformation on the Mackey-Glass input signal. $X$ corresponds to how far back the NARMA model is evaluated on. Samples with high sMC perform well. d) and e) show a periodic MSE due to the periodic nature of the Mackey-Glass equation. f-j) Predictions for each sample for tasks with requirements. For NL only tasks (f, cos(2x)sin(3x)), PW dominates. For tasks with MC and NL (g,h), WM performs best. For challenging prediction tasks (i,j) no single reservoir performs well.

We now evaluate the computational performance of each individual ASR, how this relates to the physical system dynamics and how nanoarray design can be used to engineer ASRs with diverse functionality. Throughout this work, two input datasets are used: a sine-wave and the chaotic Mackey-Glass time-series.

Figure 2 a) shows a schematic of the single-reservoir computing scheme. Input data is scaled to a suitable field-range for each reservoir (35 - 42 mT for MS, 18 - 23.5 mT for WM and 30-50 mT for PW). Input scaling is chosen to produce maximal microstate variation. Data is input by applying a minor field-loop with data encoded as field-amplitude $H_{app}$ (i.e. apply field at +$H_{app}$ then -$H_{app}$). Reservoir outputs are obtained by recording FMR spectra at -$H_{app}$. FMR readout is non-linear providing an extra layer of processing. The process is repeated across the input dataset with training performed offline.

Here we employ short training-data sets (225 data points) to mimic real-world applications with strict data-collection time and energy requirements. To avoid overfitting, we employ a sophisticated feature-selection algorithm (each FMR output channel is considered a single computational ‘feature’ with 10-fold cross validation (see Methods). Performance is evaluated via the mean squared error (MSE) between the reservoir-prediction and target.

Figure 2 b) shows the sMC and sNL for each ASR. High sMC is correlated with the presence of vortices due to their history-dependent evolution. PW possesses complex vortex dynamics and high sMC, though slightly less than WM as the lower aspect ratio PW bars more rapidly saturate to a pure vortex state. MS has no vortices, hence low sMC.

Modes in MS vary linearly with field, giving low sNL. Although WM displays non-linear amplitude changes with field, the spectral dynamics are dominated by linear macrospin-mode frequency shifts, hence sNL is low. PW has many rich, strongly nonlinear modes alongside vortex nucleation, resulting in high sNL.

Figure 2c) shows reservoir performance when transforming a sinusoidal input signal ($\sin(x)$) to a variety of target waveforms. Figure 2f-h) shows example transformations (black curves) and residuals (red shaded). Transforms shown are f) cos(2x)sin(3x), g) $\sin(x/2)$ and h) $\sin(3x) + \cos(x)$.

Signal-transformations require varying sMC and sNL depending on the target waveform. Targets which are symmetric w.r.t. the input (eg. $\sin^2(x)$, $\sin(3x)\sin(x)$) require sNL only, asymmetric waveforms w.r.t the input (eg. saw(x), $\sin(x/2)$) require both sNL and sMC as equivalent input values must be transformed to different output values across the wave-cycle. Highest sNL PW dominates for symmetric transforms. Highest sMC WM dominates for asymmetric transforms. The benefits available via nanoarray design optimisation are impressively displayed here, PW massively outperforms MS by up to $31.6 \times$ lower MSE (Figure 2 f).

For challenging tasks (e.g. $\sin(3x) + \cos(x)$, panel h)) no ASR performs exceptionally well, as highlighted by high MSEs and prominent residuals, demonstrating the intrinsic limitations of any single physical system.

Figure 2d) shows the performance of each ASR during future prediction of the Mackey-Glass time-series. These tasks require both high sMC and low sNL, seen by WM outperforming other reservoirs. Figure 2e) shows performance when performing a NARMA-transform$^{77}$ on the Mackey-Glass input, evaluated as $y(t) = Ay(t-1) + By(t-1) \sum_{n=1}^{X} y(t-n) + Cu(t-1)u(t-X) + D$ with $A = 0.3$, $B = 0.01$, $C = 2$, $D = 0.1$. $X$ varies between 1-15, defined as ‘NARMAX’. This task adds additional nonlinear complexity vs. normal Mackey-Glass prediction, favouring both sMC and sNL with WM and PW performing similarly across all tasks.

Engineering ASRs for vortex-driven history-dependent microstate evolution gives high sMC and strong future-prediction performance. This can be applied generally across any physical system. Introducing multiple states/phases with history-dependent dynamics will produce similar results. As predicted in theory$^{19}$, combining nonlinear (vortex) with linear (macrospins) input responses in WM results in a ‘mixture-reservoir’ providing optimal prediction of chaotic time-series.

For all ASRs we again observe performance breakdown for harder prediction tasks, evidenced by the periodic MSE profiles (Figure 2d,e). This is clear when predicting Mackey-Glass $t+7$ and NARMA7-transforming Mackey-Glass (Figure 2 i,j)) where predictions fail to reproduce target waveforms. The Mackey-Glass series has quasi-periodicity (approximate period of 22). MSE minima occur at $\frac{m \pi}{2}$ period intervals where behaviour at current step $t$ most closely resembles future step $t + \tau$. MSE maxima are found at $\frac{m \pi}{2}$ intervals where $m$ is odd. Here, a small input gradient must be mapped onto a large output gradient (or vice-versa) - a highly challenging task. Any meaningful computational system requires strong prediction performance.
across a broad task set. This demands a physical reservoir response exhibiting multiple memory timescales, allowing recall of a range of past input gradients to reconstruct arbitrary future behaviours\textsuperscript{22,78}. Any single reservoir typically exhibits a single resonant memory timescale, severely limiting the efficacy of single reservoir systems for complex multi-timescale predictive tasks\textsuperscript{20,78}.

**Parallel networks**

No individual ASR excels across all tasks. This is a well-known symptom of single reservoir systems\textsuperscript{17–20,22}. Multiple reservoirs with distinct responses may be combined in networks to harness the benefits of different dynamical behaviours. In
such networks, each ASR node can be viewed as a ‘hyper-neuron’ with high-dimensionality, memory and nonlinearity.

Here, we construct parallel networks, combining a synergistic suite of distinct ASRs with substantially different metrics and observe resultant performance enhancements. The enhanced output dimensionality from multiple reservoirs is a huge positive for computation, but increases the likelihood of overfitting - a common challenge in machine-learning. One way to avoid this is to increase the size of the training dataset, however, this luxury is not always available in real-world applications. In response, we have devised an intelligent feature selection methodology to avoid overfitting and provide robust, accurate performance. Code and full description of the methodology is provided for the benefit of the community (see supplementary note on Learning).

Figure 3a) shows a schematic of a parallel network. Data is input in parallel to multiple ASRs and the FMR response of each ASR measured is then combined for prediction. Figure 3b) shows the metrics of single and parallel ASR networks. When constructing parallel networks, metrics from constituent reservoirs are combined to allow a range of possible metric scores (triangular markers, Figure 3b). The feature selection algorithm naturally caps the number of outputs to below 200 for all network architectures (architecture referring to the specific arrangement of ASR nodes). Parallel sNL and sMC typically lie between the sNL and sMC from the constituent reservoirs for each network architecture (hence single PW has highest sNL, single WM has the highest sMC). sMC does not increase in parallel as no information is transferred between ASRs.

Figure 3c) shows MSEs of the best single and parallel networks for each signal-transformation task. Parallel networks are able to harness sMC and sNL characteristics from different ASRs to reduce MSE for asymmetric transforms. Introducing high sMC outputs from WM means sacrificing useful sNL outputs from PW, reducing NL-only task performance. Example transformations are shown for the best single and parallel networks for NL only (f, cos(2x)sin(3x)) and MC + NL (g, lsin(x/2)), h, sin(3x)+cos(x)) transforms. Lowest MSEs are observed when combining WM + PW, or all three ASRs with performance gains up to 4.4 × for more complex asymmetric tasks (h).

Contrastingly, the same degree of improvement is not observed for future-prediction tasks. Figure 3d) shows MSE profiles when predicting 0-22 future steps. Here, only slight improvements in MSE are seen (< 1.3 ×). The same behaviour is observed for the NARMA-transformed Mackey-Glass (e). Figure 3 i,j) shows example predictions for t+7 (i) and NARMA7 (j) for the best single and parallel networks. These observations are expected as improving performance in prediction tasks requires additional sMC. When combining reservoirs in parallel, there is no way for sMC to significantly improve, as no information flows between reservoirs.

**Deep networks**

To further enhance sMC and future-prediction, we now construct deep networks, feeding outputs from one reservoir to the input of the next.

Figure 4a) shows a schematic of a deep network architecture. Data is input to the first reservoir (R1) and FMR responses measured. The 300-dimensional R1 FMR response now must be converted to a 1D field input for R2. To accomplish this, we take the amplitude of a specific FMR frequency-channel and map it to an input field sequence for R2. The R1 output frequency-channel is selected via per-channel sMC and sNL evaluation (discussed later). The FMR spectra of R1 and R2 are combined for training and prediction.

Figure 4b) shows sMC and sNL for single (circles), parallel (triangles), two-layer (squares) and three-layered (diamonds) networks. Symbols are colour-coded depending on the architecture of ASR nodes within the network, consistent with labels in d-f). The feature selection algorithm naturally caps the number of outputs to below 200 for each architecture. For each architecture, multiple inter-layer connections (configurations) are trialled, only those with the lowest MSE’s are displayed. The resultant range of network configurations is vast, spanning a wide variety of sMC and sNL and demonstrating the huge reconfigurability afforded by this approach. Crucially for prediction tasks, sMC can be extended beyond that of single and parallel networks with corresponding performance enhancements.

The significant optimisation time and energy required to evaluate all input-output combinations is not practically feasible in experiment or simulation. For n layers, N ASRs and i output channels N\(^{n×i}\) configurations are available (here 10\(^{226}\)). An ideal solution would evaluate how changing inter-layer connectivity affects network performance.

Figure 4 c) shows this relationship when R2 is WM (other architectures shown in supplementary note 5). Here, sMC\(_{in}\) and sNL\(_{in}\) refer to the sMC and sNL of the selected R1 output-channel (i.e. sMC and sNL from Figure 1 panels f,l,r). sMC\(_{R2}\) and sNL\(_{R2}\) are the sMC and sNL of R2. Points are colour-coded depending on which reservoir acts as R1 (consistent with previous figures). Both sMC\(_{R2}\) and sNL\(_{R2}\) follow an approximately linear relationship. A linear relationship is also observed when comparing specific previous inputs (i.e. for the nth previous input, the interconnection memory sMC\(_{in,n}\) and R2 memory sMC\(_{R2,n}\) are linear, supplementary note 3). We also find that both WM and PW are able to amplify weak amounts of long-term memory in the input signal to give large long-term memory enhancements for the network. This interconnection control goes beyond conventional RC where interconnections are made at random.
Figure 4. Deep reservoir computing. a) Schematic of a deep network architecture. Specific output-channels from R1 are input into second reservoir (R2). Outputs of both reservoirs are then combined for training. b) sMC and sNL of single, parallel and deep reservoir networks. A broad range of enhanced sMC and sNL are obtained. c) Relationship between R1 output metrics used for interconnections and measured R2 metrics when R2 is WM. Lines represent linear fits. Higher R2 metric scores are correlated with higher-scoring R1 output channels. d) MSE profiles for different networks when transforming a sine input. Improvements are observed for complex asymmetric tasks. MSE vs t for e) Mackey-Glass future prediction and f) NARMA-transformed Mackey-Glass. Networks go from low sMC to high sMC reservoirs. Improvement factors are shown for t+7, t+16, NARMA7 and NARMA11. MSE profiles are flattened in deep networks, with improvements up to 4×. Example predictions are shown for g) t+7 and h) NARMA7 for single and deep networks.
Figure 4 d) shows MSE values for single, parallel and deep networks for signal-transformation. Deep networks show enhanced performance for complex asymmetric tasks requiring both MC and NL (e.g. \( \sin(2x) + \cos(3x) \)). For simpler MC+NL and NL-only tasks, single and parallel networks dominate. That our architectural flexibility enables strong performance across a broad and challenging taskset shows how critical it is to match specific physical dynamics to different tasks and demonstrates the power of a multi-system approach to neuromorphic computing.

Figure 4 e,f) shows deep network MSE vs \( t \) for future prediction e) and NARMA-transformation f) of Mackey-Glass time-series. Improvement factors for \( t+7, t+16 \) (e) and NARMA7, NARMA11 (f) are shown. Low sMC (R1) to high sMC (R2/R3) network architectures significantly improve performance, up to \( 2.7 \times \) and \( 4 \times \) in three-layer deep networks for challenging \( t+7 \) and NARMA7 tasks respectively. Reservoir ordering is critical, with high sMC to low sMC network architectures showing weak improvements of \(< 1.5 \) for all prediction tasks, matching prior theoretical work \(^{22} \) (supplementary note 4 discusses deep architecture ordering). Three-layer deep networks substantially outperform two layers, highlighting the benefits of more complex networks.

Best prediction is found for MS -> PW -> WM (3 layers, \( sMC = 6.7, sNL = 0.75 \)). While the 3-layer deep network has an sMC improvement of 1.31 vs. the best single reservoir, MSE values are up to \( 3.7 \times \) lower than the single WM sample for future prediction tasks due to the expanded range of temporal responses/timescales in the deep network. This expanded range of temporal responses results in strong flattening of MSE periodicity vs. single and parallel reservoirs, especially evident in the NARMA-transform task (f). This enhanced temporal richness is not accurately reflected in a single sMC/MC score. sMC and sNL metrics are valuable guides for basic reservoir evaluation, but lack the granularity for accurately predicting performance.

### Multilayer Physical Neural Network

We have evaluated a set of 50 two-layer deep configurations and two three-layer deep configurations. We now combine all deep networks in parallel, giving a multilayer physical neural network (PNN) where each neuron is a highly-dimensional, recurrent and nonlinear in materia reservoir. Interconnections between layers act as programmable network weights. Figure 5 a) shows a schematic of the PNN. Output combinations are selected across all 52 deep configurations for optimal performance at a given task.

Figure 5 b) compares single, parallel, deep and hybrid MSEs for \( t+7 \) Mackey-Glass prediction. The feature selection algorithm caps the number of outputs to below 200 for all architectures. The PNN substantially outperforms all other networks considered due to the broad range of memory timescales and non-linear dynamics present across its constituent reservoir neurons. Depicted as the black points and curve is the performance of software echo-state networks (ESNs) with increasing numbers of network nodes (10-100). The grey error window represents the MSE spread across 100 ESN trials. The performance of single, parallel and deep networks can be matched by adequately-sized software-based ESNs. However, the PNN substantially outperforms software ESNs, demonstrating the power of a PNN harnessing diverse and synergistic interconnected physical responses.

The PNN consistently outperforms other architectures across all future time steps with significant MSE vs \( t \) flattening for future prediction (Figure 5 c). Here we introduce an additional task - joint transformation and future-prediction of NARMA-processed Mackey-Glass. Previous architectures were evaluated on unprocessed Mackey-Glass prediction or NARMA-processed transformation only; here the PNN architecture is sufficiently powerful to NARMA-transform and future predict simultaneously (Figure 5 d), showing strong performance across a broad range of timescales. Additionally, the PNN performs well across a range of signal-transformations (Figure 5 e), outperforming other network architectures at 9/20 tasks. Which network architecture gives best signal-transform performance is highly target dependent, highlighting the importance of matching specific physical system dynamics to the desired task. PNNs are still expected to outperform all constituent sub-reservoirs with long enough training datasets, the more complex network architecture benefitting more from longer training.

PNNs combine the benefits of complex deep neural networks with the low training costs of RC approaches. Our efficient interconnection scheme allows a small suite of sub-reservoirs to be reconfigurably programmed on demand for changing task-specific needs. The PNN approach provides a bridge between the restricted, hard-coded nature of single reservoirs and highly-reconfigurable but energy-hungry deep neural networks, providing greatly enhanced flexibility and performance at low training time and energy.

The role of the intelligent feature selection algorithm is essential here. The PNN has \(~16000\) features. Gathering a sufficiently long dataset to avoid overfitting with conventional feature selection is unfeasible. Our feature selection algorithm is capable of reducing the size of this feature space whilst maintaining feature diversity to achieve high performance with no overfitting even at short training sets.
Figure 5. Multilayer physical neural network with reservoir neurons. a) Schematic of the physical neural network system. Multiple deep networks are combined in parallel with multiple interconnections between each neuron. A feature selection algorithm selects outputs across the network to optimise performance. b) MSE when predicting t+7 of the Mackey-Glass equation for single, parallel, deep and neural networks. The neural network show a significant error reduction. Shaded region and black curve shows the spread and average of MSE’s when using software ESN’s with varying number of network nodes. The neural network achieves significantly better performance whereas single, parallel and deep can be matched by a single ESN. MSE profiles and improvement factors for c) Mackey-Glass future prediction and d) future prediction of NARMA-7 processed Mackey-Glass for the best single, parallel, deep and neural networks. MSE profiles are significantly flattened for the neural network with up to $5.1 \times$ improvement. e) MSE when transforming to a sine input. Colours are consistent with b) with best networks highlighted. Neural network improvements are found for complex asymmetric tasks. f-i) Example predictions for t+9 of the NARMA7 processed Mackey-Glass signal.

Conclusion

We have engineered multiple artificial spin systems with varying microstate dynamics and spin-wave responses, evaluating their metrics and performance across a broad benchmark taskset.
Our results highlight the computational performance gained from enriched state spaces. We show the computational benefits of enhanced physical system richness including spatial disorder/parameter gradients and enhanced inter-element coupling in the disordered pinwheel ASR. The PW outperforms MS by up to $31.6 \times$, using the same amount of material and excitation and measurement energy. Our work demonstrates clearly that careful design of system geometry and dynamics is critical, with huge computational benefits available ‘for free’ via design optimisation.

Crucially, we have overcome the intrinsic performance compromises facing single reservoir systems by engineering synergistic network architectures from distinct physical systems. Parallel networks harness the different dynamic responses of individual reservoirs to enhance nonlinearity and transformation performance. Deep networks achieve enhanced memory and richer temporal responses to enable strong prediction across future steps. PNNs combine the best of both approaches, demonstrating powerful reconfigurable performance and moving physical neuromorphic computing beyond simple reservoir schemes towards flexible programmable networks where each neuron expresses high-dimensionality, recurrence and non-redundant nonlinearity while retaining the low-energy benefits of RC. Our feature selection algorithm plays a crucial role in identifying a small set of diverse, synergestic features, enabling true network performance to be realised with short datasets.

Crucially, our method of interconnecting network layers via assessing output-channel/feature metrics allows efficient design of the overall system metrics, bypassing costly iterative approaches. The approach is broadly-applicable across physical neuromorphic schemes. If the required sMC and sNL for a given task are known, metric programming allows rapid configuration of an appropriate network. If the required sMC and sNL are not known or the task depends on more than these metrics, metric programming can be used to search the sMC, sNL phase space for pockets of high performance using gradient-based approaches. The introduction of a trainable inter-layer parameter opens vast possibilities in implementing hardware neural networks with reservoirs serving as nodes and inter-layer connections serving as weights.

Our results have strong implications for the scope and versatility of physical neuromorphic computing. By harnessing a diverse suite of physical samples and network architectures, we expand high computational performance across a diverse taskset and demonstrate a flexible beyond-reservoir computational approach.

**Author contributions**

KDS and JCG conceived the work and directed the project throughout.

KDS drafted the manuscript with contributions from all authors in editing and revision stages.

JCG, KDS and AV performed FMR measurements.

JCG and HH performed MFM measurements.

JCG and KDS fabricated the samples. JCG and AV performed CAD design of the structures.

JCG performed SEM measurements.

KDS and LM implemented the reservoir computation scheme.

LM developed the cross-validation training approach for reducing overfitting on shorter training datasets.

LM developed the feature selection methodology for intelligently selecting optimal features from many reservoirs in the multilayer physical neural network architecture.

KDS designed and implemented the metric programming method.

CC and TC aided in analysis of reservoir metrics.

JL contributed task-agnostic metric analysis code.

FC helped with conceiving the work, analysing computing results and providing critical feedback.

EV provided oversight on computational architecture design.

KES and EV provided critical feedback.

WRB oversaw the project and provided critical feedback.

**Acknowledgements**

This work was supported by the Leverhulme Trust (RPG-2017-257) to WRB and the Engineering and Physical Sciences Research Council (Grant No. EP/W524335/1) to KDS

JCG was supported by the Royal Academy of Engineering under the Research Fellowship programme. AV was supported by the EPSRC Centre for Doctoral Training in Advanced Characterisation of Materials (Grant No. EP/L015277/1). This work was supported by the Laboratory Directed Research and Development program of Los Alamos National Laboratory under project number 2023038ER. The authors would like to thank David Mack for excellent laboratory management.

**Competing interests**

Authors patent applicant. Inventors (in no specific order): Kilian D. Stenning, Jack C. Gartside, Alex Vanstone, Holly H. Holder, Will R. Branford. Application number: PCT/GB2022/052501. Application filed. Patent covers the method of programming
deep networks.
Methods

Some description of our methodologies are reproduced from earlier work of several of the authors, as similar methods are employed here.

Nanofabrication

Artificial spin reservoirs were fabricated via electron-beam lithography liftoff method on a Raith eLine system with PMMA resist. 25 nm Ni_{81}Fe_{19} (permalloy) was thermally evaporated and capped with 5 nm Al_{2}O_{3}. For WM, a ‘staircase’ subset of bars was increased in width to reduce its coercive field relative to the thin subset, allowing independent subset reversal via global field. For PW, a variation in widths were fabricated across the sample by varying the electron beam lithography dose. Within a 100 µm × 100 µm write-field, the bar dimensions remain constant. The flip-chip FMR measurements require mm-scale nanostructure arrays. Each sample has dimensions of roughly 3x2 mm. As such, the distribution of nanofabrication imperfections termed ‘quenched disorder’ is of greater magnitude here than typically observed in studies on smaller artificial spin systems, typically employing 10-100 micron-scale arrays. The chief consequence of this is that the Gaussian spread of coercive fields is over a few mT for each bar subset. Smaller ASR arrays have narrower coercive field distributions, with the only consequence being that optimal applied field ranges for reservoir computation input will be scaled across a corresponding narrower field range, not an issue for typical 0.1 mT or better field resolution of modern magnet systems.

FMR measurement

Ferromagnetic resonance spectra were measured using a NanOsc Instruments cryoFMR in a Quantum Design Physical Properties Measurement System. Broadband FMR measurements were carried out on large area samples (~3 × 2 mm²) mounted flip-chip style on a coplanar waveguide. The waveguide was connected to a microwave generator, coupling RF magnetic fields to the sample. The output from waveguide was rectified using an RF-diode detector. Measurements were done in fixed in-plane field while the RF frequency was swept in 10 MHz steps. The DC field was then modulated at 490 Hz with a 0.48 mT RMS field and the diode voltage response measured via lock-in. The experimental spectra show the derivative output of the microwave signal as a function of field and frequency. The normalised differential spectra are displayed as false-colour images with symmetric log colour scale.

Reservoir computation

The reservoir training inputs are chosen to have approximately 30 data points per period for the sine and Mackey-Glass input with outputs sampled at every input. The Mackey-Glass time-delay differential equation takes the form \( \frac{dx}{dt} = \beta \frac{x(t-\tau)}{1+x^{10}} - \lambda x(t) \) and is evaluated numerically with \( \beta = 0.2, n = 10 \) and \( \tau = 17 \). The array is initially saturated in a -200 mT field in the \( \hat{z} \) direction.

A number of tasks were trialled. Signal-transformation tasks were chosen by taking a set of base transforms: \( \sin(x), |\sin(x/2)|, \sin(2x), \sin(3x), \sin^2(x), \sin^3(x) \cos(x), \cos(2x), \cos(3x), \text{saw}(x), \text{saw}(2x) \) and adding and multiplying them together. Mackey-Glass future predictions were varied from 0 - 25 future steps. NARMA tasks were evaluated using \( y(t) = Ay(t-1) + By(t-1) \sum_{i=1}^{X} y(t-X) + Cu(t-1)u(t-X) + D \) where \( u(t) \) is the Mackey-Glass equation, \( A = 0.3, B = 0.01, C = 2, D = 0.1 \) and \( X \) varies between 1-15.

Reservoir computing schemes consist of three layers: an input layer, a ‘hidden’ reservoir layer, and an output layer corresponding to globally-applied fields, the ASR and the FMR response respectively. In each case, the inputs were linearly mapped to a field range spanning 35 - 42 mT for MS, 18 - 23.5 mT for WM and 30-50 mT for PW, with the mapped field value corresponding to the maximum field of a minor loop applied to the system. After each minor loop, the FMR response is measured at the applied field \( H_{app} \) between 8- 12.5 GHz, 5 - 10.5 GHz and 5 - 10.5 GHz in 20 MHz steps for MS, WM and PW respectively. The FMR output is smoothed in frequency by applying a low-pass filter to reduce noise. Eliminating noise improves computational performance. For each input data-point, we measure between 225 - 275 distinct frequency bins and take each bin as an output giving 225 - 275 reservoir outputs. This process is repeated for the entire dataset with training and prediction performed offline.

Offline training and prediction is performed using a matrix multiplication method. We first separate the reservoir response into two datasets. A ‘train’ dataset for learning the optimum set of weights for a given task and a ‘test’ dataset to test the performance of the learned weights on previously unseen data. If we consider the training set of reservoir outputs \( \bar{x}_{\text{train}}(u) \) and the target waveform \( \bar{y} \)

\[
\begin{align*}
\bar{x}_{\text{train}}(u) &= \begin{pmatrix}
O_0(0) & O_0(1) & \cdots & O_0(m) \\
O_1(0) & O_1(1) & \cdots & O_1(m) \\
\vdots & \vdots & \ddots & \vdots \\
O_n(0) & O_n(1) & \cdots & O_n(m)
\end{pmatrix} \\
\bar{y} &= \begin{pmatrix}
y_0 \\
y_1 \\
\vdots \\
y_n
\end{pmatrix}
\end{align*}
\]
where $m$ is the number of reservoir outputs, $n$ is the number of input data points in the ‘train’ dataset and $O_n(m)$ is the $m^{th}$ reservoir output for input step $n$ (i.e. the amplitude of the $m^{th}$ frequency bin). The goal of training is to obtain a $1 \times M$ array of weights $\mathbf{W}_{\text{out}}$ which transform the reservoir outputs ($\tilde{\mathbf{y}}(u)$) into the desired target ($\tilde{\mathbf{y}}$). For this, we use ridge regression which solves the linear optimisation problem $\mathbf{W}_{\text{out}} = \arg\min_{\mathbf{W}} (||\tilde{\mathbf{y}} - \mathbf{W}\mathbf{x}_{\text{train}}(u)||^2 + \lambda ||\mathbf{W}||^2)$ with the following solution $\mathbf{W}_{\text{out}} = (\mathbf{x}_{\text{train}}(u)^T \mathbf{x}_{\text{train}}(u) + \lambda \mathbf{I})^{-1} \mathbf{x}_{\text{train}}(u)^T \tilde{\mathbf{y}}$ where $\mathbf{I}$ is the identity matrix and $\lambda$ is the ridge regularisation term - optimised via a grid-search. This can be performed with freely-available Python packages such as scikit-learn.

Once $\mathbf{W}_{\text{out}}$ is obtained, we can then multiply the ‘test’ dataset $\mathbf{x}_{\text{test}}(u)$ with $\mathbf{W}_{\text{out}}$ to obtain the reservoir prediction $\tilde{\mathbf{y}}_{\text{ASR}}$ (i.e. $\mathbf{x}_{\text{test}}(u)\mathbf{W}_{\text{out}} = \tilde{\mathbf{y}}_{\text{ASR}}$) as follows:

$$\mathbf{x}_{\text{test}}(u)\mathbf{W}_{\text{out}} = \begin{pmatrix} O_{n+1}(0) & O_{n+1}(1) & \cdots & O_{n+1}(m) \\ O_{n+2}(0) & O_{n+2}(1) & \cdots & O_{n+2}(m) \\ \vdots & \vdots & \ddots & \vdots \\ O_{n+l}(0) & O_{n+l}(1) & \cdots & O_{n+l}(m) \end{pmatrix} \begin{pmatrix} W_0 \\ W_1 \\ \vdots \\ W_m \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{y}}_{\text{ASR},n+1} \\ \tilde{\mathbf{y}}_{\text{ASR},n+2} \\ \vdots \\ \tilde{\mathbf{y}}_{\text{ASR},n+l} \end{pmatrix} = \tilde{\mathbf{y}}_{\text{ASR}}$$

where $l$ is the number of input datapoints in the ‘test’ dataset. We assess the performance of the reservoir by calculating the mean-squared error (MSE) of target waveform and the reservoir response $\text{MSE} = \frac{1}{n} \sum_{t=0}^{t=n-1} ||\mathbf{y}_{\text{ASR}}(t) - \tilde{\mathbf{y}}(t)||^2$.

To avoid overfitting to the train dataset, the number of outputs, and therefore weights, must be kept below the number of training points. Evaluating all output combinations would require $2^m - 1 \sim 10^8$ iterations for a single reservoir which increases dramatically for parallel, deep and hybrid networks. As such, we employ a feature selection algorithm which optimises which outputs to use for a given task (described later in the Learning section).

When evaluating MC and NL, it is conventional to use a random input. However in our scheme, discontinuities in the input field arising from discontinuous data input results in sharp mode-shifts between output frequency-channels, obscuring results. Our measured FMR peaks are broad and peak amplitude variations are detected across a number of output bins. As such, a single output can still track history-dependent behaviour as long as the field-induced frequency shifts remain small. As a result, we must implement an appropriate methodology to accurately reduce the dimensionality of the feature space and measure the performance of the models. The methodology adopted can be at first described as a 10 cross-validation (inner validation loop) of a 10 cross-validation approach (outer validation loop), where the outer cross-validation is used to accurately evaluate the performance and the inner loop is used to perform feature-selection.

As stated above, the high number of reservoir outputs can lead to overfitting. Thus, it was necessary to decrease the dimensionality of the space at which the learning of the read-out connectivity matrix $\mathbf{W}_{\text{out}}$ operates through feature selection. Let us consider the response of the considered reservoir $\mathbf{x}_{\text{train}}(u)$, which is the $n \times m$ dimensional matrix introduced in the previous section. The feature selection algorithm aims to use $o$ subsets of frequency bins, where $o \leq n$, that will be used for training and evaluation. We can consider feature selection as a boolean operation in the $m$-dimensional feature space, where a value of one (zero) corresponds to the considered feature being used (neglected). Given the high value of $m$, the number of possible combinations $2^m - 1$ is extremely high. As a consequence, the feature selection algorithm can lead to overfitting, therefore we must implement an appropriate methodology to accurately reduce the dimensionality of the feature space and measure the performance of the models. The methodology adopted can be at first described as a 10 cross-validation (inner validation loop) of a 10 cross-validation approach (outer validation loop), where the outer cross-validation is used to accurately evaluate the performance and the inner loop is used to perform feature-selection.

Considering a specific split of the outer validation loop (Fig. 6 a), where we select a validation set $\mathcal{V}_i$ and a test set $\mathcal{V}_j$ (comprising of the 10% of the data each), we performed 10 cross-validation on the remaining data to optimise hyperparameter values through grid-search. In this inner validation loop, each split corresponds to a test set $\mathcal{V}_j$ (comprising again of the 10% of the remaining data, without $\mathcal{V}_i$ and $\mathcal{V}_j$), where changing $j$ means to select a different test-split in the inner loop based on the $i$-th original split of the outer validation (Fig. 6 a). The remaining data, highlighted with the blue contour of Fig. 6 a, are used for training to optimise the read-out weights and minimise an error function $E$. At this stage, we performed a grid-search methodology on hyperparameters $\theta$ and $\lambda$ which control directly and indirectly the features being adopted for training (Fig. 6 b). The hyperparameter $\theta$ acts as a threshold on the correlation matrix of the features. Simply, if the correlation among two features exceeds the specific value of $\theta$ considered, one of these two features is removed for training
The idea behind this methodology is to discard features that are highly correlated, since they would contribute in a similar way during training. This emphasizes diversity in the reservoir response. The hyperparameter $\lambda$ is the penalty term in ridge-regression. Higher values of $\lambda$ lead to a stronger penalization on the magnitude of the read-out weights. As such, $\lambda$ can help prevent overfitting and controls indirectly the number of features being adopted. We should use a high value of $\lambda$ if the model is more prone to overfitting the training dataset, a case that occurs when the number of features adopted is high.

Calling $E\{T_{ij}'|W_o^*,\theta,\lambda\}$ the error computed on the test set $T_{ij}'$ with weights $W_o^*$ optimised on the corresponding training data ($W_o^* = \arg\min_{W_o \in \mathbb{R}^m} E\{ T_{ij}|W_o,\theta,\lambda \}$) in the algorithm of Fig. 6 b) and with hyperparameter values $\theta$ and $\lambda$ respectively, we select the values of the hyperparameters that correspond to the minimum average error over the test sets in the inner validation loop. Otherwise stated, we select the optimal $\theta^*_i$ and $\lambda^*_i$ for the $i$-th split in the outer loop from the test average error in the inner 10 cross-validation as $\theta^*_i, \lambda^*_i = \arg\min_{\theta,\lambda \in \Theta, \Lambda} \sum_{j=1}^{10} E\{ T_{ij}'|W_o^*,\theta,\lambda \}$. This methodology permits to find $\theta^*_i$ and $\lambda^*_i$ that are not strongly dependent on the split considered, while maintaining the parts of the dataset $\mathcal{Y}_i$ and $\mathcal{F}_i$ unused during training and hyperparameter selection. The sets $\Theta$ and $\Lambda$ correspond to the values explored in the grid-search. In our case, $\Theta = \{1, 0.999, 0.99, 0.98, 0.97, \ldots\}$ and $\Lambda = \{1e^{-4}, 1e^{-3}, 1e^{-2}, 2.5e^{-2}, 1e^{-1}\}$. Repeating this procedure for each split of the outer loop, we found the optimal $\theta^*_i$ and $\lambda^*_i$, for $i = 1, \ldots, 10$. This concludes the algorithm (Hyperparameter Selection) described in Fig. 6 b. Selection of the hyperparameters $\theta^*_i$ permit to find subsets of features based on correlation measures. However, promoting diversity of reservoir measures does not necessarily correspond to the highest performance achievable. Thus, we adopted an evolutionary algorithm to better explore the space of possible combination of measurements (algorithm of Fig. 6 c).

It is necessary now to notice how a value of $\theta^*_i$ corresponds to a $m$-dimensional boolean vector $f^{(i)}$, whose $j$-th dimension is
zero if its \(j\)-th feature \(f^j\) is correlated more than \(\theta^j\) with at least one other output. For each split \(i\) in the outer loop, we adopted an evolutionary algorithm that operates over the \(m\)-dimensional boolean space of feature-selection, where each individual corresponds to a specific vector \(f\). At each evolutionary step, we performed operations of crossover and mutation over a set of \(N_p\) parents \(F_p = \{f(n)\}_{n=1,...,N_p}\). For each split \(i\) in the outer loop and at the first evolutionary step, we initialised the parents of the algorithm to \(f^0\). We defined a crossover operation \(CO\) among two individuals \(f(i)\) and \(f(j)\) as \(f = CO(f(i), f(j))\) where the \(k\)-th dimension of the new vector \(f\) is randomly equal to \(f(i)_k\) or \(f(j)_k\) with the same probability. A mutation operation of a specific \(f(i)\) is defined as \(f = M(f(i))\) by simply applying the operator \(not\) to each dimension of \(f(i)\) with a predefined probability \(p_m\). The application of crossovers and mutations permits the definition of a set of \(N_c\) children \(F_c = \{f(n)\}_{n=1,...,N_c}\) from which we select the \(N_p\) models with the highest performance over the test sets of the inner loop as parents for the next iteration. Otherwise stated, we selected the \(F_p\) vectors corresponding to the lowest values of the average error as \(F_p = \arg \min_{f \in F_c} \sum_{j=1}^{10} E \left\{ \mathcal{F}_j \parallel \mathcal{W}_0, f, \lambda^*_1 \right\} \) where \(\arg \min_{N_p}\) selects the \(N_p\) arguments of the corresponding function with minimal values. We notice how a step of the evolutionary approach aims to minimise an error estimated in the same fashion as in the algorithm of Fig. 6 b, but this time searching for the best performing set \(F_p\), rather than the best performing couple of hyperparameter values \(\lambda^*_1\) and \(\theta^*_1\). Finally, we stopped the evolutionary algorithm at the iteration instance where the average performance of \(F_p\) over \(\mathcal{V}_i\) is at minimum and selected the model \(f_i^*\) with the lowest error on \(\mathcal{V}_i\). The utilisation of a separate set \(\mathcal{V}_i\) for the stop-learning condition was necessary to avoid overfitting of the training data. Indeed, it was possible to notice how the performance on \(\mathcal{V}_i\) would improve for the first iterations of the evolutionary algorithm and then become worse. This concludes the evolutionary algorithm depicted in Fig. 6 c. At last, the overall performance of the model is computed as the sum of the mean-squared errors over the outer validation loop as \(\mathcal{E} = \sum_{i=1}^{10} E \left\{ \mathcal{F}_i \parallel \mathcal{W}_0, f_i^*, \lambda^*_1 \right\} \). Summarising, we can think the overall methodology as an optimisation of relevant hyperparameters followed by a fine-tuning of the set of features used through an evolutionary algorithm. The final performance and its measure of variation reported in the paper are computed as average and standard deviation over ten repetitions of the evolutionary algorithm respectively.

**MFM measurement**
Magnetic force micrographs were produced on a Dimension 3100 using commercially available normal-moment MFM tips.

**Data availability statement**
The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

**Code availability statement**
The code used in this study is available from the corresponding author on reasonable request.

**References**
1. Andrae, A. S. & Edler, T. On global electricity usage of communication technology: trends to 2030. *Challenges* 6, 117–157 (2015).
2. Jones, N. How to stop data centres from gobbling up the world’s electricity. *Nature* 561, 163–167 (2018).
3. Tanaka, G. *et al.* Recent advances in physical reservoir computing: A review. *Neural Networks* 115, 100–123 (2019).
4. Nakajima, K. Physical reservoir computing—an introductory perspective. *Jpn. J. Appl. Phys.* 59, 060501 (2020).
5. Milano, G. *et al.* In materia reservoir computing with a fully memristive architecture based on self-organizing nanowire networks. *Nat. Mater.* 21, 195–202 (2022).
6. Dawidek, R. W. *et al.* Dynamically driven emergence in a nanomagnetic system. *Adv. Funct. Mater.* 31, 2008389 (2021).
7. Torrejon, J. *et al.* Neuromorphic computing with nanoscale spintronic oscillators. *Nature* 547, 428–431 (2017).
8. Gartside, J. C. *et al.* Reconfigurable training and reservoir computing in an artificial spin-vortex ice via spin-wave fingerprinting. *Nat. Nanotechnol.* 17, 460–469 (2022).
9. Lee, O. *et al.* Task-adaptive physical reservoir computing. *arXiv preprint arXiv:2209.06962* (2022).
10. Vidamour, I. *et al.* Reservoir computing with emergent dynamics in a magnetic metamaterial. *arXiv preprint arXiv:2206.04446* (2022).
11. Zhong, Y. *et al.* A memristor-based analogue reservoir computing system for real-time and power-efficient signal processing. *Nat. Electron.* 5, 672–681 (2022).
12. Marković, D., Mizrahi, A., Querlioz, D. & Grollier, J. Physics for neuromorphic computing. *Nat. Rev. Phys.* **2**, 499–510 (2020).
13. Milano, G. *et al.* In materia reservoir computing with a fully memristive architecture based on self-organizing nanowire networks. *Nat. Mater.* **1–8** (2021).
14. Chumak, A. *et al.* Roadmap on spin-wave computing concepts. *IEEE Transactions on Quantum Eng.* (2021).
15. Papp, Á., Porod, W. & Csaba, G. Nanoscale neural network using non-linear spin-wave interference. *Nat. communications* **12**, 1–8 (2021).
16. Cucchi, M., Abreu, S., Ciccone, G., Brunner, D. & Kleemann, H. Hands-on reservoir computing: a tutorial for practical implementation. *Neuromorphic Comput. Eng.* (2022).
17. Love, J., Mulkers, J., Bourianoff, G., Leliaert, J. & Everschor-Sitte, K. Task agnostic metrics for reservoir computing. arXiv preprint arXiv:2108.01512 (2021).
18. Inubushi, M. & Yoshimura, K. Reservoir computing beyond memory-nonlinearity trade-off. *Sci. reports* **7**, 1–10 (2017).
19. Goldmann, M., Köster, F., Lüdge, K. & Yanchuk, S. Deep time-delay reservoir computing: Dynamics and memory capacity. *Chaos: An Interdiscip. J. Nonlinear Sci.* **30**, 093124 (2020).
20. Jaeger, H. Discovering multiscale dynamical features with hierarchical echo state networks. *Deutsche Natl.* (2007).
21. Manneschi, L. *et al.* Exploiting multiple timescales in hierarchical echo state networks. *Front. Appl. Math. Stat.* **6**, 76 (2021).
22. Moon, J., Wu, Y. & Lu, W. D. Hierarchical architectures in reservoir computing systems. *Neuromorphic Comput. Eng.* **1**, 014006 (2021).
23. Gallicchio, C., Micheli, A. & Pedrelli, L. Deep reservoir computing: A critical experimental analysis. *Neurocomputing* **268**, 87–99 (2017).
24. Gallicchio, C., Micheli, A. & Pedrelli, L. Deep reservoir computing: A critical experimental analysis. *Neurocomputing* **268**, 87–99 (2017).
25. Gallicchio, C. & Micheli, A. Echo state property of deep reservoir computing networks. *Cogn. Comput.* **9**, 337–350 (2017).
26. Gallicchio, C., Micheli, A. & Pedrelli, L. Design of deep echo state networks. *Neural Networks* **108**, 33–47 (2018).
27. Ma, Q., Shen, L. & Cottrell, G. W. Deep-r-ens: A deep projection-encoding echo-state network. *Inf. Sci.* **511**, 152–171 (2020).
28. Thomas, A. Memristor-based neural networks. *J. Phys. D: Appl. Phys.* **46**, 093001 (2013).
29. Caravelli, F. & Carbajal, J. P. Memristors for the curious outsiders. *Technologies* **6**, DOI: 10.3390/technologies6040118 (2018).
30. Moon, J. *et al.* Temporal data classification and forecasting using a memristor-based reservoir computing system. *Nat. Electron.* **2**, 480–487 (2019).
31. Du, C. *et al.* Reservoir computing using dynamic memristors for temporal information processing. *Nat. communications* **8**, 1–10 (2017).
32. Yao, P. *et al.* Fully hardware-implemented memristor convolutional neural network. *Nature* **577**, 641–646 (2020).
33. Li, C. *et al.* Efficient and self-adaptive in-situ learning in multilayer memristor neural networks. *Nat. communications* **9**, 1–8 (2018).
34. Sui, X., Wu, Q., Liu, J., Chen, Q. & Gu, G. A review of optical neural networks. *IEEE Access* **8**, 70773–70783 (2020).
35. Grollier, J. *et al.* Neuromorphic spintronics. *Nat. Electron.* 1–11 (2020).
36. Romera, M. *et al.* Vowel recognition with four coupled spin-torque nano-oscillators. *Nature* **563**, 230–234 (2018).
37. Jensen, J. H., Folven, E. & Tufte, G. Computation in artificial spin ice. In *Artificial Life Conference Proceedings*, 15–22 (MIT Press, 2018).
38. Jensen, J. H. & Tufte, G. Reservoir computing in artificial spin ice. In *Artificial Life Conference Proceedings*, 376–383 (MIT Press, 2020).
39. Hon, K. *et al.* Numerical simulation of artificial spin ice for reservoir computing. *Appl. Phys. Express* **14**, 033001 (2021).
40. Heyderman, L. J. Spin ice devices from nanomagnets. *Nat. Nanotechnol.* **17**, 435–436, DOI: 10.1038/s41565-022-01088-2 (2022).
41. Saccone, M. & et al. Direct observation of a dynamical glass transition in a nanomagnetic artificial hopfield network. *Nat. Phys.* 18, 517–521, DOI: 10.1038/s41567-022-01538-7 (2022).

42. Arava, H. *et al.* Computational logic with square rings of nanomagnets. *Nanotechnology* 29, 265205 (2018).

43. Luo, Z. *et al.* Current-driven magnetic domain-wall logic. *Nature* 579, 214–218 (2020).

44. Sheldon, F., Kolchinsky, A. & Caravelli, F. Computational capacity of Irc, memristive and hybrid electronic reservoirs. *Phys. Rev. E* 106 (2022).

45. Wang, Z. *et al.* Resistive switching materials for information processing. *Nat. Rev. Mater.* 5, 173–195 (2020).

46. Stenning, K. D. *et al.* Magnonic bending, phase shifting and interferometry in a 2d reconfigurable nanodisk crystal. *ACS nano* (2020).

47. Haldar, A., Kumar, D. & Adeyeye, A. O. A reconfigurable waveguide for energy-efficient transmission and local manipulation of information in a nanomagnetic device. *Nat. nanotechnology* 11, 437 (2016).

48. Gartside, J. C. *et al.* Current-controlled nanomagnetic writing for reconfigurable magnonic crystals. *Commun. Phys.* 3, 1–8 (2020).

49. Grundler, D. Reconfigurable magnonics heats up. *Nat. Phys.* 11, 438 (2015).

50. Chumak, A., Serga, A. & Hillebrands, B. Magnonic crystals for data processing. *J. Phys. D: Appl. Phys.* 50, 244001 (2017).

51. Wang, Q., Chumak, A. V. & Pirro, P. Inverse-design magnonic devices. *Nat. communications* 12, 1–9 (2021).

52. Gartside, J. C. *et al.* Reconfigurable magnonic mode-hybridisation and spectral control in a bicomponent artificial spin ice. *Nat. Commun.* 12, 1–9 (2021).

53. Barman, A., Mondal, S., Sahoo, S. & De, A. Magnetization dynamics of nanoscale magnetic materials: A perspective. *J. Appl. Phys.* 128, 170901 (2020).

54. Kaffash, M. T., Lendinez, S. & Jungfleisch, M. B. Nanomagnonics with artificial spin ice. *Phys. Lett. A* 402, 127364 (2021).

55. Barman, A. *et al.* The 2021 magnonics roadmap. *J. Physics: Condens. Matter* (2021).

56. Dion, T. *et al.* Observation and control of collective spin-wave mode-hybridisation in chevron arrays and square, staircase and brickwork artificial spin ices. *arXiv preprint arXiv:2112.05354* (2021).

57. Arroo, D. M., Gartside, J. C. & Branford, W. R. Sculpting the spin-wave response of artificial spin ice via microstate selection. *Phys. Rev. B* 100, 214425 (2019).

58. Dion, T. *et al.* Tunable magnetization dynamics in artificial spin ice via shape anisotropy modification. *Phys. Rev. B* 100, 054433 (2019).

59. Vanstone, A. *et al.* Spectral-fingerprinting: Microstate readout via remanence ferromagnetic resonance in artificial spin systems. *arXiv preprint arXiv:2106.04406* (2021).

60. Chaurasiya, A. K. *et al.* Comparison of spin-wave modes in connected and disconnected artificial spin ice nanostructures using brillouin light scattering spectroscopy. *ACS nano* (2021).

61. Iacocca, E., Gliga, S. & Heinonen, O. G. Tailoring spin-wave channels in a reconfigurable artificial spin ice. *Phys. Rev. Appl.* 13, 044047 (2020).

62. Iacocca, E., Gliga, S., Stamps, R. L. & Heinonen, O. Reconfigurable wave band structure of an artificial square ice. *Phys. Rev. B* 93, 134420 (2016).

63. Gliga, S., Kákay, A., Hertel, R. & Heinonen, O. G. Spectral analysis of topological defects in an artificial spin-ice lattice. *Phys. review letters* 110, 117205 (2013).

64. Gliga, S., Iacocca, E. & Heinonen, O. G. Dynamics of reconfigurable artificial spin ice: Toward magnonic functional materials. *APL Mater.* 8, 040911 (2020).

65. Jungfleisch, M. *et al.* Dynamic response of an artificial square spin ice. *Phys. Rev. B* 93, 100401 (2016).

66. Jungfleisch, M. B. *et al.* High-frequency dynamics modulated by collective magnetization reversal in artificial spin ice. *Phys. Rev. Appl.* 8, 064026 (2017).

67. Yildiz, I. B., Jaeger, H. & Kiebel, S. J. Re-visiting the echo state property. *Neural networks* 35, 1–9 (2012).
68. Van der Sande, G., Brunner, D. & Soriano, M. C. Advances in photonic reservoir computing. *Nanophotonics* **6**, 561–576 (2017).

69. Liang, X. *et al.* Rotating neurons for all-analog implementation of cyclic reservoir computing. *Nat. communications* **13**, 1–11 (2022).

70. Skjærvø, S. H., Marrows, C. H., Stamps, R. L. & Heyderman, L. J. Advances in artificial spin ice. *Nat. Rev. Phys.* **2**, 13–28 (2020).

71. Schultheiss, K. *et al.* Excitation of whispering gallery magnons in a magnetic vortex. *Phys. review letters* **122**, 097202 (2019).

72. Macêdo, R., Macauley, G., Nascimento, F. & Stamps, R. Apparent ferromagnetism in the pinwheel artificial spin ice. *Phys. Rev. B* **98**, 014437 (2018).

73. Li, Y. *et al.* Superferromagnetism and domain-wall topologies in artificial “pinwheel” spin ice. *ACS nano* **13**, 2213–2222 (2018).

74. Mackey, M. C. & Glass, L. Oscillation and chaos in physiological control systems. *Science* **197**, 287–289 (1977).

75. Gauthier, D. J., Boltt, E., Griffith, A. & Barbosa, W. A. Next generation reservoir computing. *Nat. communications* **12**, 1–8 (2021).

76. Manneschi, L., Lin, A. C. & Vasilaki, E. Sparce: Improved learning of reservoir computing systems through sparse representations. *IEEE Transactions on Neural Networks Learn. Syst.* (2021).

77. Jaeger, H. Adaptive nonlinear system identification with echo state networks. *Adv. neural information processing systems* **15** (2002).

78. Gallicchio, C. & Micheli, A. Why layering in recurrent neural networks? a deepesn survey. In *2018 International Joint Conference on Neural Networks (IJCNN)*, 1–8 (IEEE, 2018).
Supplementary Information

Supplementary note 1 - Sample dimension variation

Figure 7. Variation of dimensions across the MS (a,b) and PW (c-f) samples.

Figure 7 shows SEM images of variation of dimensions across the MS (a,b) and PW (c-f) samples. In MS, two separate regions with widths of 124 nm and 137 nm are present. This manifests as slightly different coercive fields and resonant frequencies. In PW, a gradient of sample dimensions were fabricated across the surface of the chip. Bar subsets have slightly different widths to further enhance sample complexity. Widths range from 212 - 226 nm (thin) and 243 - 267 nm (wide) giving a broad distribution of resonant frequencies and coercive fields, enhancing output nonlinearity. Furthermore, some bars are connected (panel f) giving rise to complex magnetisation profiles and FMR spectral evolution.

Supplementary note 2 - $k_{max}$ selection for metric calculation.

Figure 8. a) sMC and b) sNL when varying how many previous inputs are used in the metric calculation. sMC profiles show characteristic humps and continual rising indicative of correlation between current and past input profiles from the periodic nature of the Mackey-Glass input equation.
Memory-capacity (MC) and nonlinearity (NL) are typically obtained by inputting a random input signal, with no correlation between consecutive inputs, to the reservoir. When calculating MC, this ensures that any observed memory effects arise from reservoir states alone. However, our measurement scheme precludes this method. Data is encoded in magnetic field amplitude and readout in-field resulting in field-dependent frequency shifts in the FMR response. For random inputs, the sharp jumps between consecutive inputs causes information to shift between outputs (e.g. shifts of 0.5 GHz spanning 25 output frequencies are observed over a 5 mT input range). Linear regression is unable to process this type of shifting leading to misleadingly low calculated MC.

As such, we use the smoothly varying Mackey-Glass equation as an input signal to calculate the metrics. This ensures that any field shifts between consecutive input are minimised. Our FMR peaks are broad with microstate information held across multiple neighbouring outputs (e.g. 7 GHz and 7.02 GHz will be collinear). A disadvantage of this approach is that the Mackey-Glass equation is quasi-periodic. In addition to memory from the ASR states, this can lead to a memory arising from the similarity between current and previous inputs. As such, the value of $K_{\text{max}}$ with which the metrics are calculated must be chosen carefully to minimise the effects of the periodic input.

Figure 8 shows the sMC and sNL of the three ASR samples. sMC of the input signal with itself is also shown. All curves show characteristic humps and continuous rising in the sMC profile due to the periodic nature of the MG input. sNL shows no variation with $K_{\text{max}}$. We chose $K_{\text{max}} = 8$ as this is at the end of the flat region on the input sMC curve i.e. this is the maximum value before self-correlation from the next period starts.

**Supplementary note 3 - Vortex induced memory amplification.**

Figure 9a-c) show the relationship between sMC and sNL. sMC and sNL are evaluated using 60 outputs against the original Mackey-Glass input. Figure 9d-f) shows the relationship for sMC$_{\text{in}}$ vs sMC$_{\text{out}}$ when calculating sMC for specific previous inputs. In all cases, a linear trend is observed. For MS, the gradient of sMC$_{\text{in}}$ vs sMC$_{\text{out}}$ stays approximately the same as sMC is evaluated to further previous inputs. For WM and PW, strong sMC$_{\text{out}}$ is observed throughout. The gradient between input and output sMC increases for further previous inputs. As such, the memory of the input signal is effectively amplified. The effect is strongest for WM. Small hints of long-term memory in the input signal translate to large contributions in the second reservoir. This provides additional information into why the ordering of reservoirs is crucial, as memory amplification at the end of the network is key for improved performance.

**Supplementary note 4 - Reservoir ordering**

Figure 10 shows the future prediction and NARMA transformation performance in a deep network when going from a reservoir with high sMC to a reservoir to low sMC. Low improvements are observed throughout. The initial reservoir response obscures short term information. The second reservoir simply mimics this response. As such, short term information is lost.

The ordering of ASR’s plays an important role: if R1 has low sMC, its response is captures short term behaviour. When this information is transferred, a high sMC R2 receives information about short-term behaviour which it can retain for longer. As such, short-term memory is retained in R1 and long-term memory is retained in R2. For high-to-low sMC, R1’s output will be obscured by history-dependence, limiting the amount of information retained about short term behaviour. When this information is transferred, R2 simply mimics this information. As such, neither R1 or R2 retain a reasonable short-term memory, only long-term correlations are present, reducing the overall sMC.

Complex MC + NL signal-transformations require higher harmonic generation and responses shifted in phase w.r.t the input. By going from a high sMC reservoir (WM) to a high sNL reservoir (sNL), both a linear and non-linear representation of the input signal is present in the overall network (linear from high sMC, non-linear from high sNL) producing a diverse set of lagged responses and higher harmonics.

**Supplementary note 5 - PNN performance for NARMA-processed Mackey-Glass**

Figure 11 shows MSE profiles for when performing a NARMA-transform on the Mackey-Glass input signal. Here, only minor improvements are observed for the PNN architecture as the MSE vs. $t$ curve is already linear for the best deep architecture.
Figure 9. sMC_{in} vs sMC_{out} when R2 is a) MS, b) WM and c) PW. High sMC and sNL are achieved when the interconnection sMC and sNL are high. d-f) sMC_{in} vs sMC_{out} when evaluating sMC on specific previous inputs from t-0 (current input) to t-6. Dashed line represents a linear fit.
**Figure 10.** Mackey-Glass future prediction and NARMA transformation in a deep network when going from a reservoir with high sMC to a reservoir with low sMC. Low improvements are seen across all tasks.

**Figure 11.** MSE profiles when performing a NARMA-transform of a Mackey-Glass input for the best single, parallel, deep and PNN architectures.