Neural networks are currently transforming the field of computer algorithms. With increasingly demanding applications it becomes clear that their execution on current computing substrates is severely limited. Addressing this bottleneck, reservoir computing was successfully implemented on a large variety of novel physical substrates. However, demonstration of hierarchical multilayer systems such as deep neural networks is lagging behind. We introduce multilayer reservoirs comprising coupled nonlinear time-delay oscillators with different dynamical responses. Such systems are excellently suited for a realization in hardware, and we analyze their performance focusing on the impact of unidirectional versus bidirectional coupling. Compared to a single layer system of equal size, three unidirectionally coupled reservoir layers improved the system’s performance in long-term prediction of the Mackey-Glass and Lorenz timeseries by more than one and three orders of magnitude, respectively. Finally, we show that such architectures conceptually correspond to deep convolutional neural networks.

Neural networks have emerged as the current disruptive computational concept. When cascading multiple neural network layers, the computational performance of such systems sets the benchmark in multiple challenging tasks. In such deep neural networks, the output of previous layers commonly serves as input of consecutive layers. Crucially, such hierarchical arrangement dramatically enhances the computational performance.

In the wake of deep-neural network’s success, it was realized that their emulation on Turing / von Neumann machines is highly inefficient. This stimulated strong interest in the realization of neural networks in physical substrates, whose architecture submits to the networks’ architectural topology. In contrast to many algorithms, such hardware-networks can be efficiently parallelized. Particularly photonic systems, which offer key advantages for such parallelization, are considered a promising future alternative.

Among the various neural network architectures, reservoir computers have emerged as especially promising for hardware implementations. Conceptually, a reservoir’s hidden layer is a complex recurrent neural network and corresponds to a high-dimensional nonlinear dynamical system. Training is restricted to the final readout layer, and hence the nonlinear dynamical system’s topology remains constant. This strongly relaxes conditions for implementations in physical substrates, resulting in a large number of realizations in nonlinear photonic and other physical systems. Yet, precisely this simplicity raised fundamental concerns regarding their cascadeability, something which only recently has partially been addressed. Comparable to deep convolutional networks, a continuous decrease of spatial-frequency in the response of higher layers appears beneficial. Until now, only hardware reservoirs consisting of a single hidden layer have been implemented physically. Here, the workhorse of the field have been nonlinear delay systems. Such time-delay reservoirs offer a compromise between good computing performance and exceptional ease of hardware implementation. Furthermore, they serve as model-systems for more complex hardware substrates.

We report on a deep reservoir scheme comprising hierarchically coupled nonlinear oscillators exhibiting dynamics on multiple different timescales. Crucially, coupling between different layers is fixed, adhering to the motive of conceptual simplicity. We find that such cascading significantly improves computational performance. Our work identifies promising avenues for increasing the computational performance of hardware-implemented neural networks, ensuring that architectural simplicity curbs the particular challenges of physically implementing networks.

In Fig. we schematically illustrate our deep TDR
concept for the case of cascading two layers. The deep TDR is governed by the following set of equations:

\[ \tau_i \dot{x}_i(t) = -x_i(t) - \delta_i y_i(t) + \beta_i \sin^2[d_i(t) + b_i] \]  
\[ \dot{y}_i(t) = x_i(t) \]  
\[ d_i(t) = x_i(t - \tau_{Di}) + \sum_{p=\pm 1} w_{i+p,i} x_{i+p}(t) + \rho_i u(t) \]  
\[ \rho_i \neq 0, \delta_1 = 0. \]

The state of the delay-coupled node in layer \( i \in \{1, 2, \ldots, l\} \) is given by \( x_i(t) \), and we use the \( \sin^2 \) nonlinearity often employed in photonic TDRs [12, 13]. Due to inertia, dynamics of hardware nonlinear nodes always experience low-pass (LP) filtering according to a fast-time constant \( \tau_i \), which can be extended to band-pass (BP) filtering when a slow-time constant \( \delta_i \) is added [14]. Each layer’s nonlinearity is weighed by bifurcation parameter \( \beta_i \), and the nonlinearity’s input contains constant bias \( b_i \) and a time-dependent drive \( d_i(t) \), see Eq. (3).

Drive \( d_i(t) \) features self-feedback delayed by \( \tau_{Di} \) and potentially bidirectional coupling to adjacent layers according to coefficients \( w_{i+1,i} \). According to Eq. (1), only the first layer is coupled to the external signal \( u(t) \). This injection signal \( u(t) \) encodes the external input \( s(t) \) according to the temporal masking procedure of TDRs [8, 15]. Crucially, here we have employed a de-synchronized information injection procedure in which each value of \( s(t) \) is set for an input-masking length of \( 0.8 \cdot \tau_{Di} \).

Essential for a hardware implementation of such a cascaded system is the underlying architecture. According to Fig. 1 and Eq. (3), nonlinear nodes \( i \) and \( i \pm 1 \) can be bidirectionally coupled; coupling to \( i \pm 1 < 1 \) is naturally eliminated. Such bidirectional architectures are common motives in delayed feedback oscillators [16]. Coefficient \( w_{i+1,i} \) therefore scales the flow of information from the subsequent (previous) layer into layer \( i \). According to Eq. (3) coupling is instantaneous and, crucially, is constant in time.

The fact that TDR-layers can be termed convolutional originates from a nonlinear dynamical node’s response to perturbations. The state of a nonlinear node in layer \( i \) is given by the temporal convolution product between its impulse response function \( h_i(t) \) and its input. Combined with a normalization of the system’s continuous time \( t \) by feedback delay \( \tau_{Di} \), this becomes

\[ \frac{t}{\tau_{Di}} = n + \frac{\sigma_i}{N_i}, \sigma_i \in \{1, N_i\}, n \in \{1, 2, \ldots\}, \]  
\[ x_i^{\sigma_i}(n) = \int_{-\infty}^{n+\sigma_i} h_i (n + \sigma_i - \xi) \sin^2 [d_i(\xi - 1) + b_i] d\xi. \]

Firstly, Eqs. (5) and (6) map continuous time \( t \) onto positions of the virtual nodes \( \sigma_i \) and discrete time \( n \) via the temporal node-duration \( \theta_i = \tau_{Di}/N_i \) [15]. Details of this temporal embedding technique can also be found in [7, 8]. Such coupling through a temporal convolution can directly be mapped to a coupling matrix of spatio-temporal systems more commonly used in deep neural networks [8, 17]. Second, according to Eq. (6), node \( \sigma_i \) convolves its input \( d_i(t) + b_i \) with its impulse response \( h_i(t) \), which therefore acts as the layer’s convolution kernel.

The analogy between cascaded TDRs and deep convolutional networks goes further. Besides hierarchy, deep convolutional neural networks commonly convolve increasingly larger neuron-populations of previous layers [1]. This operation is often associated with generalization: the importance of local features in the previous layer is reduced, while more general aspects are highlighted. In TDRs, increasing in the convolution kernel’s width corresponds to widening \( h_i(t) \), see Eq. (1). In TDRs, widening a layer’s convolution kernel can be realized by an additional low-frequency cut-off according to timescale \( \delta_i \) in Eqs. 1 and 2, and we enforce widening kernels by \( \delta_{i+1} \leq \delta_i \). In Fig. 2 we show the response of a two-layer deep TDR driven by the chaotic Mackey-Glass sequence (Mackey-Glass parameters identical to [2]). Responses are plotted in spatio-temporal representation as previously introduced in [18, 19], and more details about this representation and TDRs can be found in [8], and parameters are \( \beta_{2,3} = 1.1, \tau_1 = 6 \cdot 10^{-3}, \tau_{2,3} = 7 \cdot 10^{-3}, \tau_{Di} = 17.85, \Phi_{10} = 0.2, \delta_{2,3} = 0.01, w_{1,2} = 0.7, w_{2,3} = 0.8, w_{2,1} = w_{3,2} = 0 \). Nodes are arranged horizontally along \( \sigma_i \) and their temporal evolution is arranged vertically along \( n \). As we move into higher layers, consecutive layers do highlight dynamics on different timescales along \( n \) and node locations \( \sigma_i \). Our deep TDR therefore hosts features much like those taken into consideration in the design of deep convolutional networks.

Finally, creating the computational result requires to connect the deep TDR’s high-dimensional state to the output via weights adjusted during learning. Our readout layer has access to all virtual nodes of all network layers.
and the system’s output is created according to

\[ y^\text{out}_j(n) = \sum_{i} \sum_{\sigma_i} W^\text{out}_{i,\sigma_i} x^\sigma_i(n). \]  

Here, \( j \) is the dimension of the system’s output, which depends on the particular task. Common methods to obtain \( W^\text{out} \) are based on linear or ridge regression \([2, 8]\). In experimental systems, these methods can be implemented in auxiliary hardware like field-programmable gate arrays \([20]\), or can to a degree be replaced by Boolean reinforcement learning \([10]\). The recurrent neural networks discussed here are primarily relevant for processing temporal information. We therefore task the system to predict chaotic sequences \( \Delta n \) timesteps into the future, where we compare output \( y^\text{out} \) with target \( y^T(n) = s(n + \Delta n) \) and quantify the prediction’s quality via the normalized mean square error (NMSE).

First we predict the chaotic Mackey-Glass delay equation with the same parameters as in \([2]\), in particular with a delay of 17 timesteps. By predicting twice the original system’s delay into the future (\( \Delta n = 34 \)) we test our system in long-term prediction. For ease of interpretability we cascade two TDR layers consisting of \( N_1 = 600 \) nonlinear nodes, and exhaustively scan system parameters to obtain a systematic understanding. The resulting NMSEs are shown in Fig. 3. We keep \( \tau_1 = 0.6 \cdot 10^{-3}, \tau_2 = 0.6 \cdot 10^{-3}, \tau_{D1.2} = 12, b_{1,2} = 0.2, \rho_1 = 8 \) and \( \delta_2 = 0.01 \) constant, with their values mostly based on empirical observations. In order to provide a baseline-reference for other topologies, we evaluate uncoupled layers (\( w_{1,2} = w_{2,1} = 0 \)) and scan the bifurcation parameter-plane \((\beta_1, \beta_2)\), see Fig. 3(a). Importantly, for this test we set \( \rho_2 = \rho_1 \), hence coupling the BP-layer to the same input as the LP layer. We find a clear optimum for \( \beta_1 \), while performance dependence on \( \beta_2 \) is less pronounced as long as it remains sufficiently small to avoid bifurcation. The lowest error obtained is NMSE=0.8 \cdot 10^{-5} at \( \beta_1 = 1.4 \) and \( \beta_2 = 1.2 \).

We now turn to different coupling topologies and disconnect the second layer from the system’s input information (\( \rho_2 = 0, w_{1,2} = 0.7, w_{2,1} = 0.6 \)). Figure 3(b) shows that bidirectional coupling significantly alters the optimal bifurcation parameters and results in an equally pronounced \( \beta_2 \) dependency. We obtain NMSE=0.9 \cdot 10^{-5} at \( \beta_1 = 1.4 \) and \( \beta_2 = 1.2 \), and for this topology and set of parameters the performance benefit is negligible. Continuing with the optimized value of \( \beta_i \), we now focus on the coupling topology by exhaustively scanning \( w_{1,2} \) and \( w_{2,1} \), see Fig. 3(c). The NMSE reveals some performance sensitivity upon the coupling-strength from the first to the second layer. The most important finding is, however, that there is a clear and systematic dependency upon \( w_{2,1} \): the clear global performance optimum is found for unidirectional coupling with \( w_{2,1} = 0 \). Performance is strongly increased, with a prediction error NMSE=1.6 \cdot 10^{-6} being \( \sim 5 \) times smaller than for the bidirectional and the uncoupled system, confirming the benefit of the hierarchical arrangement between consecutive layers of nonlinear oscillators also for TDRs.

To further generalize our finding, we turn to predicting the chaotic Lorenz system. The Lorenz system is a three-dimensional set of ordinary differential equations, and we used the same parameters as in \([21]\). The input information was the Lorenz system’s first dimension \( x(n) \), and the prediction target was \( y^T(n) = x(n + 1) \), hence \( \Delta n = 1 \). Results are listed in Tab. 4. Prediction performance is again enhanced by the addition of two layers in a unidirectional configuration. However, on a first glance the positive benefit appears to be smaller.

The previously implemented prediction only evaluates the systems via predicting a constant distance located \( \Delta n \) in the future. A more suited approach to determine the capacity of approximating a chaotic system’s behavior is based on the so called teacher forcing \([2]\). After training the system to predict target \( y^T(n) = s(n + 1) \) for \( n \in \{1, \ldots, 5000\} \), its input becomes \( s(\tilde{n}) = y^\text{out}(\tilde{n} - 1) \),
| Nodes p/layer | Coupling strength | LZ NMSE |
|---------------|------------------|---------|
| 1200 lp       | -                | 7.6 \cdot 10^{-5} |
| 600 lp, 600 bp| $w_{1,2} = 1.1$  | 5.7 \cdot 10^{-7} |
| 400 lp, 400 bp, 400 bp | $w_{1,2} = w_{2,3} = 1.1$ | 2.5 \cdot 10^{-7} |

TABLE I. Comparison for different architectures with total number of neurons $N = 1200$. Layers are lp=low-pass, bp=band-pass. LZ: Lorenz chaotic time series one step prediction parameters: $\tau_1 = 0.006$, $\tau_2 = \tau_3 = 0.007$, $\delta_2 = \delta_3 = 0.01$, $\beta_1 = 1.5$, $\beta_2 = \beta_3 = 1.2$.

Since now exclusively driven by its own output, the system becomes an autonomous predictor of the learned temporal sequence [2], and the autonomous evolution enables comparing its output with the original chaotic sequence over longer intervals. Such comparison creates a measure of how well the chaotic system as a whole is approximated, and not just simply a single value $\Delta n$-steps in its future.

Figure 4 shows the results for autonomous evolution for Lorenz and Mackey-Glass prediction using three cascaded TDR-layers coupled unidirectionally. Parameters are identical for the three-layer network use for Fig. 2. Under these more demanding test-conditions, the positive impact of cascaded deep TDR (red dashed data) over the single-layer TDR (blue dotted data) is apparent and is particularly strong when predicting the Lorenz system, see Fig. 3(a). For the uncoupled layers the trajectories do not only rapidly diverge, but the system simply cannot reproduce the overall character of the target system. Rather than reproducing the chaotic excursions along an attractor, the autonomous TDR quickly converges to a dynamical state resembling a limit-cycle. However, the network is capable of an excellent approximate of Lorenz chaos as soon as the three employed layers are coupled in a deep, uni-directional topology. This is also visible from the temporal divergence between the output $y^{out}(\tilde{n})$ and target $y^T(\tilde{n})$, see Fig. 4(c). The deep network shows a systematic, on average exponential divergence, while output and target are fully separated after less than one oscillatory period for the uncoupled TDR. Prediction of the chaotic Mackey-Glass system also strongly benefits from the deep unidirectional network, see Fig. 4(b,d).

To conclude, we have introduced an elegant scheme for deep convolutional networks in a simple architecture of coupled nonlinear oscillators with delay. Based on cascading TDRs featuring convolutional filters of increasing width we imitate information processing conditions conceptually comparable to deep convolutional neural networks. Consecutive layers feature impulse response functions of increasing width, implementing correspondingly widening convolution kernels. This enables modifying intra- and inter-layer connectivity simply by adjusting the oscillator’s time scales.

Applied to both, Mackey-Glass and Lorenz chaos prediction, our concept significantly improves the long-term prediction quality. In the case of Lorenz prediction, our introduced topology proofs to be essential. We find remarkable that we were able to predict both benchmarks without adjusting the dynamical parameters $\beta_i$, $\tau_i$, $\delta_i$, etc. for the Lorenz task.

Recently, reservoirs have been demonstrated to infer a chaotic oscillator’s hidden degrees of freedom [21] and to predict the evolution of chaotic spatio-temporal systems far into the future [22]. Our architecture could significantly improve performances and simultaneously illustrates realistic strategies for implementing advanced systems for such tasks in hardware.

We shall point out a limitation of our approach. In our system the range of possible kernel shapes is limited by physical constraints. Moreover, here we take the kernels fixed due to the constant filtering parameters, i.e. they are not learned during the training stage. However, such optimization is possible in principle. While the significant performance increase does not yet elevate the TDR (NMSE(84)=$10^{-4.4}$) to the accuracy of the original spatio-temporal reservoir (NMSE(84)=$10^{-8.3}$) [2], multiple simple additions to the current concept could still significantly improve performance [23, 24].

Finally, we would like to point out the large variety of possible hierarchical TDR networks. Hybrid systems where for some or all layers self-feedback is removed, will incorporate feed-forward layers [23]. The large variety of dynamical states accessible with such systems offers another great opportunity. Layers featuring excitable solitons can potentially create long term memory [20], and, when combined with the reported LP and BP-layers, physically implement powerful long-shot term memory networks [27]. Deeper networks, combining layers host-
The discussed variety of different dynamical states open possibilities in new domains, among others natural language processing and sequence generation. This work was supported by the EUR EIPHI program (Contract No. ANR-17-EURE-0002), by the BiPhoProc ANR project (No. ANR-14-OHRI-0002-02), by the Volkswagen Foundation NeuroQNet project and the ENERGETIC project of Bourgogne Franche-Comté. X.P. has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 713694 (MULTIPLY).

[1] Y. LeCun, Y. Bengio, and G. Hinton, Nature 521, 436 (2015).
[2] H. Jaeger and H. Haas, Science (New York, N.Y.) 304, 78 (2004).
[3] G. Van der Sande, D. Brunner, and M. C. Soriano, Nanophotonics 6, 561 (2017).
[4] G. Tanaka, T. Yamane, J. B. Héroux, R. Nakane, N. Kanazawa, S. Takeda, H. Numata, D. Nakano, and A. Hirose, arXiv:1808.0496 (2018).
[5] C. Gallicchio and A. Micheli, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 27 (2016).
[6] C. Gallicchio and A. Micheli, arXiv preprint arXiv:1712.04323 (2017).
[7] L. Largér, A. Baylón-Fuentes, R. Martinenghi, V. S. Udaltsov, Y. K. Chembo, and M. Jacquot, Physical Review X 7, 011015 (2017).
[8] D. Brunner, B. Penkovsky, B. A. Marquez, M. Jacquot, I. Fischer, and L. Largér, Journal of Applied Physics 124, 152004 (2018).
[9] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljacic, Nature Photonics 11, 441446 (2017).
[10] J. Bueno, S. Maktoobi, L. Froehly, I. Fischer, M. Jacquot, L. Largér, and D. Brunner, Optica 5, 756 (2018).
[11] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, M. Jarrahi, and A. Ozcan, Science 26, 1 (2018).
[12] L. Largér, M. C. Soriano, D. Brunner, L. Appeltant, J. M. Gutierrez, L. Pesquera, C. R. Mirasso, and I. Fischer, Optics express 20, 3241 (2012).
[13] Y. Paquot, F. Duport, A. Smerieri, J. Dambre, B. Schrauwen, M. Haeltner, and S. Massar, Scientific reports 2, 287 (2012).
[14] L. Largér, B. Penkovsky, and Y. Maistrenko, Physical Review Letters 111, 054103 (2013).
[15] L. Appeltant, M. C. Soriano, G. V. D. Sande, J. Danckaert, S. Massar, J. Dambre, B. Schrauwen, C. R. Mirasso, I. Fischer, G. Van der Sande, J. Danckaert, S. Massar, J. Dambre, B. Schrauwen, C. R. Mirasso, and I. Fischer, Nature communications 2, 468 (2011).
[16] M. Soriano, J. García-Ojalvo, C. Mirasso, and I. Fischer, Reviews of Modern Physics 85, 421 (2013).
[17] J. D. Hart, L. Largér, T. E. Murphy, and R. Roy, arXiv preprint arXiv:1808.04596 (2018).
[18] F. T. Arecchi, G. Giacomelli, A. Lapucci, and R. Meucci, Physical Review A 45, R4225 (1992).
[19] F. Marino and G. Giacomelli, Physical Review E 98, 06201(R) (2018).
[20] M. Hermans, P. Antonik, M. Haeltner, and S. Massar, Physical Review Letters 117, 128301 (2016).
[21] Z. Lu, J. Pathak, B. Hunt, M. Girvan, R. Brockett, and E. Ott, Chaos: An Interdisciplinary Journal of Nonlinear Science 27, 041102 (2017).
[22] J. Pathak, B. Hunt, M. Girvan, Z. Lu, and E. Ott, Physical Review Letters 120, 241102 (2018).
[23] R. Martinenghi, S. Rybalko, M. Jacquot, Y. K. Chembo, and L. Largér, Physical Review Letters 108, 244101 (2012).
[24] L. Grigoryeva, J. Henriques, L. Largér, and J.-P. Ortega, Neural Networks 55, 59 (2014).
[25] S. Ortín, M. C. Soriano, L. Pesquera, D. Brunner, D. San-Martín, I. Fischer, C. R. Mirasso, and J. M. Gutiérrez, Scientific Reports 5, 14945 (2015).
[26] B. Romeira, R. Avó, J. M. L. Figueiredo, S. Barland, and J. Javaloyes, Scientific Reports 6, 19510 (2016).
[27] S. Hochreiter and J. Schmidhuber, Neural Computation 9, 1735 (1997).