Tailoring the dynamics of spintronic neural networks

Anatole Moureaux

Neuromorphic Engineering Group
Institute of Condensed Matter and Nanosciences
Université catholique de Louvain

NEnG NEUROMORPHIC ENGINEERING GROUP

IMCN

San Diego | August 18, 2024

UCLouvain

SPIE. OPTICS+ PHOTONICS

In this presentation

Spintronic neural
networks

In this presentation

What? Why? How?

Spintronic neural networks

In this presentation

What? Why? How?

Applications

Spintronic neural networks

In this presentation

What? Why? How?

Applications

Spintronic neural networks

Simulation & optimization

In this presentation

What? Why? How?

Applications

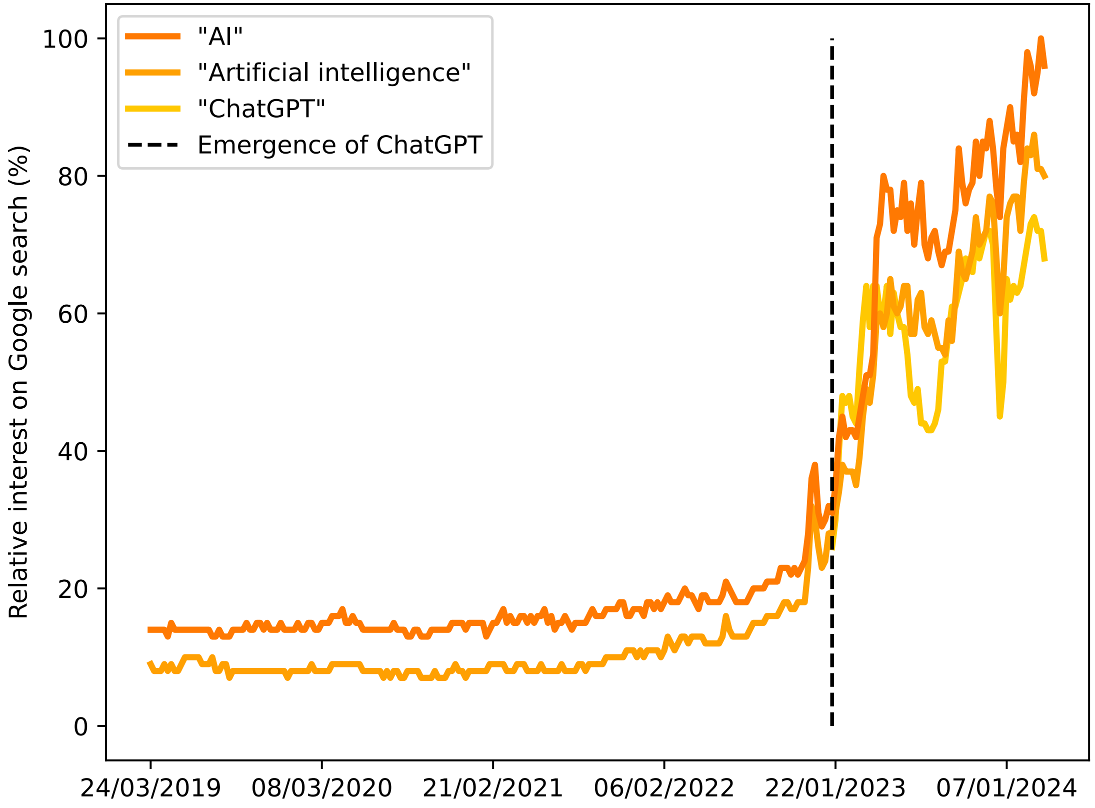
Spintronic neural networks

Perspectives

Simulation & optimization

The artificial intelligence (AI) landscape in 2024

Unprecedented popularity and use



Google Trends (2024)

Chatbots

Image & video generation

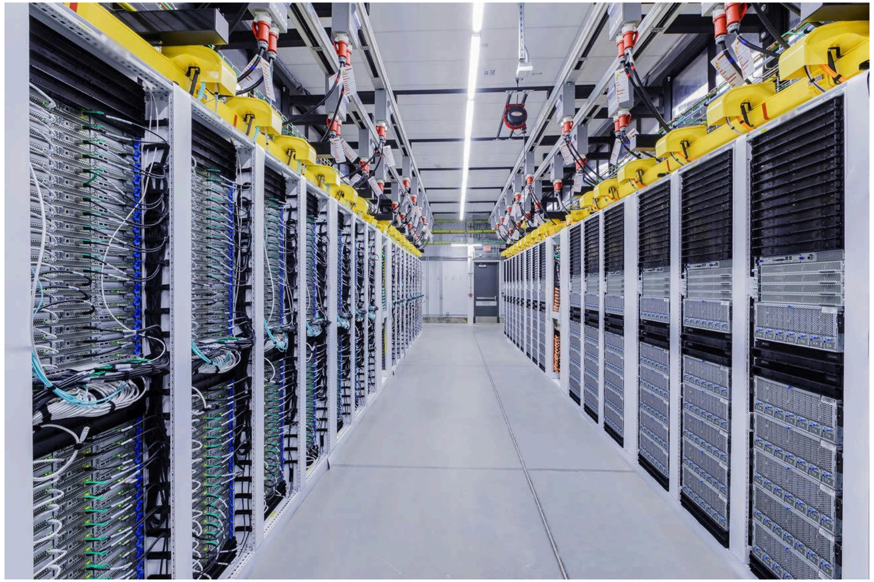
Agriculture

Healthcare

Industry

...

The shadows in the AI landscape



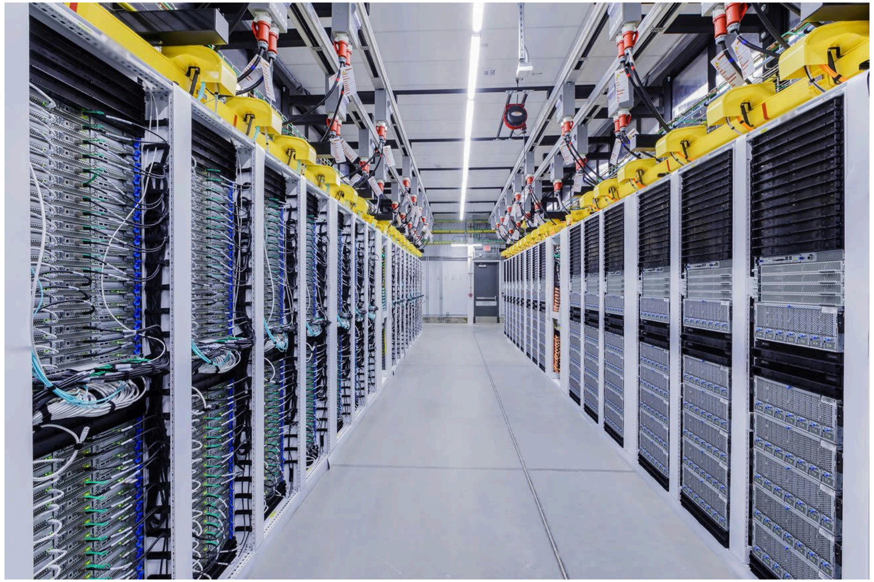
Energy consumption

- **Training** >1300 MWH (= 130 US homes for 1 year) for GPT-3
- **Inference:** 1 image with Dall-E 3 = charging a smartphone

A. Luccioni, S. Viguier, and A. Ligozat, *arXiv preprint 2211.02001* (2022)

A. Luccioni, Y. Jernite, and E. Strubell, *arXiv preprint 2311.16863* (2023)

The shadows in the AI landscape



Energy consumption

- **Training** >1300 MWH (= 130 US homes for 1 year) for GPT-3
- **Inference:** 1 image with Dall-E 3 = charging a smartphone

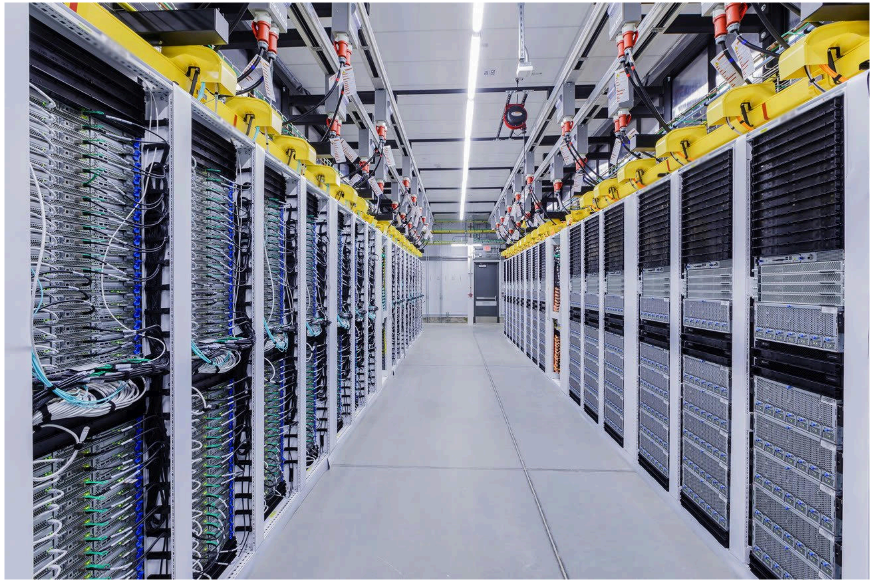
Centralization and privacy

- Queries to **remote servers** → Internet connection
- **Proprietary servers** → data protection ?

A. Luccioni, S. Viguier, and A. Ligozat, *arXiv preprint 2211.02001* (2022)

A. Luccioni, Y. Jernite, and E. Strubell, *arXiv preprint 2311.16863* (2023)

The shadows in the AI landscape



Energy consumption

- **Training** >1300 MWH (= 130 US homes for 1 year) for GPT-3
- **Inference:** 1 image with Dall-E 3 = charging a smartphone

Centralization and privacy

- Queries to **remote servers** → Internet connection
- **Proprietary servers** → data protection ?

Exponential development

- **Energy consumption > global production in 2040**
- IoT, monitoring in remote areas, agriculture → **low connectivity and power resources**

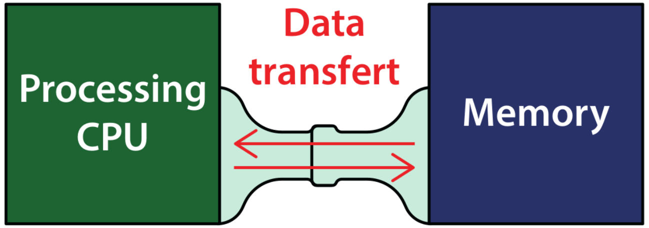
A. Luccioni, S. Viguier, and A. Ligozat, *arXiv preprint 2211.02001* (2022)

A. Luccioni, Y. Jernite, and E. Strubell, *arXiv preprint 2311.16863* (2023)

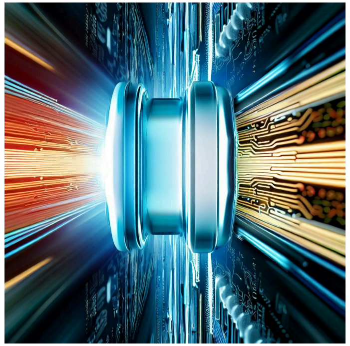
V. Sangwan, and M. Hersam, *Nature nanotechnology* **15**, 7 (2020)

Computing beyond von Neumann

von Neumann architecture



Conventional AI schemes



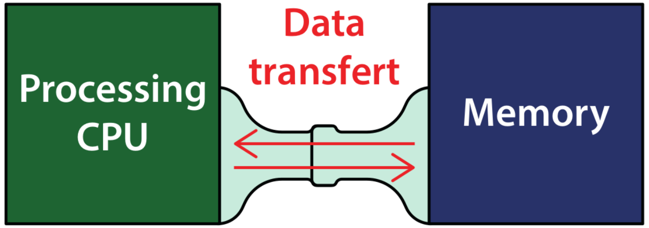
DALL·3 prompt:

"A bottleneck representing limitations due to the separation between the CPU and the RAM"

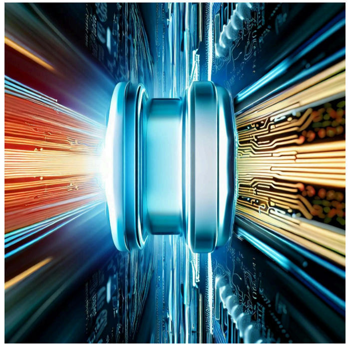
130 kW !

Computing beyond von Neumann

von Neumann architecture



Conventional AI schemes



DALL·3 prompt:

"*A bottleneck representing limitations due to the separation between the CPU and the RAM*"

130 kW !

Energy-efficient machine learning

Unconventional ML schemes

Reservoir computing, Ising machines, spiking neural networks…

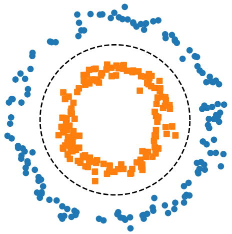
Physical implementations

Dynamical, nonlinear, probabilistic devices and systems

Outline

1. **Random nonlinear projections learning**
2. Hardware implementation using spintronic devices
3. Applications
4. High-througput modeling of spintronic neural networks
5. Tuning and optimization
6. Towards fully spintronic networks using a hardware MAC operation

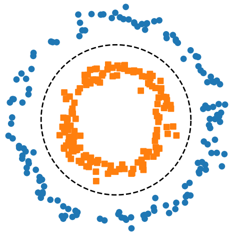
Nonlinear random projections learning



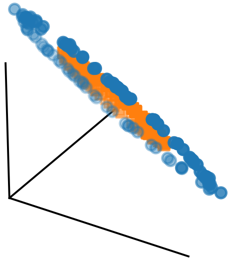
Nonlinearly separable data

G. Huang, Q. Zhu, and C. Siew, *Neurocomputing* **70**, 1-3 (2006)

Nonlinear random projections learning



$\Rightarrow$

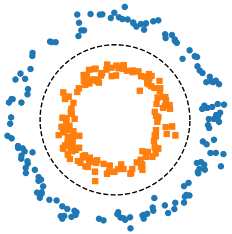


Nonlinearly separable data

Random projection in higher-dimension space

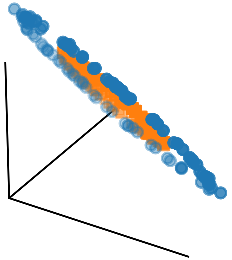
G. Huang, Q. Zhu, and C. Siew, *Neurocomputing* **70**, 1-3 (2006)

Nonlinear random projections learning



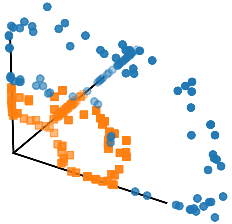
Nonlinearly separable data

⇒



Random projection in higher-dimension space

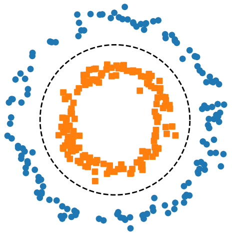
⇒



Nonlinear transformation (e.g. reLU)

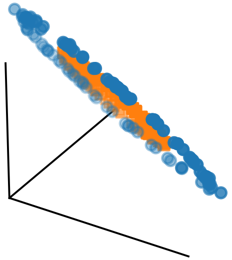
G. Huang, Q. Zhu, and C. Siew, *Neurocomputing* **70**, 1-3 (2006)

Nonlinear random projections learning



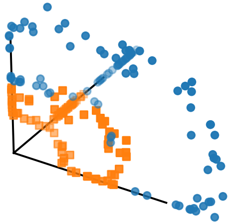
Nonlinearly separable data

⇒



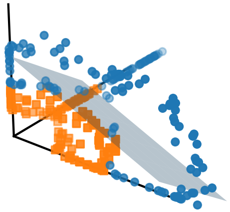
Random projection in higher-dimension space

⇒



Nonlinear transformation (e.g. reLU)

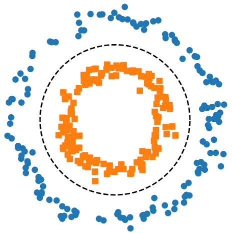
⇒



Linear regression & classification

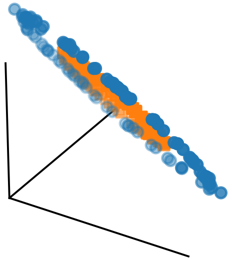
G. Huang, Q. Zhu, and C. Siew, *Neurocomputing* **70**, 1-3 (2006)

Nonlinear random projections learning



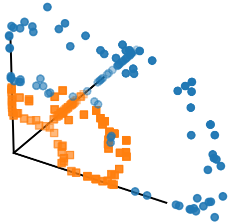
Nonlinearly separable data

⇒



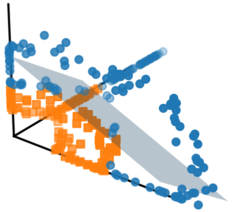
Random projection in higher-dimension space

⇒



Nonlinear transformation (e.g. reLU)

⇒



Linear regression & classification

In a nutshell

- **Classification**, **clustering**, **time series prediction**, …
- **Simple neural architectures**
- **Random and fixed** hidden weights (only the **readout weights** are learned)
- **Fast training** (linear regression through least squares optimization)
- **Global optimum** is found

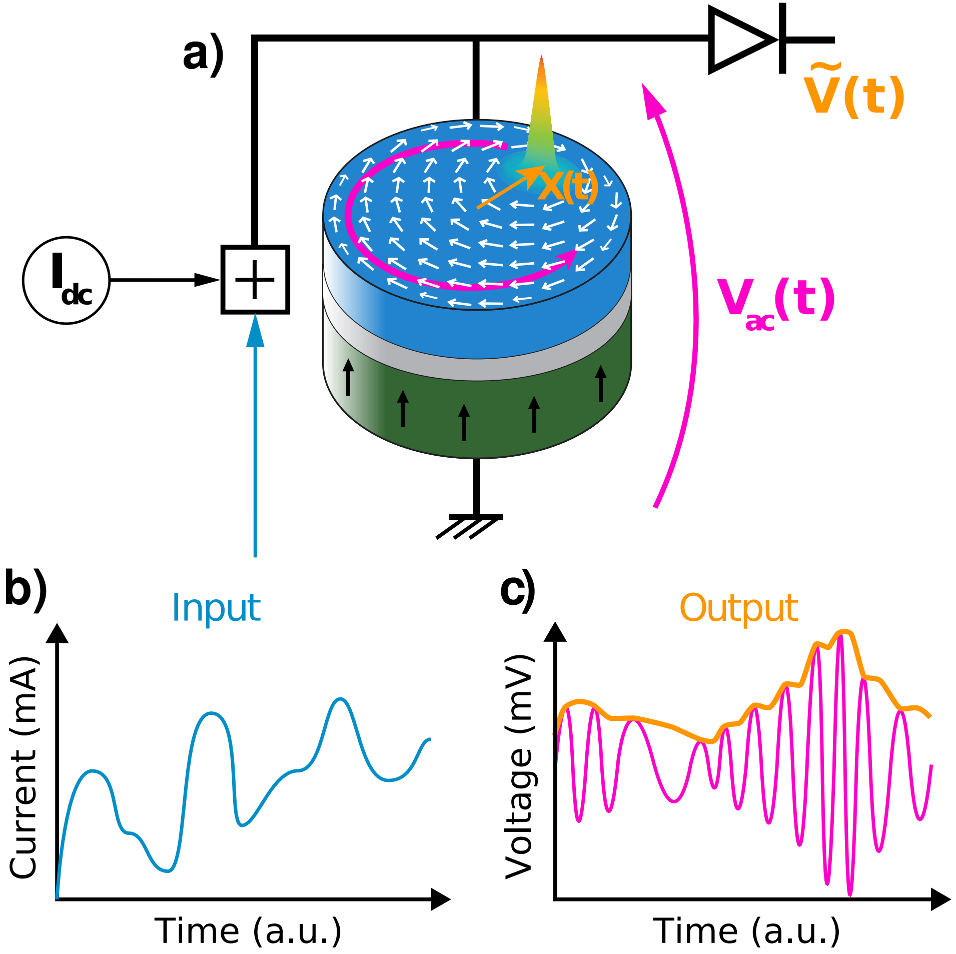
G. Huang, Q. Zhu, and C. Siew, *Neurocomputing* **70**, 1-3 (2006)

Outline

1. Random nonlinear projections learning
2. **Hardware implementation using spintronic devices**
3. Applications
4. High-througput modeling of spintronic neural networks
5. Tuning and optimization
6. Towards fully spintronic networks using a hardware MAC operation

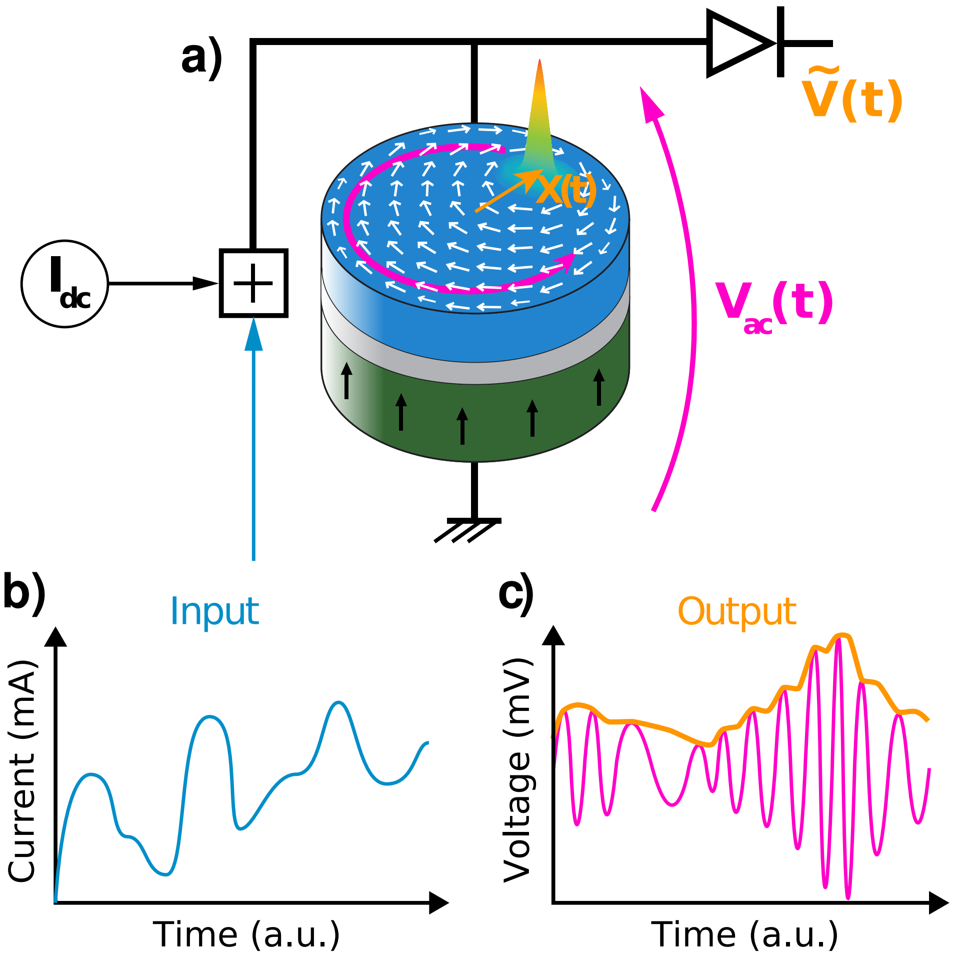
Spin-torque vortex oscillators (STVOs)

- **Magnetic tunnel junctions (MTJs):** 3-layered magnetic nanostructures
  1. Spin polarization
  2. Spin-transfer torque (STT)
  3. Tunnel magnetoresistance (TMR)
- **Nanoscale DC to AC conversion:**
- **Versatility and integrability:**



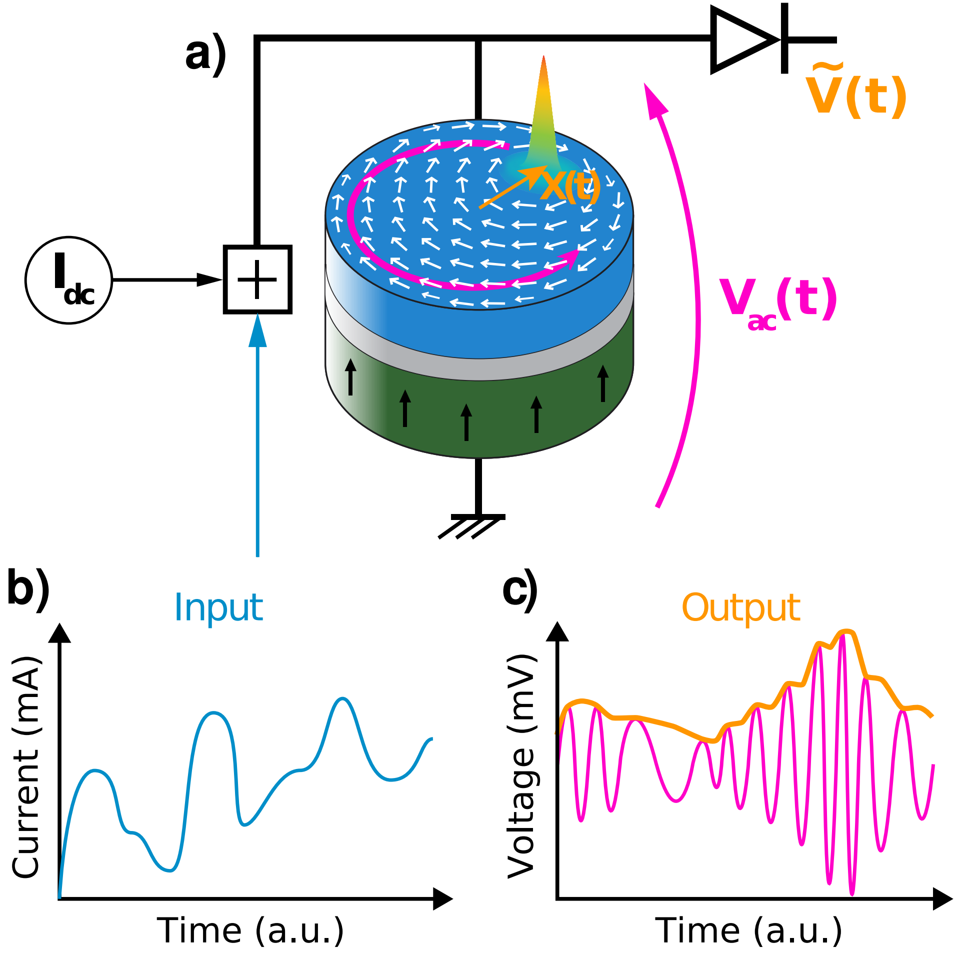
Spin-torque vortex oscillators (STVOs)

- **Magnetic tunnel junctions (MTJs):** 3-layered magnetic nanostructures
  1. Spin polarization
  2. Spin-transfer torque (STT)
  3. Tunnel magnetoresistance (TMR)
- **Nanoscale DC to AC conversion:** DC current injection triggers RF voltage oscillations through nonlinear magnetization dynamics.
- **Versatility and integrability:**

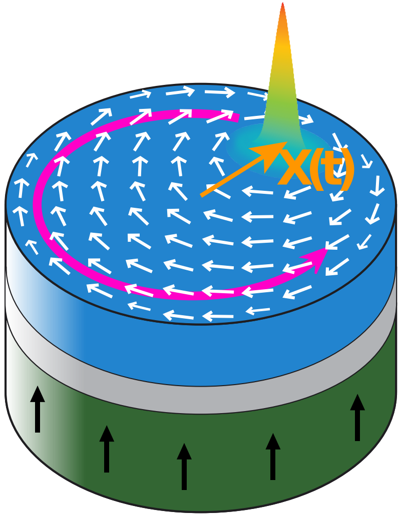


Spin-torque vortex oscillators (STVOs)

- **Magnetic tunnel junctions (MTJs):** 3-layered magnetic nanostructures
  1. Spin polarization
  2. Spin-transfer torque (STT)
  3. Tunnel magnetoresistance (TMR)
- **Nanoscale DC to AC conversion:** DC current injection triggers RF voltage oscillations through nonlinear magnetization dynamics.
- **Versatility and integrability:** CMOS-compatibility, fabrication and characterization well controlled (MRAM). Other physical phenomena exploitable.



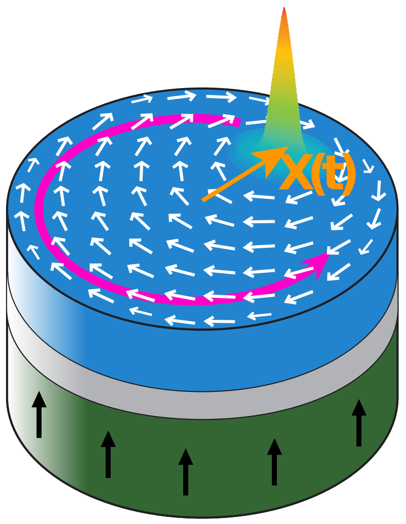
The STVO nonlinearity as an activation function



- Magnetization ground state is a *vortex*.
- STT $\Rightarrow$ circular oscillations of the *vortex core*.
- Nonlinear voltage oscillations $\propto$ orbit $s(t) = \frac{X(t)}{R}$
- Nonlinear transformation of the input signal amplitude

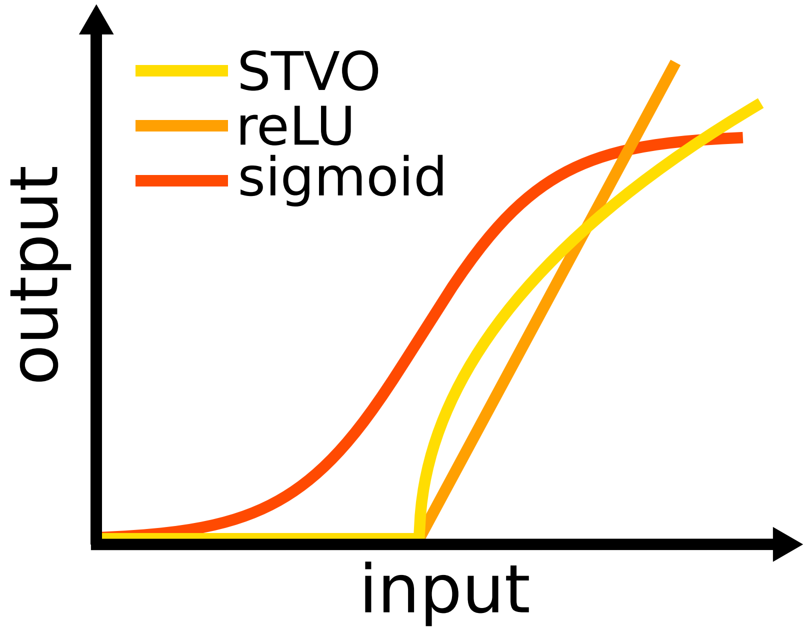
M. Romera, P. Talatchian, S. Tsunegi, F. Abreu Araujo *et al.*, *Nature* **563**, 7730 (2018)
F. Abreu Araujo, M. Riou, J. Torrejon, S. Tsunegi *et al.*, *Scientific Reports* **10**, 1 (2020)

The STVO nonlinearity as an activation function



- Magnetization ground state is a *vortex*.
- STT $\Rightarrow$ circular oscillations of the *vortex core*.
- Nonlinear voltage oscillations $\propto$ orbit $s(t) = \frac{X(t)}{R}$
- Nonlinear transformation of the input signal amplitude

Use as an activation function



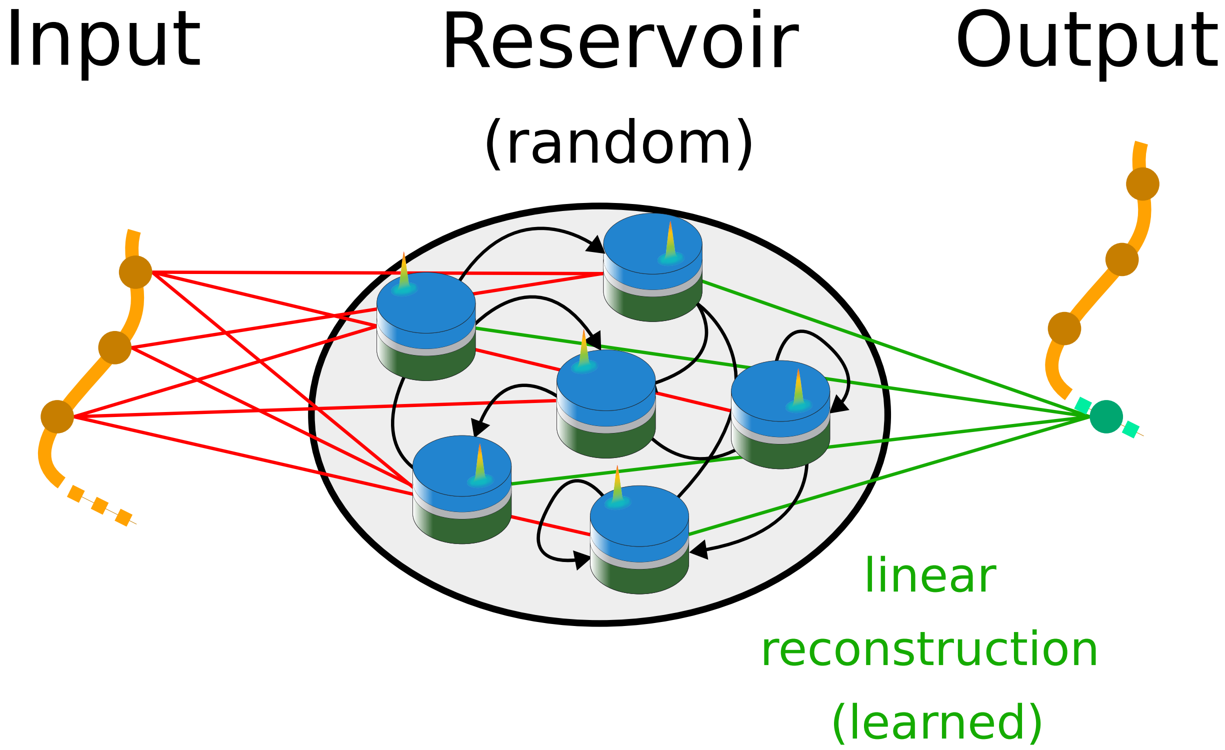
... assuming an appropriate **scaling** of the input and output signals

M. Romera, P. Talatchian, S. Tsunegi, F. Abreu Araujo *et al.*, *Nature* **563**, 7730 (2018)
F. Abreu Araujo, M. Riou, J. Torrejon, S. Tsunegi *et al.*, *Scientific Reports* **10**, 1 (2020)

Outline

1. Random nonlinear projections learning
2. Hardware implementation using spintronic devices
3. **Applications**
4. High-througput modeling of spintronic neural networks
5. Tuning and optimization
6. Towards fully spintronic networks using a hardware MAC operation

Timeseries forecasting using reservoir computing (RC)



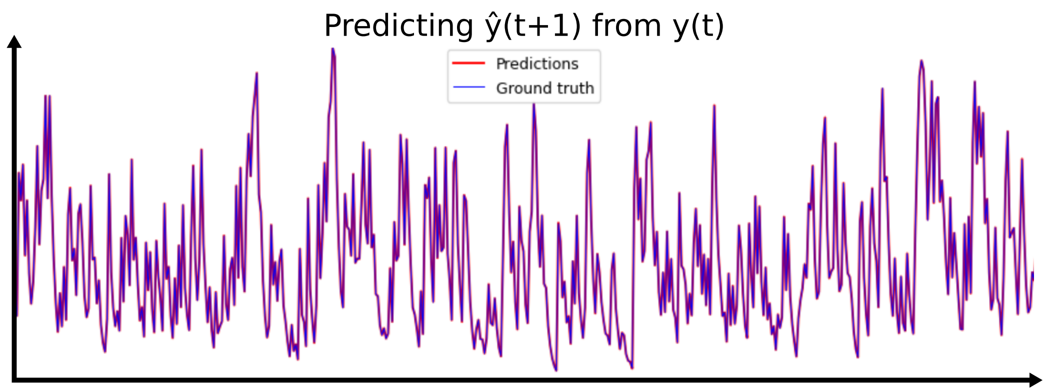
1. **Projection** of the data in a random **recurrent network** of STVOs
2. STVOs allow **short-term memory** and **nonlinear transformation**
3. Output weights are obtained by **linear regression**

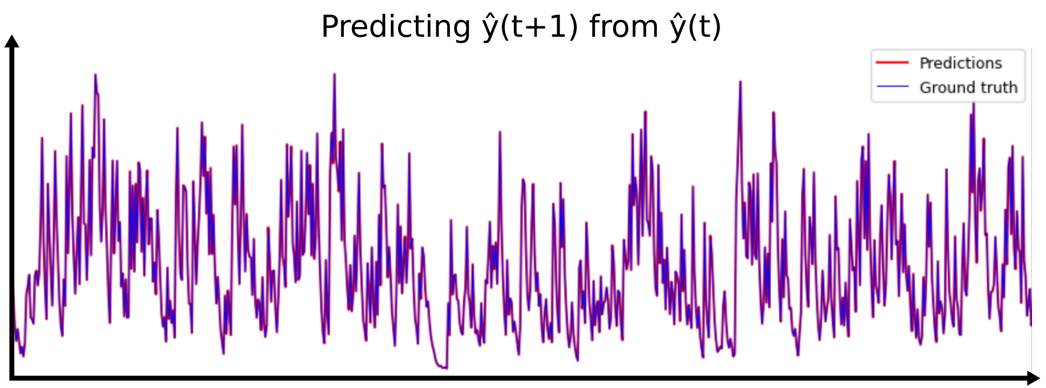
S. Shahi *et al.*, *Machine learning with applications* **8**, (2022)

Timeseries forecasting using reservoir computing (RC)

NARMA2: benchmark **memory capacity**, **predictive power** and **universal approximation capability**:

Predict $y(t+1) = \alpha y(t) + \alpha y(t)y(t-1) + \beta u(t)$ with $u(t)$ a noise contribution.





NMSE: $1.36e-04$

NMSE: $1.84e-04$

State-of-the-art performance for simulation results

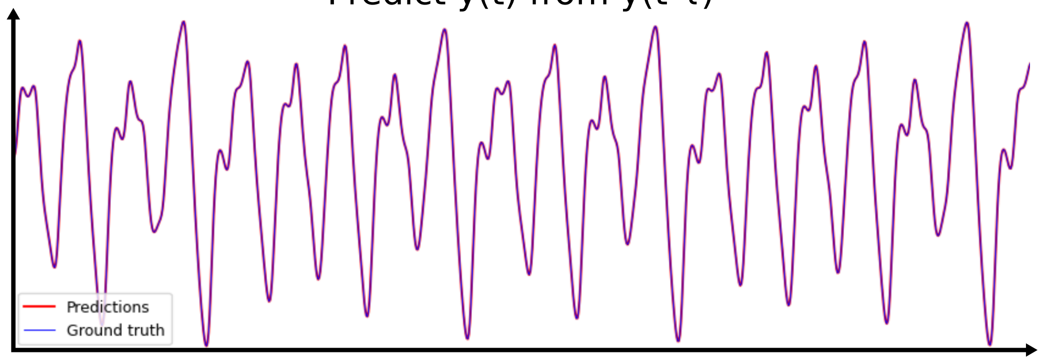
Work of our MSc. student Aurian David (aurian.david@student.uclouvain.be)

Timeseries forecasting using reservoir computing (RC)

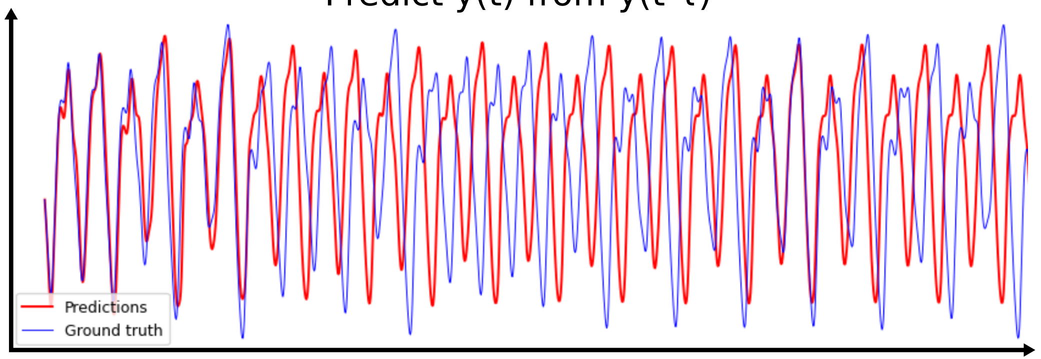
Mackey-Glass chaotic timeseries prediction: Mimicking the following chaotic system:

$$\frac{dy(t)}{dt} = \beta \frac{y(t-\tau)}{1+y(t-\tau)^n} - \gamma y(t)$$

Predict ŷ(t) from y(t-τ)



Predict ŷ(t) from ŷ(t-τ)



NMSE: $2.525e-06$

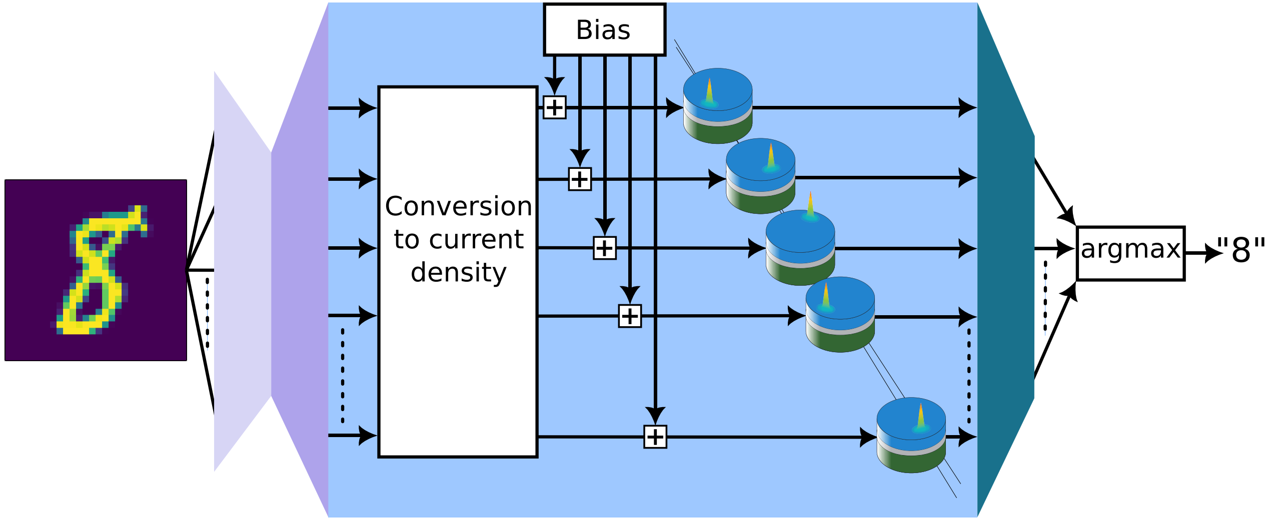
NMSE increases due to the chaotic nature of the timeseries

Work of our MSc. student Aurian David (aurian.david@student.uclouvain.be)

Data classification using extreme learning machines (ELMs)

Classification of static data

1. 1-hidden layer **perceptrons**
2. Train the readout weights by **linear regression**

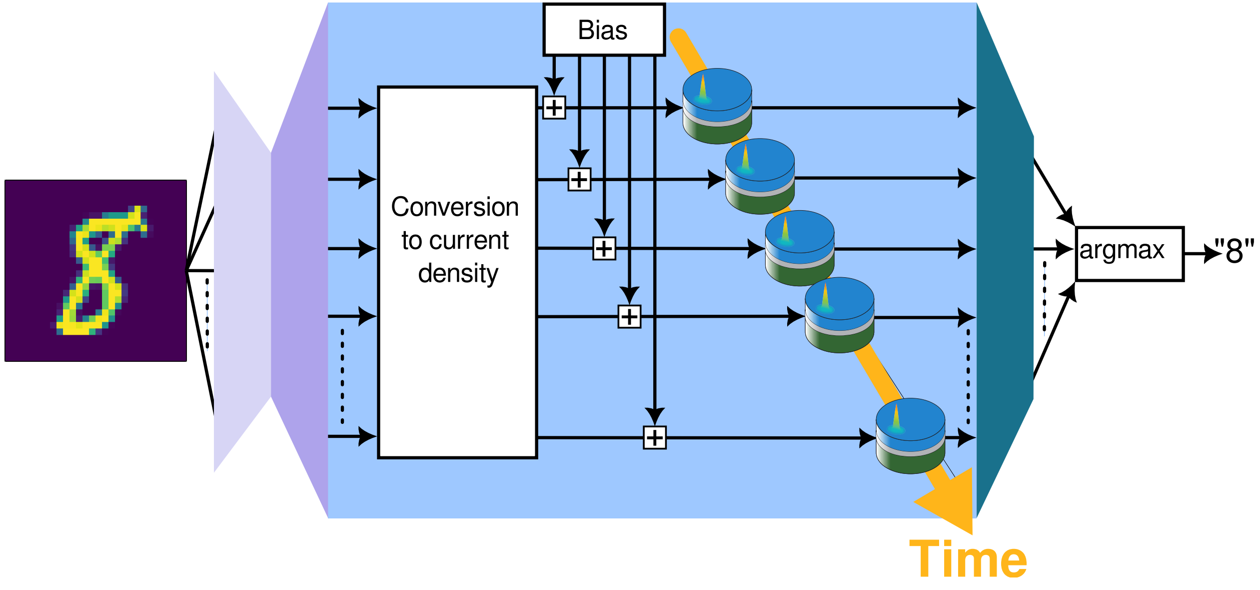


A. Moureaux, C. Chopin, S. de Wergifosse, L. Jacques, and F. Abreu Araujo, *arXiv preprint 2308.05810* (2023)

Data classification using extreme learning machines (ELMs)

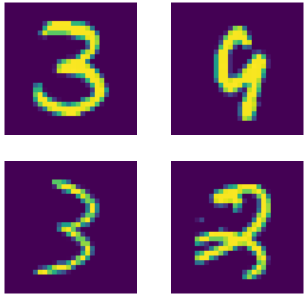
Classification of static data

1. 1-hidden layer **perceptrons**
2. Train the readout weights by **linear regression**
3. Implementation using a single STVO by leveraging **time-multiplexing**



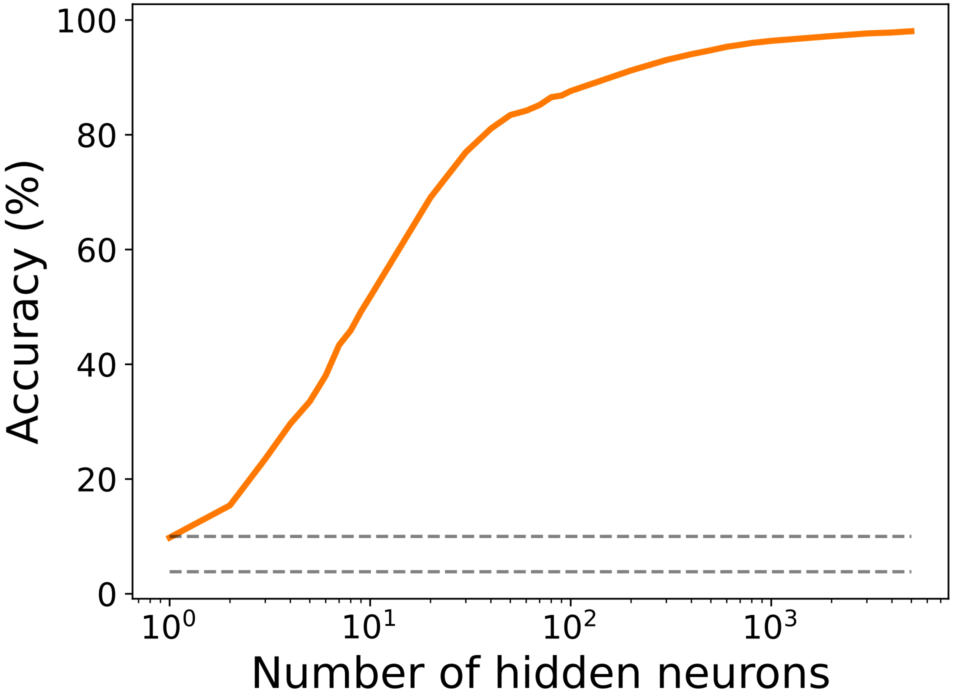
A. Moureaux, C. Chopin, S. de Wergifosse, L. Jacques, and F. Abreu Araujo, *arXiv preprint 2308.05810* (2023)

Image recognition



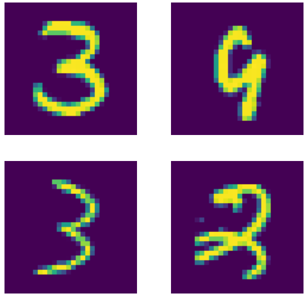
MNIST

98.1% accuracy



A. Moureaux, C. Chopin, S. de Wergifosse, L. Jacques, and F. Abreu Araujo, *arXiv preprint 2308.05810* (2023)

Image recognition

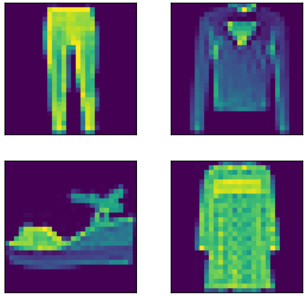


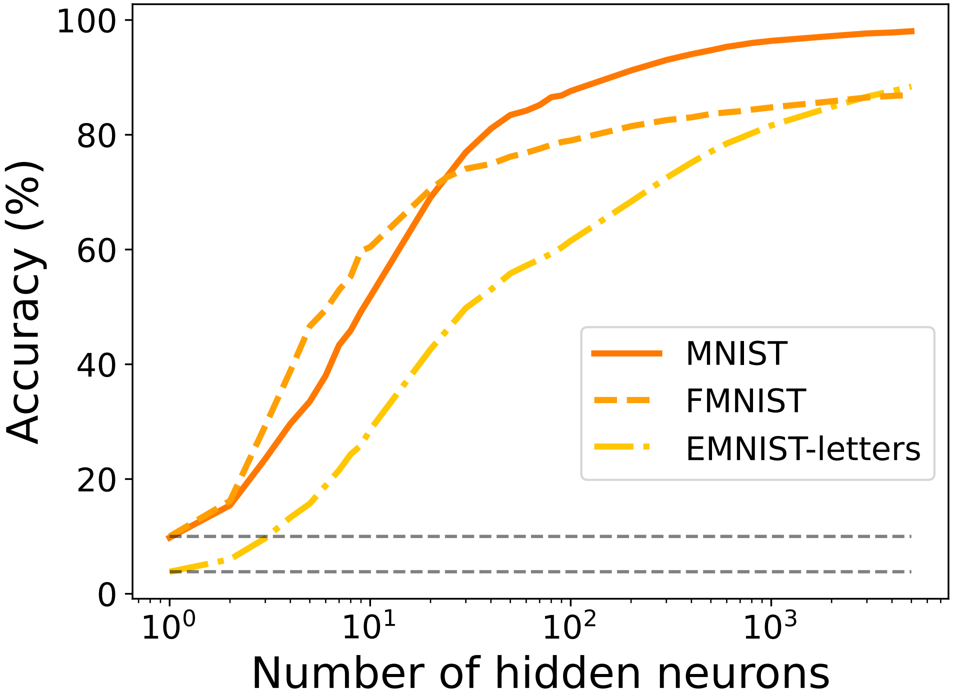
MNIST

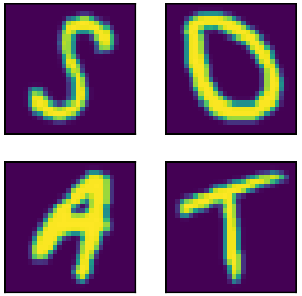
98.1% accuracy

Fashion-MNIST

86.9% accuracy







Extended-MNIST-letters

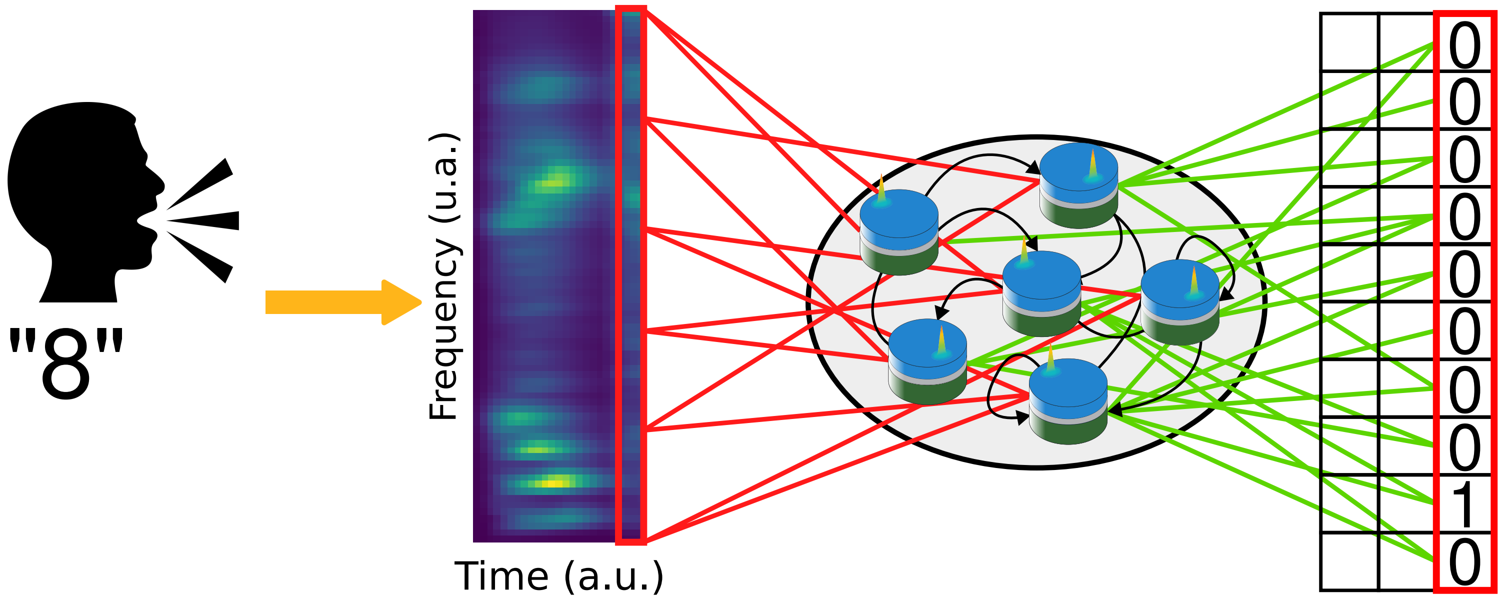
88.4% accuracy

Generalization on other datasets:
- More complex data (**FMNIST**)
- Higher number of categories (**EMNIST-letters**)

A. Moureaux, C. Chopin, S. de Wergifosse, L. Jacques, and F. Abreu Araujo, *arXiv preprint 2308.05810* (2023)

An hybrid case: speech recognition

Using frequency filtering, one can **convert speech into images** containing **temporal dependency**.



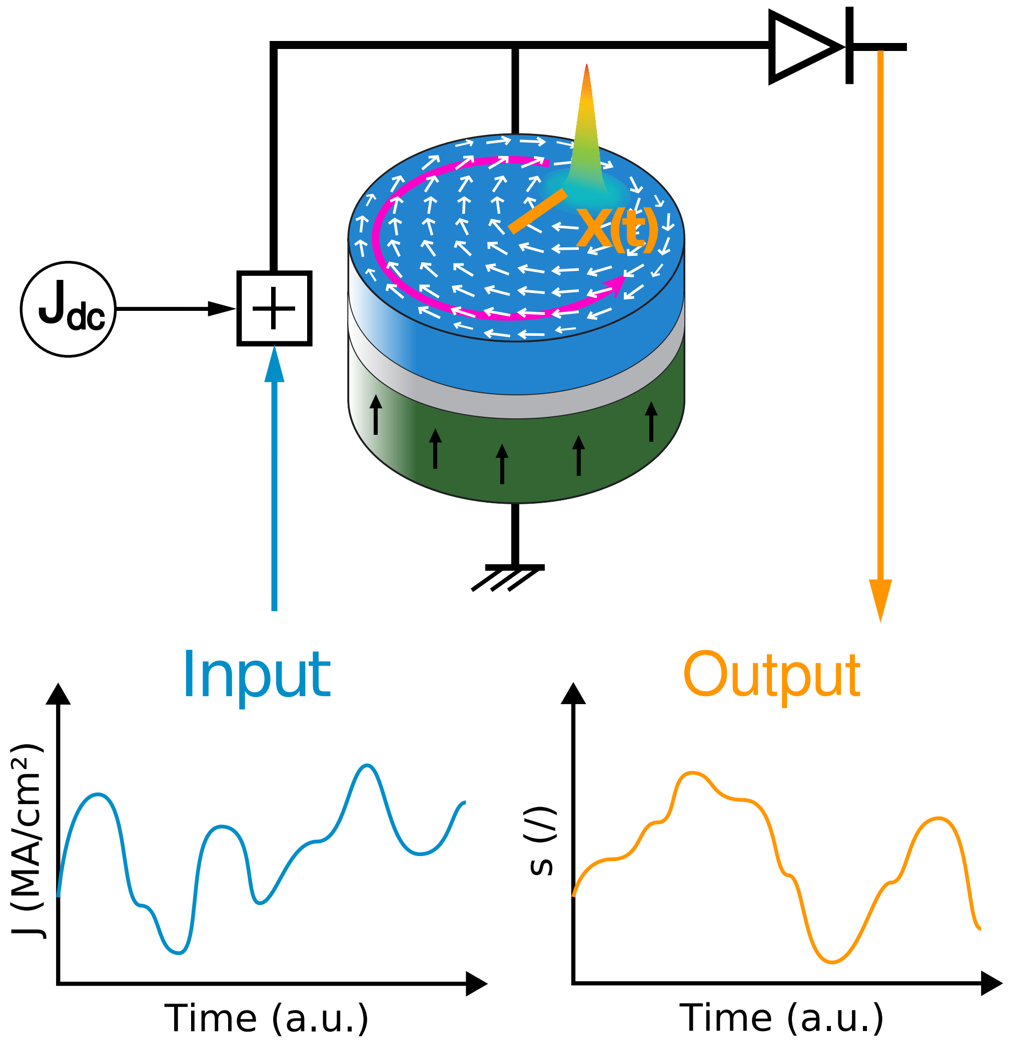
Allows classification of speech from **partial inputs**.

F. Abreu Araujo, M. Riou, J. Torrejon, S. Tsunegi, D. Querlioz *et al.*, *Scientific Reports* **10**, 1 (2020)

Outline

1. Random nonlinear projections learning
2. Hardware implementation using spintronic devices
3. Applications
4. **High-througput modeling of spintronic neural networks**
5. Tuning and optimization
6. Towards fully spintronic networks using a hardware MAC operation

Simulating the STVO dynamics



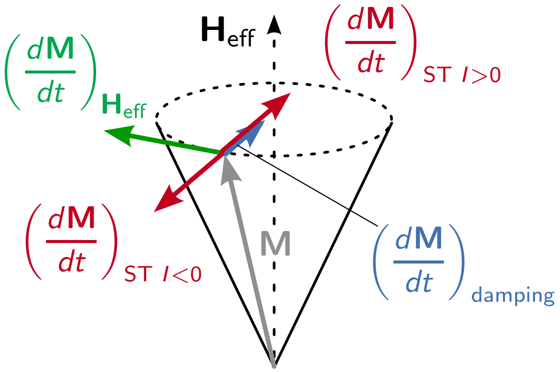
How to simulate the STVO nonlinearity?

$\Rightarrow$ Consider $s(J(t)) = \frac{X(J(t))}{R}$ the reduced position of the vortex core as a function of the time-varying input signal $J(t)$.

Numerical and analytical frameworks

Micromagnetic simulations (MMS)

Each STVO is
divided into
$\sim$ 12800 cells



then we solve:

$$\frac{d\mathrm{M}}{dt} = -\gamma \mathrm{M} \times \mathrm{H}_{\text{eff}} + \frac{\alpha}{M_s} \mathrm{M} \times (\mathrm{M} \times \mathrm{H}_{\text{eff}}) + \mathbf{\Gamma}_{\text{spin-torque}}$$

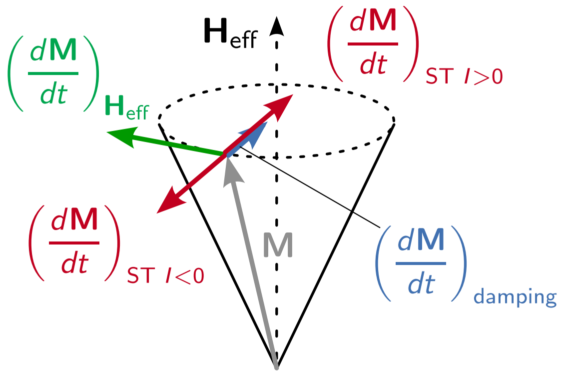
Accurate but very slow:

1s of dynamics: 440 years of simulation

Numerical and analytical frameworks

Micromagnetic simulations (MMS)

Each STVO is divided into $\sim$ 12800 cells



then we solve:

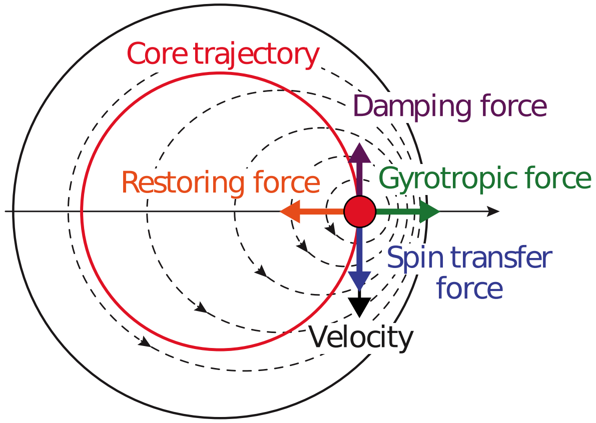
$$\frac{d\mathbf{M}}{dt} = -\gamma\mathbf{M}\times\mathbf{H}_{\mathrm{eff}} + \frac{\alpha}{M_s}\mathbf{M}\times(\mathbf{M}\times\mathbf{H}_{\mathrm{eff}}) + \mathbf{\Gamma}_{\text{spin-torque}}$$

Accurate but very slow:

1s of dynamics: 440 years of simulation

Thiele equation approach (TEA)

Vortex core considered as a quasi-particle in a balance of forces



Vortex core motion is described by:

$$G\left(\mathbf{e}_z\times\dot{\mathbf{X}}\right) + D\dot{\mathbf{X}} = \frac{\partial W}{\partial \mathbf{X}} + \mathbf{F}^{\mathrm{ST}} \text{ and } \begin{pmatrix}\dot{X}\\ \dot{Y}\end{pmatrix} = \begin{pmatrix}\Gamma & -\omega\\ \omega & \Gamma\end{pmatrix}\cdot\begin{pmatrix}X\\ Y\end{pmatrix}$$

Fast but inaccurate

The data-driven Thiele equation approach (DD-TEA)

MMS + TEA

Fit the Γ and ω parameters from the **TEA** using **MMS** results.

$\Rightarrow$ Renders the quantitative accuracy lacking in the **TEA framework**.

$\Rightarrow$ 2.4G$\times$ faster than **MMS**.

F. Abreu Araujo, C. Chopin, and S. de Wergifosse, *arXiv preprint 2206.13596* (2022)

The data-driven Thiele equation approach (DD-TEA)

MMS + TEA

Fit the Γ and ω parameters from the **TEA** using **MMS** results.

$\Rightarrow$ Renders the quantitative accuracy lacking in the **TEA framework**.

$\Rightarrow$ 2.4G$\times$ faster than **MMS**.

	MMS	**TEA**	**DD-TEA**
Accuracy			
Speed			
Neuromorphic spintronics			

F. Abreu Araujo, C. Chopin, and S. de Wergifosse, *arXiv preprint 2206.13596* (2022)

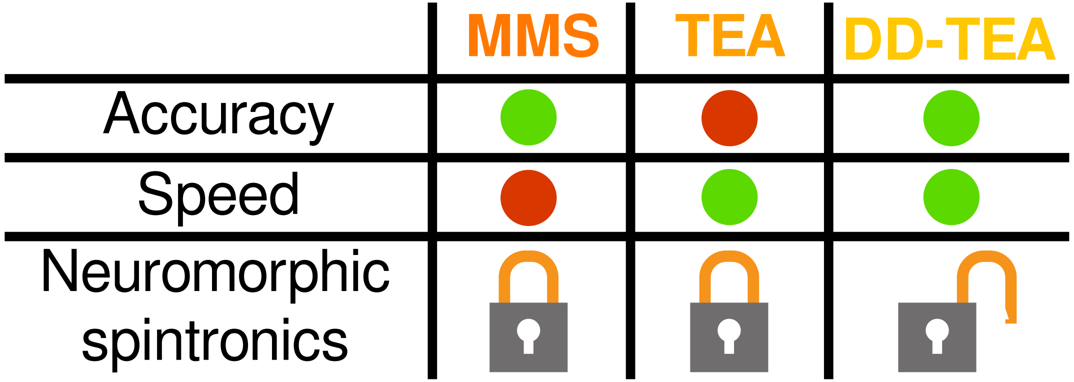
The data-driven Thiele equation approach (DD-TEA)

MMS + TEA

Fit the Γ and ω parameters from the **TEA** using **MMS** results.

$\Rightarrow$ Renders the quantitative accuracy lacking in the **TEA framework**.

$\Rightarrow$ 2.4G$\times$ faster than **MMS**.



F. Abreu Araujo, C. Chopin, and S. de Wergifosse, *arXiv preprint 2206.13596* (2022)

What's next ?

1. Use the DD-TEA model to simulate the vortex core trajectory under a given input signal.
2. Simulate STVO-based computing architectures.

3. Assess the influence of operating parameters and seek optimization.

Outline

1. Random nonlinear projections learning
2. Hardware implementation using spintronic devices
3. Applications
4. High-througput modeling of spintronic neural networks
5. **Tuning and optimization**
6. Towards fully spintronic networks using a hardware MAC operation

Parameters of interest

Physical parameters

affecting the STVO dynamics.

- **Peak-to-peak amplitude** of the input signal (mean = 0)
- **Bias** of the input signal
- **Sampling rate** of the input signal
- External magnetic field, temperature, noise, ...

Parameters of interest

Physical parameters

affecting the STVO dynamics.

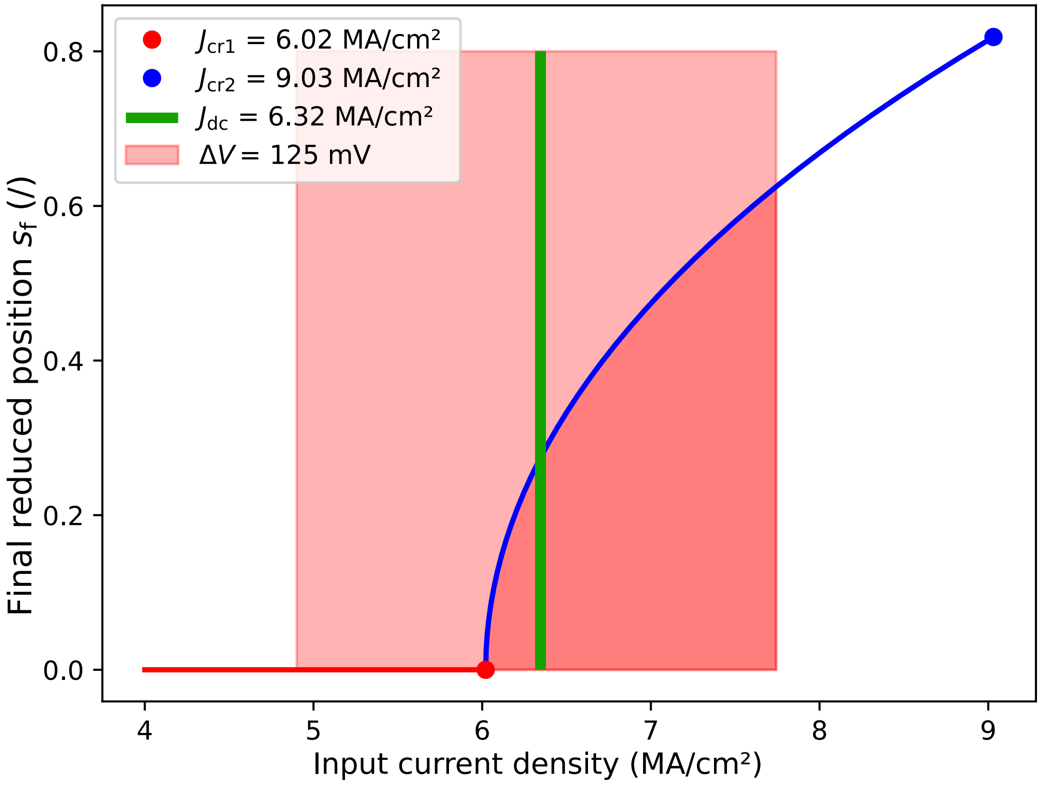
- **Peak-to-peak amplitude** of the input signal (mean = 0)
- **Bias** of the input signal
- **Sampling rate** of the input signal
- External magnetic field, temperature, noise, ...

Hyperparameters

affecting the architecture of the network

- Number of **nonlinear units** (#neurons $\propto$ #parameters)
- **Input weights** (fixed randomly)
- **Reservoir computing**: Connectivity matrix, spectral radius, ...

Physical parameters



- Input signal distributed around the bias J_{dc}
- Below J_{cr1}: no oscillations/damping regime
- Above J_{cr2}: vortex is expelled (hard limit)

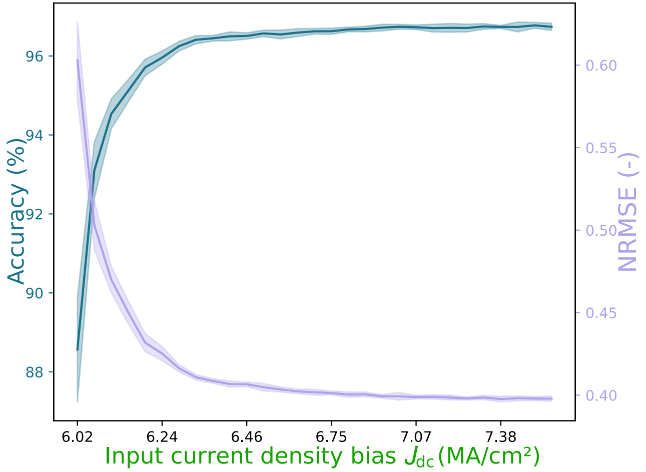
▶ **Bias current density J_{dc}**: controls the **position of the input signal** between the two critical values J_{cr1} and J_{cr2}.

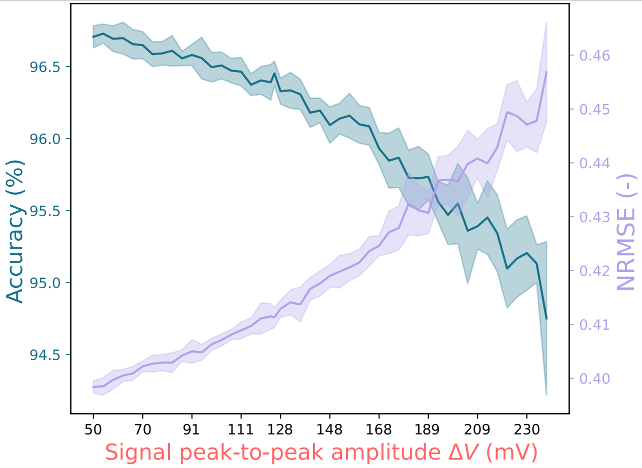
⇒ which part of the dynamics is sounded by the STVO.

▶ The **amplitude of the input signal** defines the **width of the range** of the dynamics sounded by the STVO.

⇒ "richness" of the transformation.

Image recognition (MNIST handwritten digits)

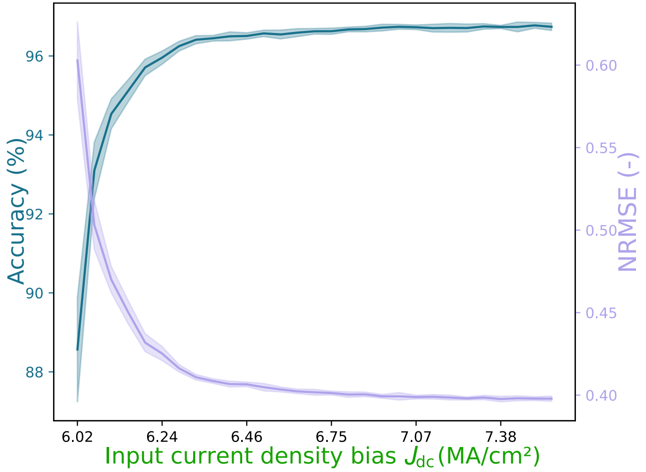


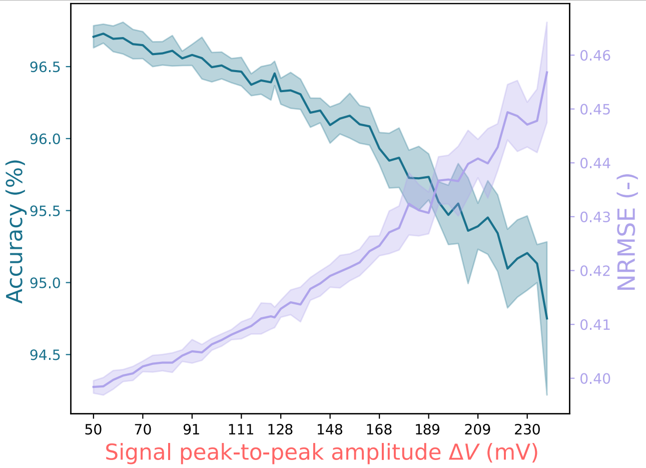


▶ Better performance (higher accuracy and lower error) at **high bias current density** and **low signal amplitude**.

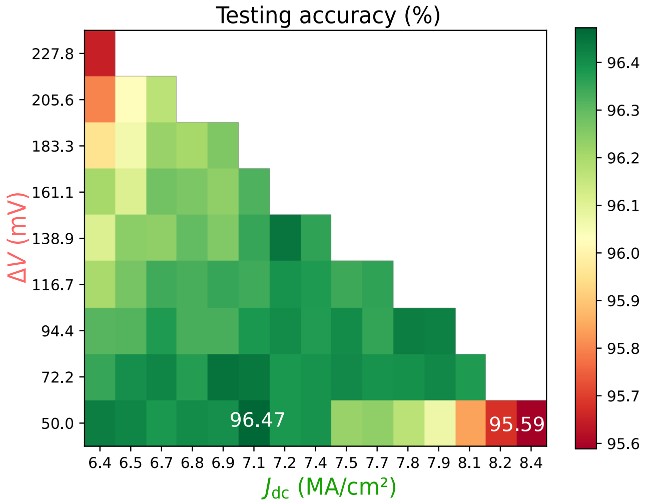
A. Moureaux, C. Chopin, S. de Wergifosse, L. Jacques, and F. Abreu Araujo, *arXiv preprint 2308.05810* (2023)

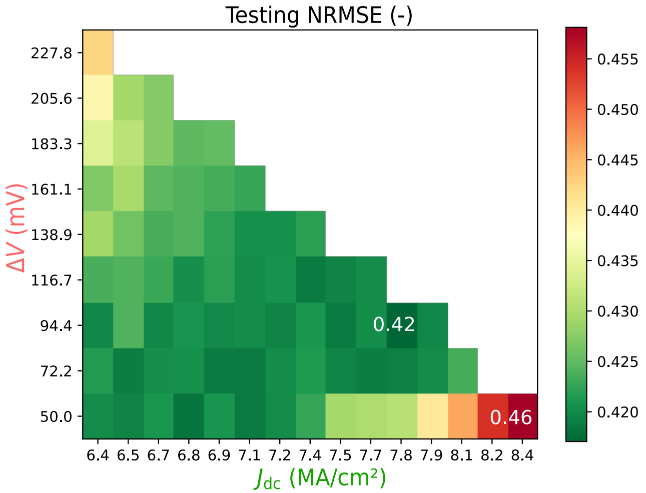
Image recognition (MNIST handwritten digits)





▶ Better performance (higher accuracy and lower error) at **high bias current density** and **low signal amplitude**.





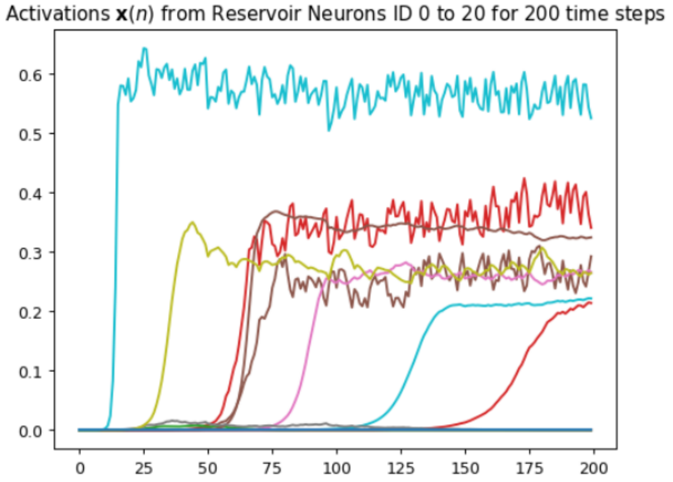
▶ However, interplay between these two parameters !

⇒ Need **joint grid search** for efficient optimization.

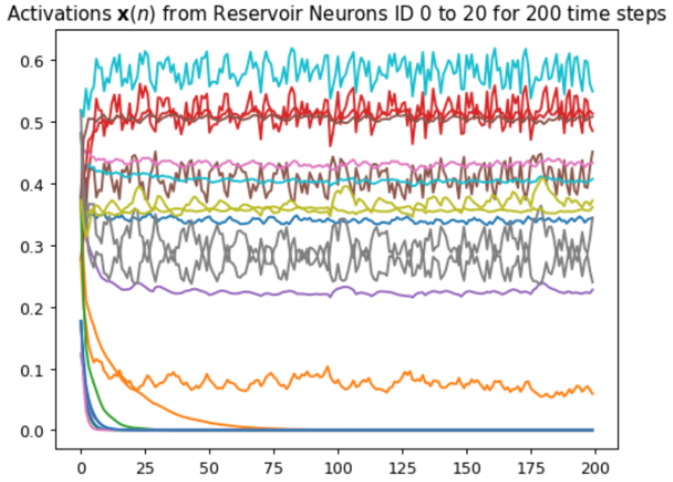
A. Moureaux, C. Chopin, S. de Wergifosse, L. Jacques, and F. Abreu Araujo, *arXiv preprint 2308.05810* (2023)

Timeseries forecasting (NARMA2)

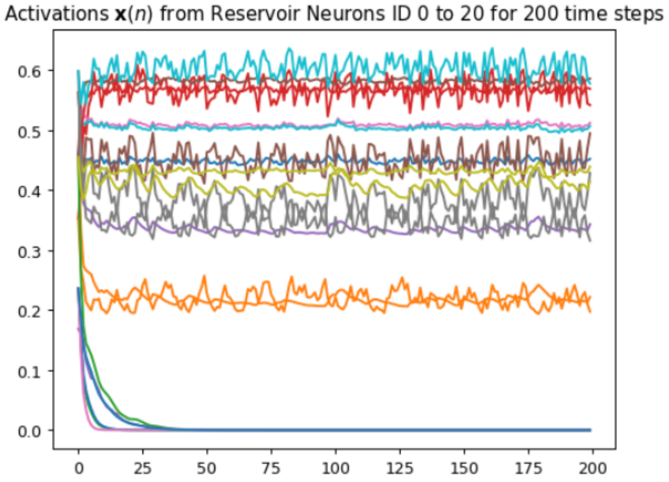
Influence of the bias current density J_{dc}



$J_{dc}/J_{cr1}=1$



$J_{dc}/J_{cr1}=1.1$



$J_{dc}/J_{cr1}=1.15$

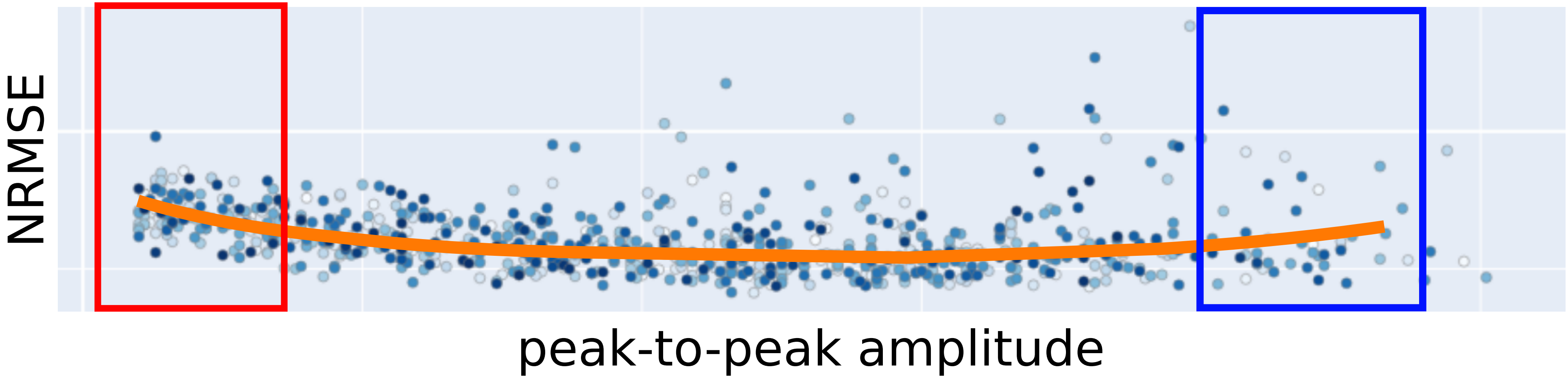
$\Rightarrow$ Lower J_{dc} values tend to delay the triggering of the oscillations.

$\Rightarrow$ Higher J_{dc} values increase the accuracy reached for the NARMA2 forecasting.

Work of our MSc. student Aurian David (aurian.david@student.uclouvain.be)

Timeseries forecasting (NARMA2)

Influence of the peak-to-peak amplitude of the input signal

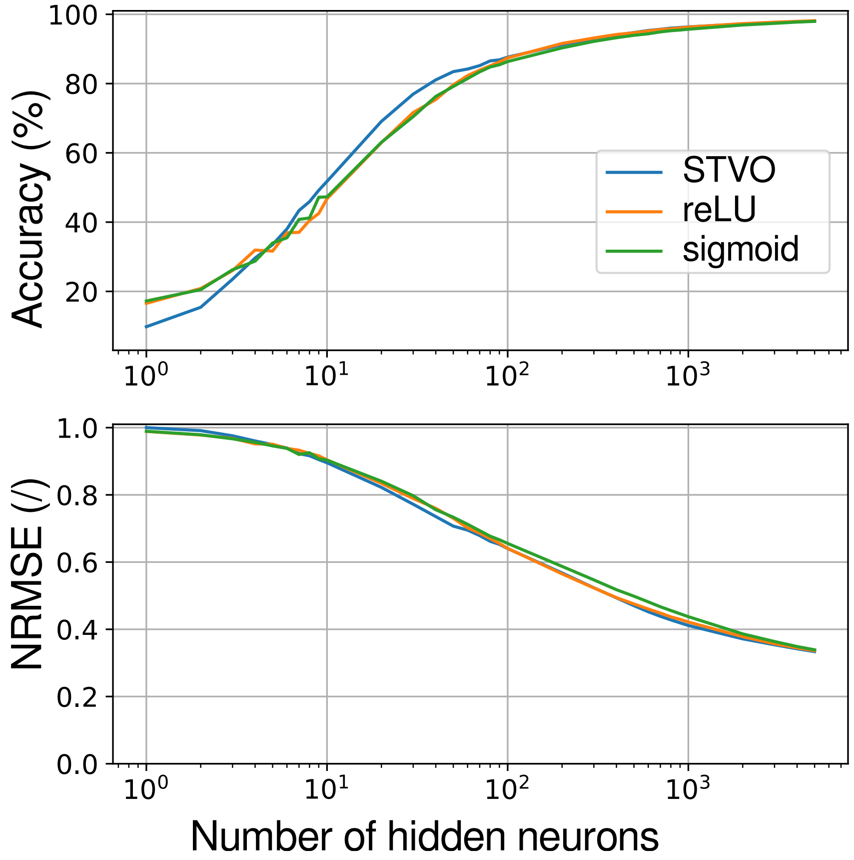


- At **small peak-to-peak amplitude**, the transformation becomes linear ⇒ prediction error ↗.
- At **high peak-to-peak amplitude**, the variance of the output ↗ ⇒ prediction error ↗.

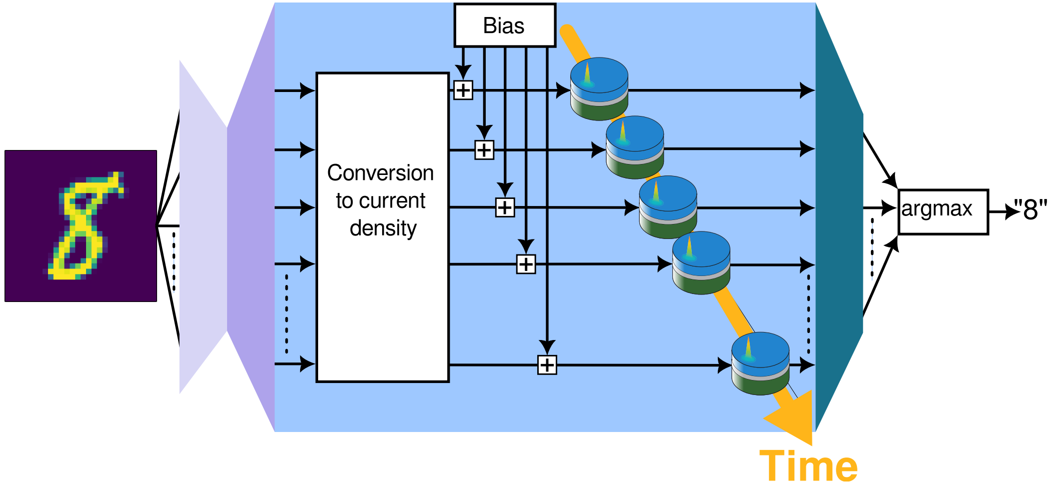
Work of our MSc. student Aurian David (aurian.david@student.uclouvain.be)

Hyperparameters

⇒ Assess the influence of architecture-related parameters



EX: MNIST handwritten digits images recognition

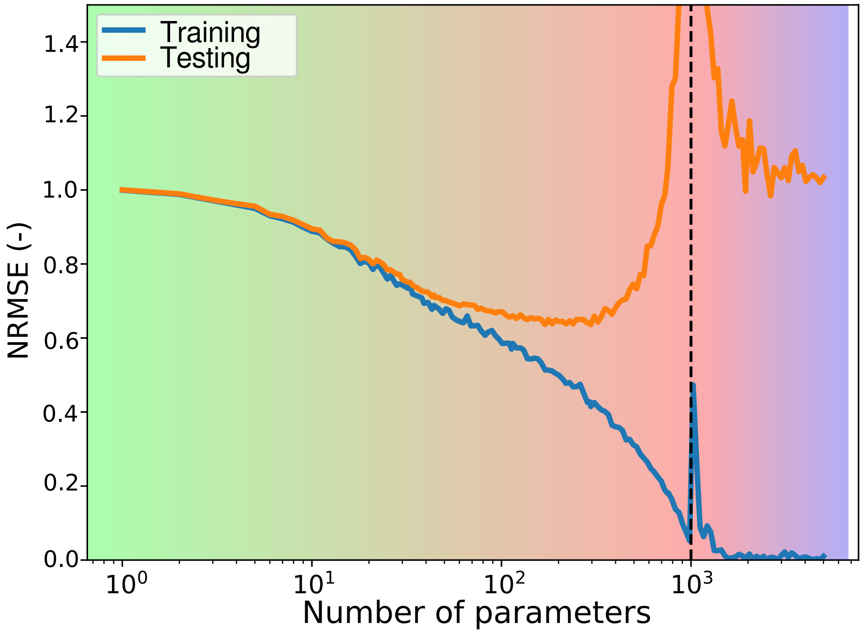


Influence of **#parameters** and comparison to **conventional nonlinear functions.**

A. Moureaux, C. Chopin, S. de Wergifosse, L. Jacques, and F. Abreu Araujo, *arXiv preprint 2308.05810* (2023)

Hyperparameters

⇒ Assess the influence of architecture-related parameters



Double-descent phenomenon:
2 successive decreases of the training NRMSE when the complexity of the network ↗

▶ **Underparametrized regime**:
NRMSE ↘ a first time when complexity ↗ as the variance of the model increases

▶ **Interpolation regime**:
NRMSE ↗ due to the maximal variance of the model (*overfitting*)

▶ **Overparametrized regime**:
NRMSE ↘ a second time due to the decrease of the model's variance (intrinsic regularization). (*undergoing work*)

P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak *et al.*, *arXiv preprint 1912.02292* (2019)

Outline

1. Random nonlinear projections learning
2. Hardware implementation using spintronic devices
3. Applications
4. High-througput modeling of spintronic neural networks
5. Tuning and optimization
6. **Towards fully spintronic networks using a hardware MAC operation**

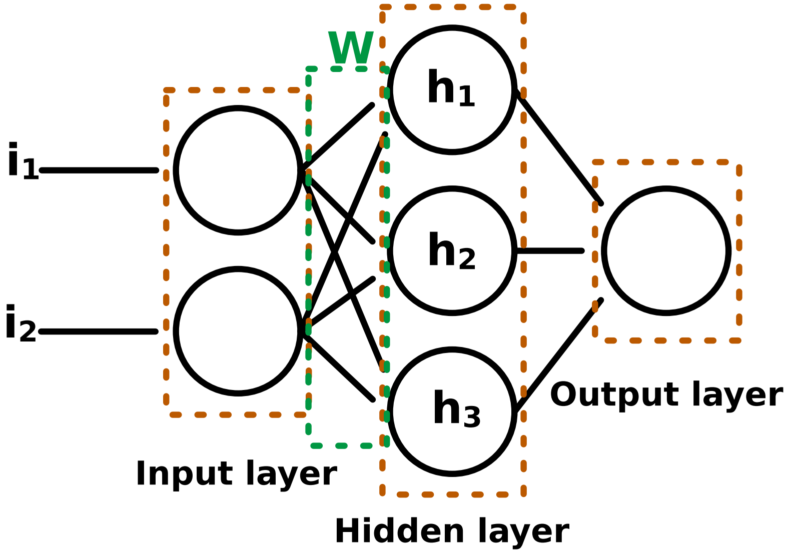
Maths in neural networks:

Nonlinear transformations

⇒ Data processing inside the nodes

Multiply-and-accumulate (MAC) operation

⇒ Propagate the information in the network



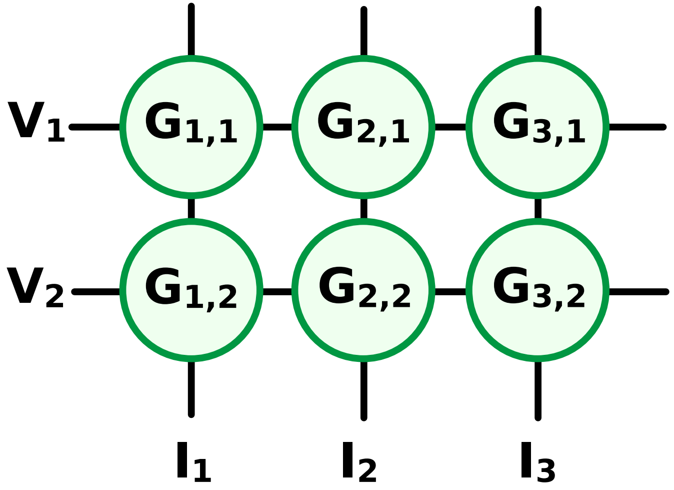
MAC operation: weighted sum for the input of a given node:

$$h_k = \boldsymbol{\sigma}\left(\sum_j w_{jk} i_j\right)$$

$$\underline{h} = \boldsymbol{\sigma}\left(\underline{\underline{w}} \cdot \underline{i}\right)$$

Implement the MAC operation in hardware

In-memory computing: perform the MAC operation in dedicated hardware where the information is already stored.



Weights are stored as conductance states G_{ij} in separate cells.

S. Jung, H. Lee *et al.*, *Nature* **601**, 7892 (2022)

1. **Input** is encoded as a voltage vector $\underline{V}$
2. **Multiplication** in each cell through Ohm's law: $I_{ij} = G_{ij} V_j$
3. **Accumulation** through Kirchhoff's law:

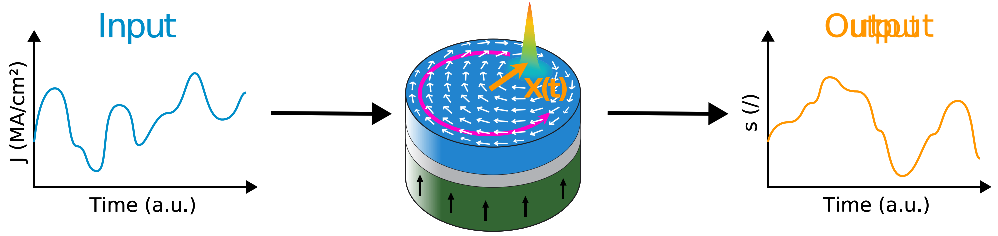
$$I_i = \sum_j I_{ij}$$

$$\underline{I} = \underline{\underline{G}} \cdot \underline{V}$$

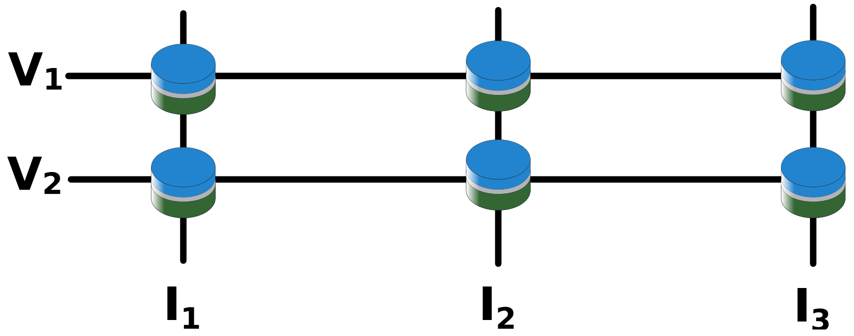
MTJs can store distinct conductance states !

Perspectives

MTJs for nonlinear transformations:

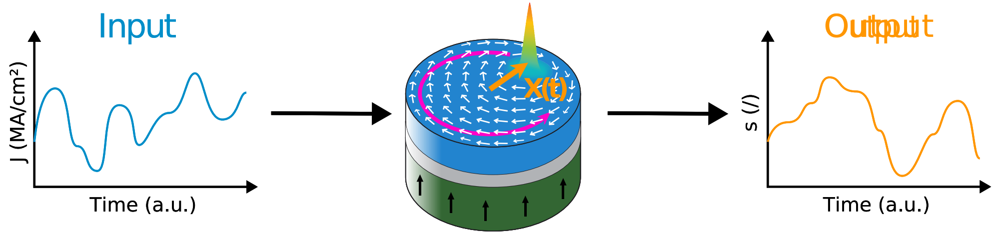


MTJs for MAC operation:

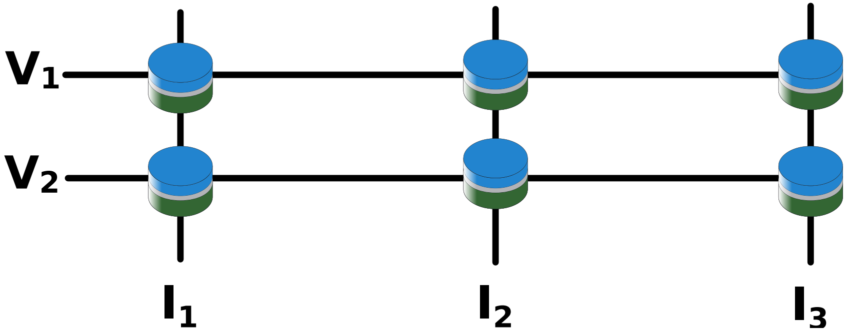


Perspectives

MTJs for nonlinear transformations:



MTJs for MAC operation:



MTJ-based coprocessor

- **Offline** training and scaling through **simulation**
- Low-power and fast **inference on the chip**
- **Versatility and adaptability**: signal forecasting, data classification, clustering, ...

Take-home message

MTJs can be used as **nonlinear units** in hardware intelligent computing systems for **timeseries forecasting** (RC) and **image classification** (ELM) at low power.

Take-home message

MTJs can be used as **nonlinear units** in hardware intelligent computing systems for **timeseries forecasting** (RC) and **image classification** (ELM) at low power.

A **data-driven approach** can be used to perform **high-throughput simulations** of entier STVO-based neural networks.

Take-home message

MTJs can be used as **nonlinear units** in hardware intelligent computing systems for **timeseries forecasting** (RC) and **image classification** (ELM) at low power.

A **data-driven approach** can be used to perform **high-throughput simulations** of entier STVO-based neural networks.

This allows to tune **physical parameters** and optimize the performance of said networks, as well as investigating the influence of **hyperparameters.**

Take-home message

MTJs can be used as **nonlinear units** in hardware intelligent computing systems for **timeseries forecasting** (RC) and **image classification** (ELM) at low power.

A **data-driven approach** can be used to perform **high-throughput simulations** of entier STVO-based neural networks.

This allows to tune **physical parameters** and optimize the performance of said networks, as well as investigating the influence of **hyperparameters.**

MTJs can also store distinct conductance states. This can be leveraged to implement a **hardware MAC operation**, the other type of computation involved in neural networks.

Acknowledgments

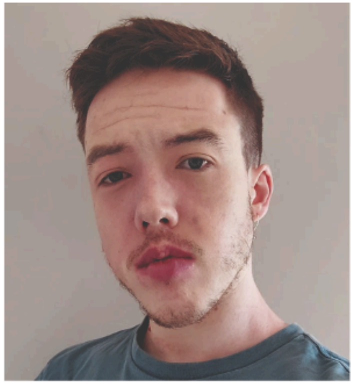
The Neuromorphic Engineering group @ UCLouvain, Belgium

Prof. Flavio ABREU ARAUJO
Group leader

Dr. Lindiomar BORGES DE AVILA Jr
Post-doctoral researcher

Dr. Tristan DA CÂMARA SANTA CLARA GOMES
Post-doctoral researcher
(currently at INESC-MN)

Simon DE WERGIFOSSE
Doctoral researcher



Anatole MOUREAUX
Doctoral researcher

NEnG
NEUROMORPHIC ENGINEERING GROUP

IMCN

UCLouvain

MULTISPIN.AI
AI Co-Processor Research Project