Contrastive Consolidation of Top-Down Modulations achieves Sparsely Supervised Continual Learning

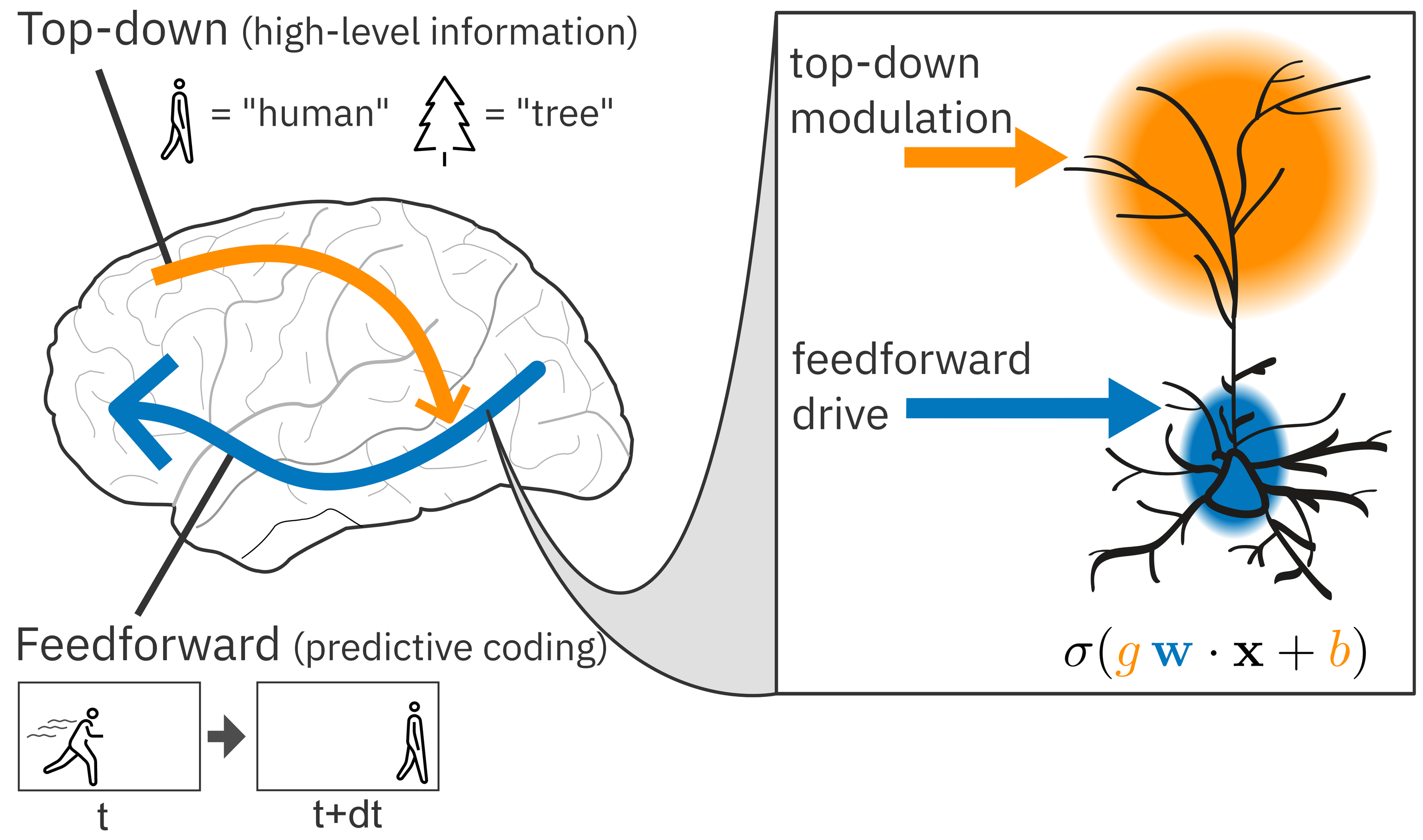
Viet Anh Khoa Tran[1,2], Emre O. Neftci[1,2], Willem A. M. Wybo[1]

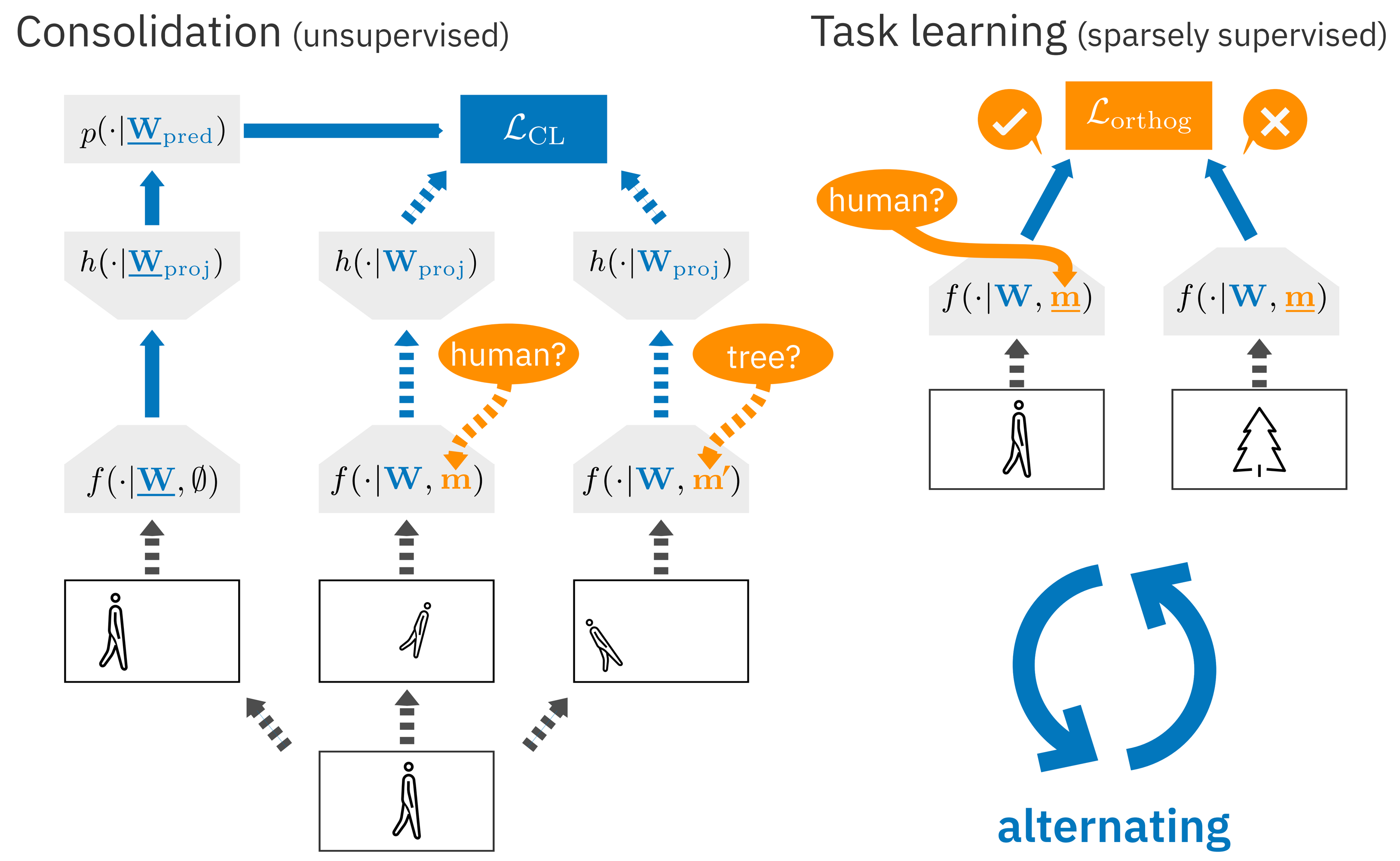
[1]Forschungszentrum Jülich [2]RWTH Aachen University

bluesky preprint

Cortical Learning is predictive coding integrating top-down modulations?

TMCL (Task-Modulated Contrastive Learning) is contrastive learning integrating task modulations.





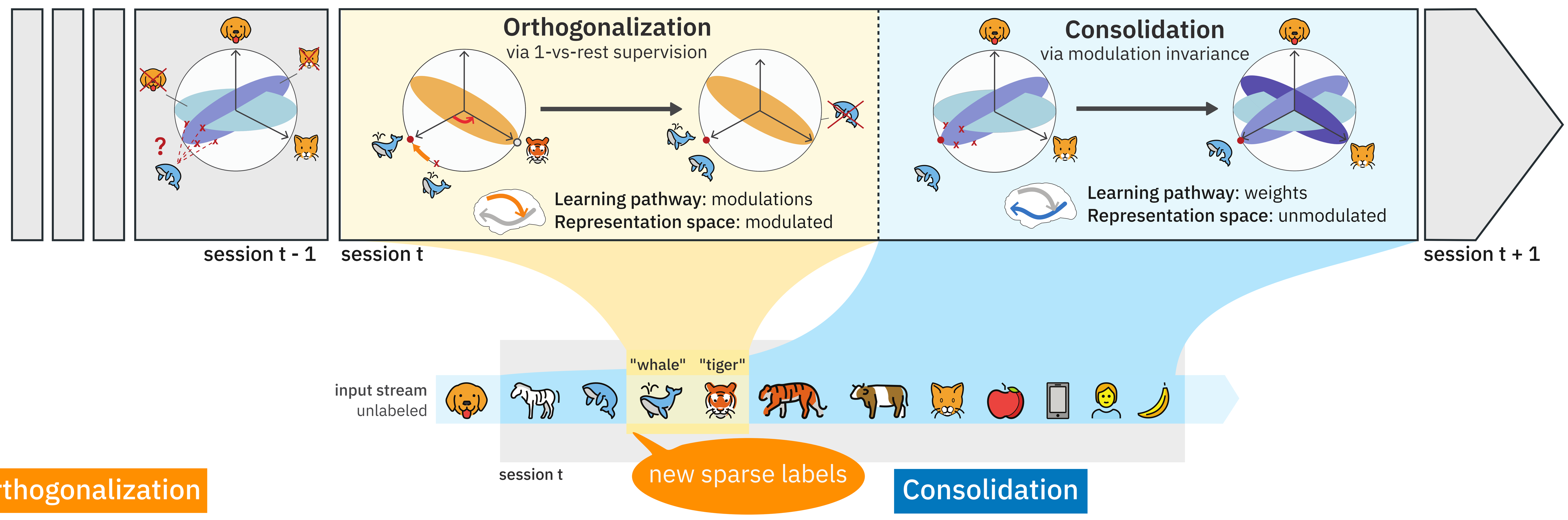
Learning in the cortex

- Cortical neurons separated into
  - a proximal, perisomatic zone (feedforward inputs)
  - and into a distal, apical region (top-down modulations).
- Segregation allows implementation of different learning rules, i.e. predictive coding and supervised learning.

Learning in machines

- The classic approach is to pre-train via unsupervised contrastive learning.
  - The pre-trained model is then usually fine-tuned with supervision on a particular task or domain, leading to catastrophic forgetting.
- Instead, TMCL **fine-tunes** supervised modulations on top of contrastively learned feedforward weights, then **consolidates** these modulations into the feedforward weights.

Continually integrating sparsely supervised modulations into weights



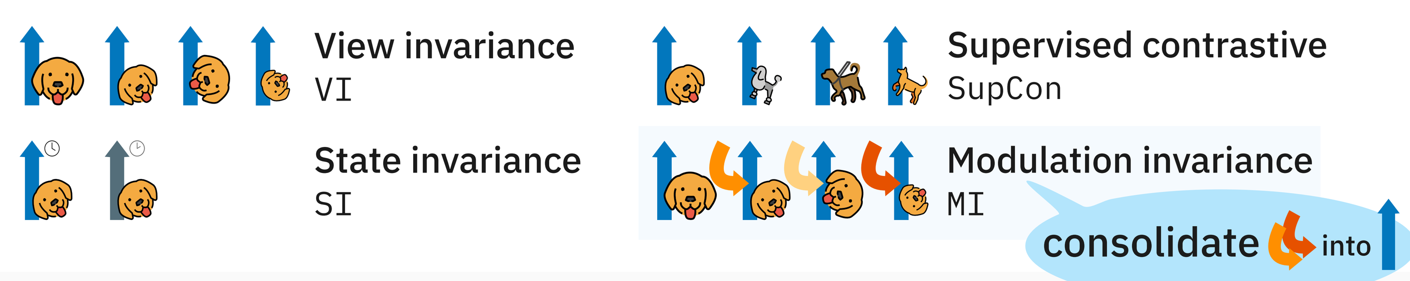
Orthogonalization

- As label for class c arrives, learn **one-vs-all modulations** on top of frozen feedforward weights.
- The orthogonal projection loss collapses samples of class c, and orthogonalizes these samples from other samples.

$$\sum_{\mathbf{p},\mathbf{p}'\in\mathbf{X}^{(c)}} \underbrace{\left(1-s_{\mathbf{m}}(\mathbf{p},\mathbf{p}')\right)}_{\text{collapse}} + \sum_{\substack{\mathbf{p}\in\mathbf{X}^{(c)}\\ \mathbf{n}\in\mathbf{X}^{(\neg c)}}} \underbrace{\left|s_{\mathbf{m}}(\mathbf{p},\mathbf{n})\right|}_{\text{orthogonalization}}.$$

Consolidation

- Use **contrastive learning** with different positives to achieve different invariances (e.g. Barlow Twins):



▶ TMCL outperforms comparable methods in **label-sparse class-incremental learning.**

incremental CIFAR-100 with 5 sessions (mean , ±std)

	Method	100% labels		10% labels		1% labels		no labels	
		kNN	linear	kNN	linear	kNN	linear	kNN	linear
sup.	SupCon	53.7±0.3	58.4±0.7	38.1±0.9	47.7±0.9	28.3±1.0	39.1±1.4	-	-
	+ SI (CaSSLe)	53.4±1.1	58.7±0.9	38.0±0.6	47.2±0.4	28.5±0.5	39.9±0.3	-	-
	CE [103]	58.9±0.5	60.1±0.5	48.3±0.4	50.7±0.3	33.9±0.3	41.6±0.5	-	-
unsup. or semi-sup.	VI	-	-	-	-	-	-	55.0±0.4	59.3±0.2
	+ SupCon	58.4±0.3	62.2±0.2	55.4±0.3	59.6±0.4	54.8±0.2	59.3±0.2	-	-
	+ CE [103]	56.8±0.5	60.6±0.6	55.4±0.4	59.5±0.3	55.4±0.5	59.8±0.2	-	-
	+ MI	56.3±0.2	60.7±0.4	56.3±0.4	**61.1**±0.3	56.1±0.5	60.7±0.3	-	-
	+ SI (PNR)	-	-	-	-	-	-	**57.1**±0.1	**60.2**±0.2
	+ SupCon	**60.6**±0.3	**62.7**±0.2	58.2±0.3	60.7±0.3	57.5±0.4	60.2±0.1	-	-
	+ CE [103]	58.8±0.1	61.2±0.3	57.8±0.1	60.3±0.6	57.5±0.1	60.1±0.3	-	-
	+ MI	58.2±0.1	60.9±0.2	**58.4**±0.3	60.7±0.2	**58.3**±0.2	**60.9**±0.2	-	-

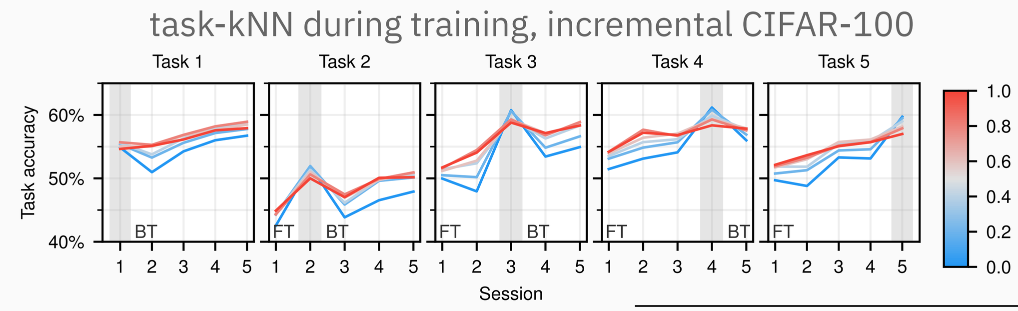
▶ TMCL continually learns generalizable representations for **transfer learning.**

kNN after incremental CIFAR-100 (mean , ±std)

	Method	Aircraft	CIFAR-10	CUBirds	DTD	EuroSAT	GTSRB	STL-10	SVHN	VGGFlower
1% CIFAR labels	SupCon	8.3±1.1	52.0±2.0	3.3±0.3	15.5±0.6	64.1±3.3	38.5±3.5	44.9±1.4	46.6±2.1	18.9±2.0
	+ SI (CaSSLe)	11.6±3.1	51.9±1.3	3.4±0.3	16.3±1.4	66.4±4.3	38.3±4.9	44.8±1.8	46.3±1.1	21.9±5.3
	VI	27.5±0.7	77.0±0.3	10.0±0.2	27.6±0.8	86.0±0.4	67.8±0.2	65.4±0.5	48.3±0.4	58.5±0.6
	+ SupCon	27.4±0.6	77.3±0.3	9.8±0.2	27.9±0.5	85.8±0.1	68.2±1.3	64.9±0.6	49.8±0.4	58.4±0.6
	+ MI	28.0±0.5	78.0±0.3	10.7±0.2	29.4±0.8	87.1±0.2	68.2±0.9	66.3±0.3	49.0±0.7	61.7±0.5
	+ SI (PNR)	28.5±0.5	78.3±0.1	11.1±0.1	28.6±0.7	87.0±0.2	69.4±0.4	67.0±0.4	49.1±0.8	64.7±0.5
	+ SupCon	29.1±0.2	78.3±0.2	10.5±0.4	28.2±0.4	87.0±0.5	**70.2**±0.2	66.8±0.3	**49.9**±0.6	64.4±0.3
	+ MI	**29.9**±0.8	**79.0**±0.2	**11.8**±0.3	**29.5**±0.3	**87.5**±0.2	69.8±0.9	**67.6**±0.3	49.7±0.7	**66.2**±0.4
100% C.l.	CE [103]	**29.0**±0.6	78.4±0.5	10.0±0.1	28.5±0.8	83.4±0.4	64.0±0.4	66.2±0.5	**53.2**±0.6	58.2±0.8
	VI	27.5±0.7	77.0±0.3	10.0±0.2	27.6±0.8	86.0±0.4	67.8±0.2	65.4±0.5	48.3±0.4	58.5±0.6
	+ SupCon	28.1±0.7	**79.1**±0.3	10.4±0.5	28.9±0.7	86.6±0.1	68.6±0.6	**66.8**±0.2	49.8±0.8	58.2±0.4
	+ MI	28.9±0.3	78.2±0.2	**10.7**±0.2	**29.2**±0.2	**87.0**±0.1	**71.0**±0.8	66.7±0.3	50.7±0.6	**62.8**±0.5

▶ Strength of TMCL (λ_{MI}) controls the **stability-plasticity tradeoff.**

task-kNN during training, incremental CIFAR-100



▶ TMCL improves backwards and forwards transfer.

average task-kNN difference vs. task-specific training (mean , ±std)

Methods	BT	FT
SupCon	-18.5±0.4	-27.7±0.4
+ SI (CaSSLe)	-18.1±0.2	-27.4±0.4
VI	-1.7±0.4	-6.5±0.4
+ MI	-0.3±0.2	-3.7±0.4
+ SupCon	-1.9±0.1	-5.9±0.3
+ SI (CaSSLe)	1.0±0.1	-4.4±0.4
+ MI	**1.3**±0.2	**-3.0**±0.3
+ SupCon	1.0±0.1	-3.7±0.1

Acknowledgements

The authors gratefully acknowledge computing time on the supercomputers JURECA and JUWELS at Forschungszentrum Jülich.

WEST AI

JÜLICH Forschungszentrum

Neuromorphic Software Ecosystems (PGI-15)

Dendritic Learning Group