Self-supervised Learning: What we can learn from nonlinear dynamics and neuroscience

FIMI2025@Okinawa, Mar 1st 2025

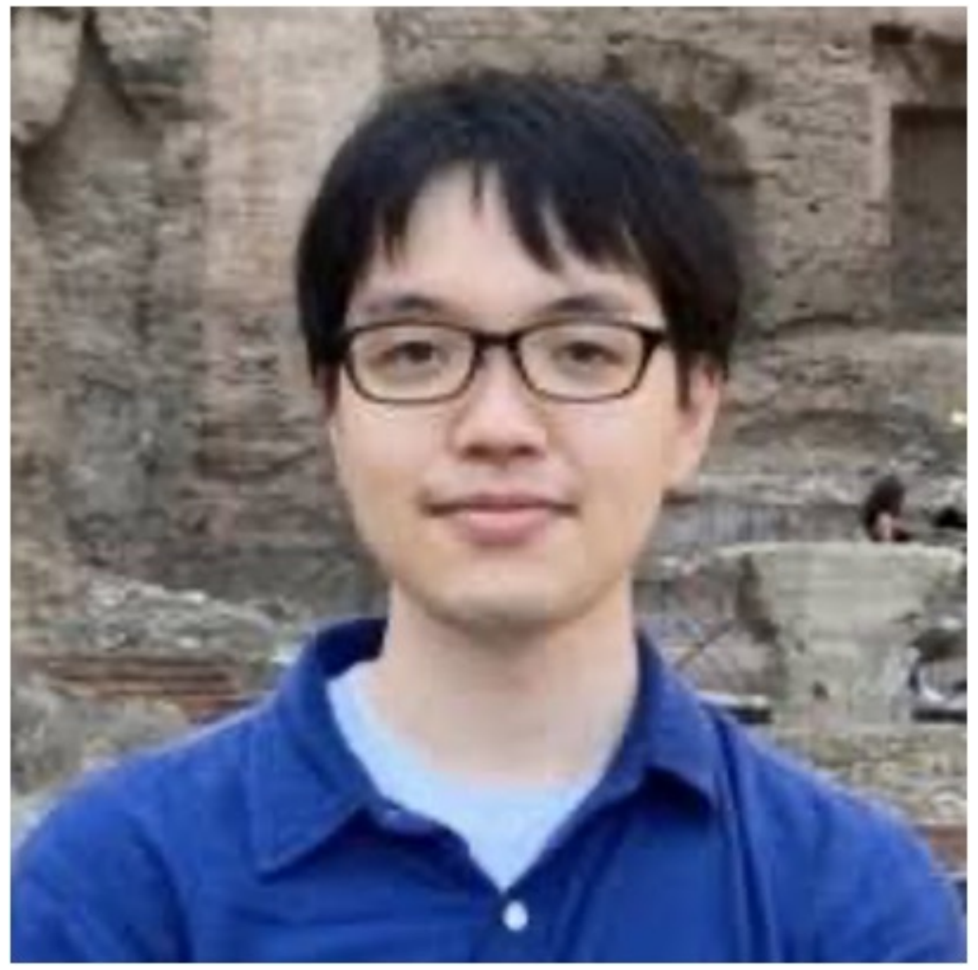
Han Bao

(Kyoto University → The Institute of Statistical Mathematics)

where most of the work has been done

This work was …

- Collaboration with great members!

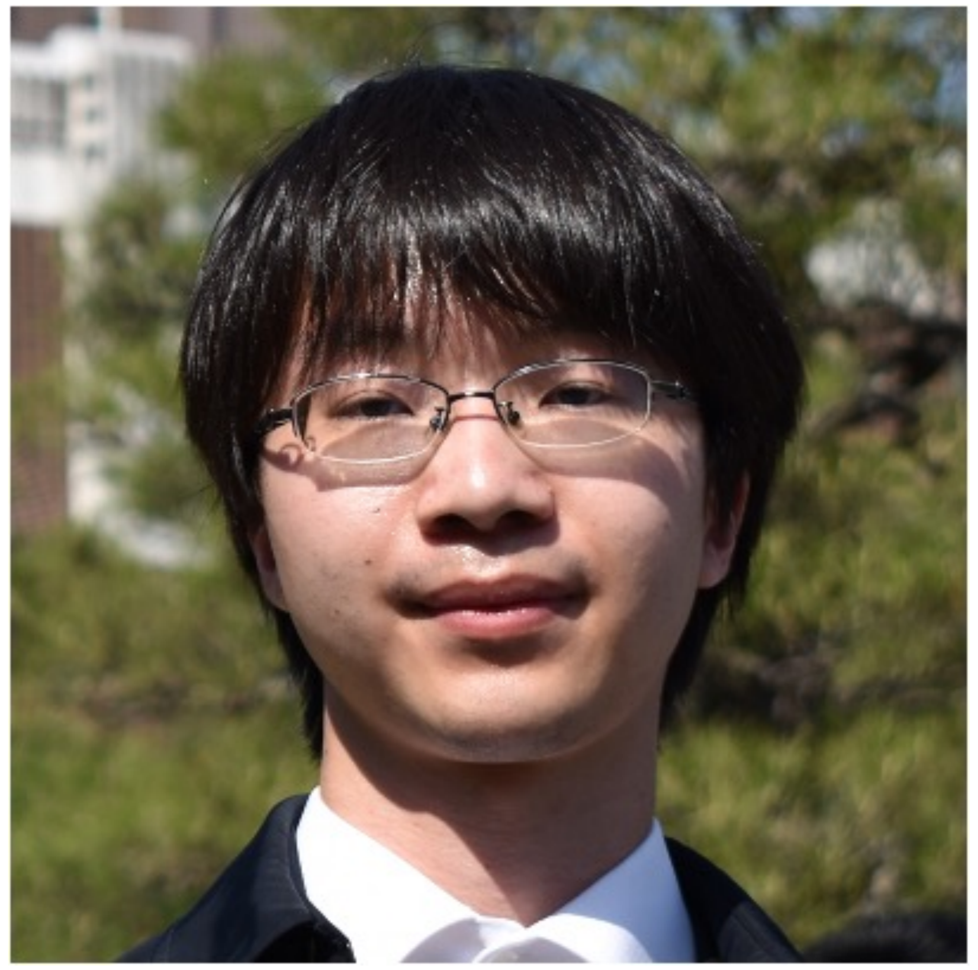


Satoki Ishikawa

Science Tokyo

Makoto Yamada

OIST



Yuki Takezawa

Kyoto University

Self-supervised Learning: What we can learn from **nonlinear dynamics** and **neuroscience**

Part I

learning dynamics, stability, adaptivity, ···

Part II

predictive coding, hippocampal model

Once upon a time …

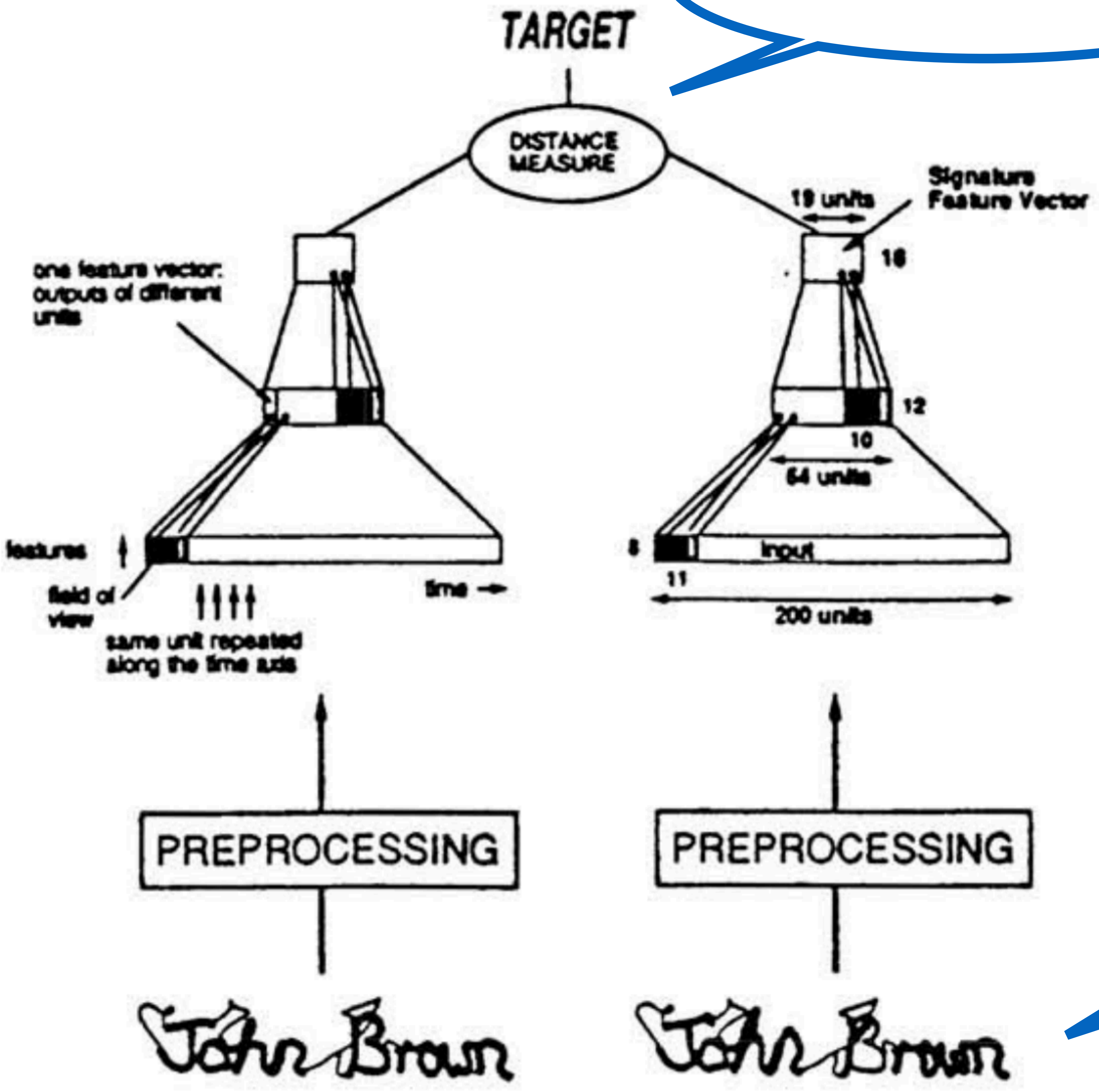
Published in Advances in Neural Information Processing Systems 6 (NIPS **1993**)

Signature Verification using a "Siamese" Time Delay Neural Network

Jane Bromley, Isabelle Guyon, Yann LeCun,
Eduard Säckinger and Roopak Shah
AT&T Bell Laboratories
Holmdel, NJ 07733
jbromley@big.att.com

Once upon a time …

Cosine distance



2005: supervised, metric learning

Learning a Similarity Metric Discriminatively, with Application to Face Verification

Sumit Chopra   Raia Hadsell   Yann LeCun

Courant Institute of Mathematical Sciences
New York University
New York, NY, USA

2010: unsupervised, density estimation

Noise-contrastive estimation: A new estimation principle for unnormalized statistical models

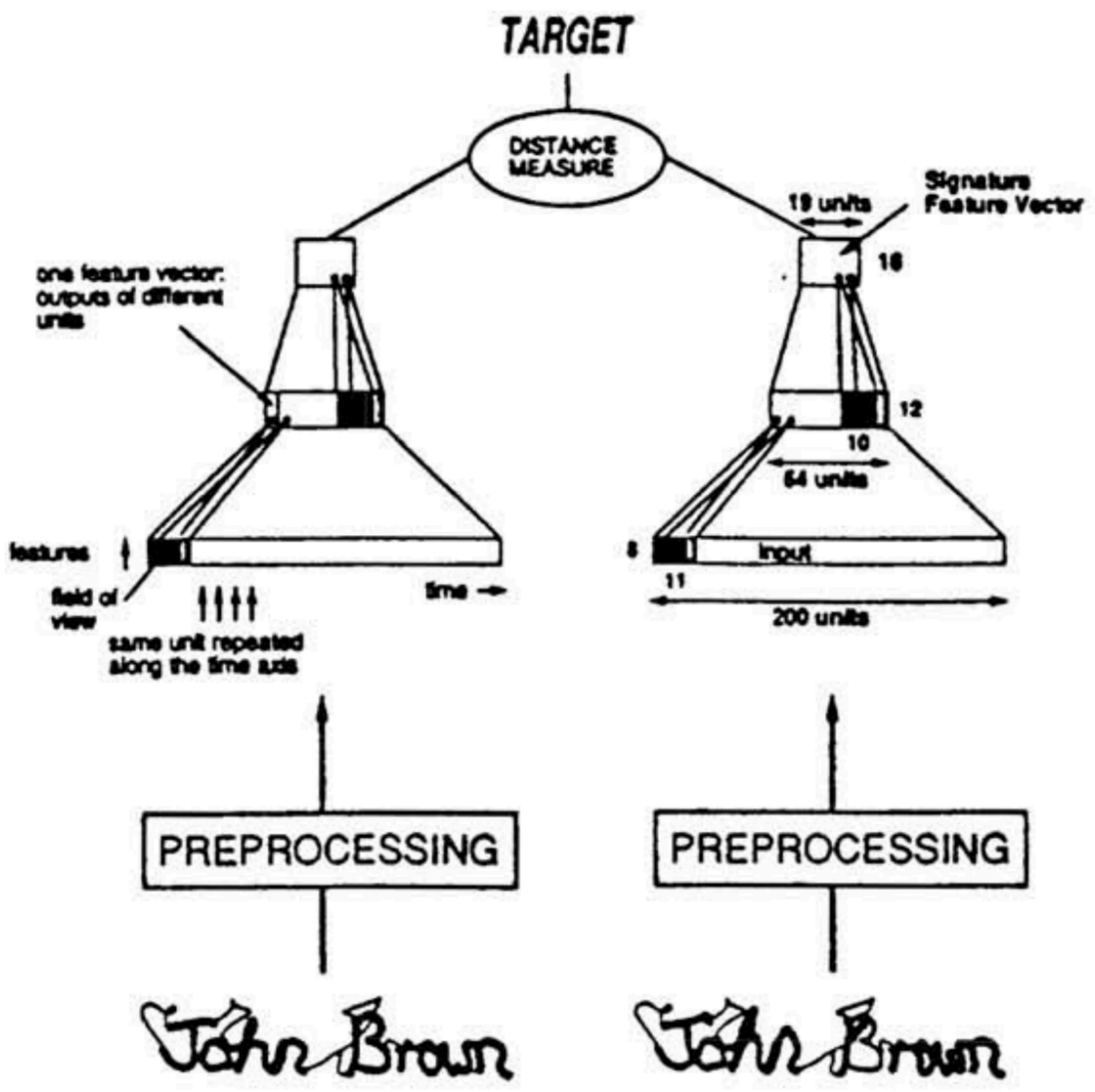
Michael Gutmann
Dept of Computer Science
and HIIT, University of Helsinki

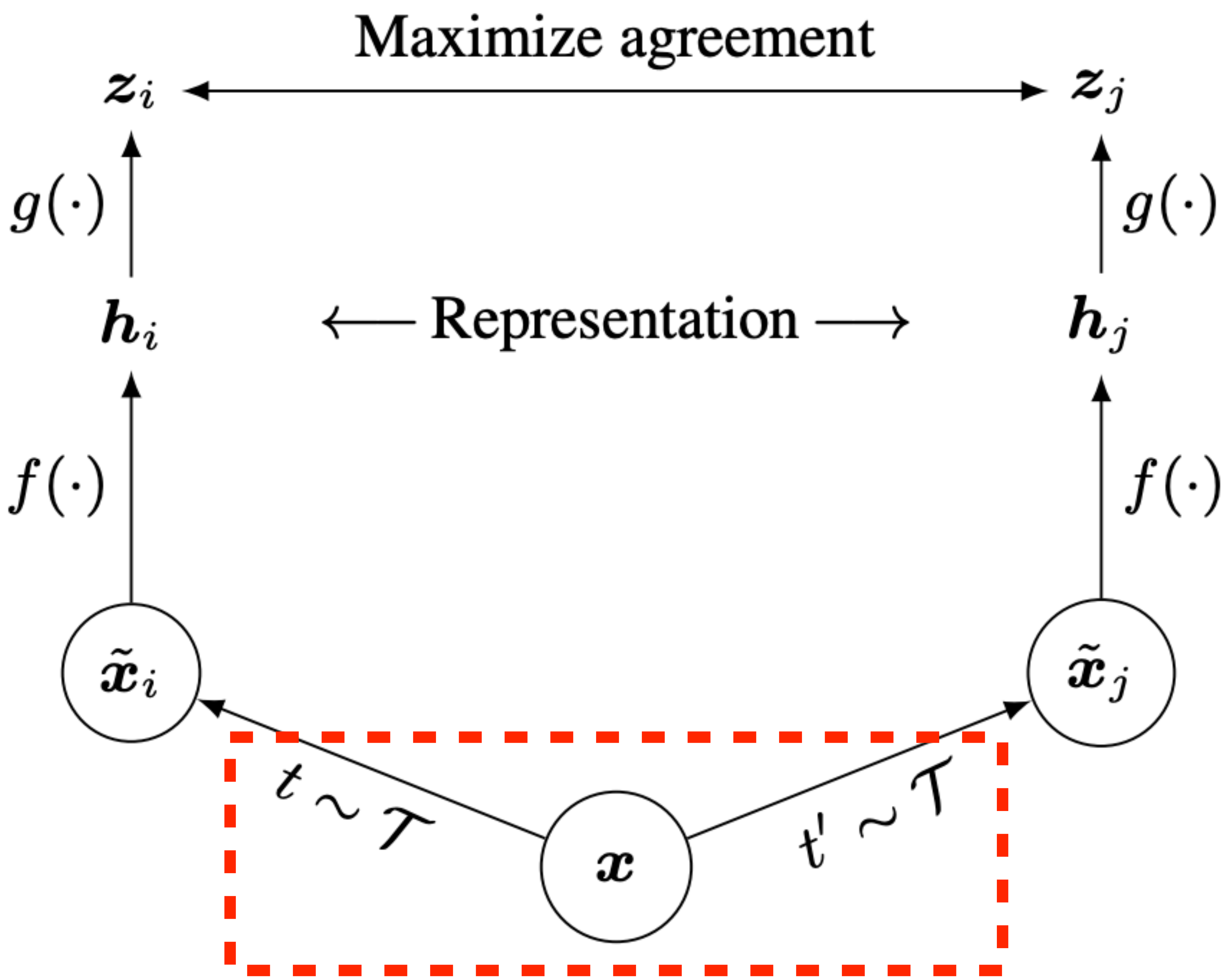
Aapo Hyvärinen
Dept of Mathematics & Statistics, Dept of Computer
Science and HIIT, University of Helsinki

From supervised to unsupervised

1993: supervised

2020: **un**supervised
SimCLR [Chen+ 20]



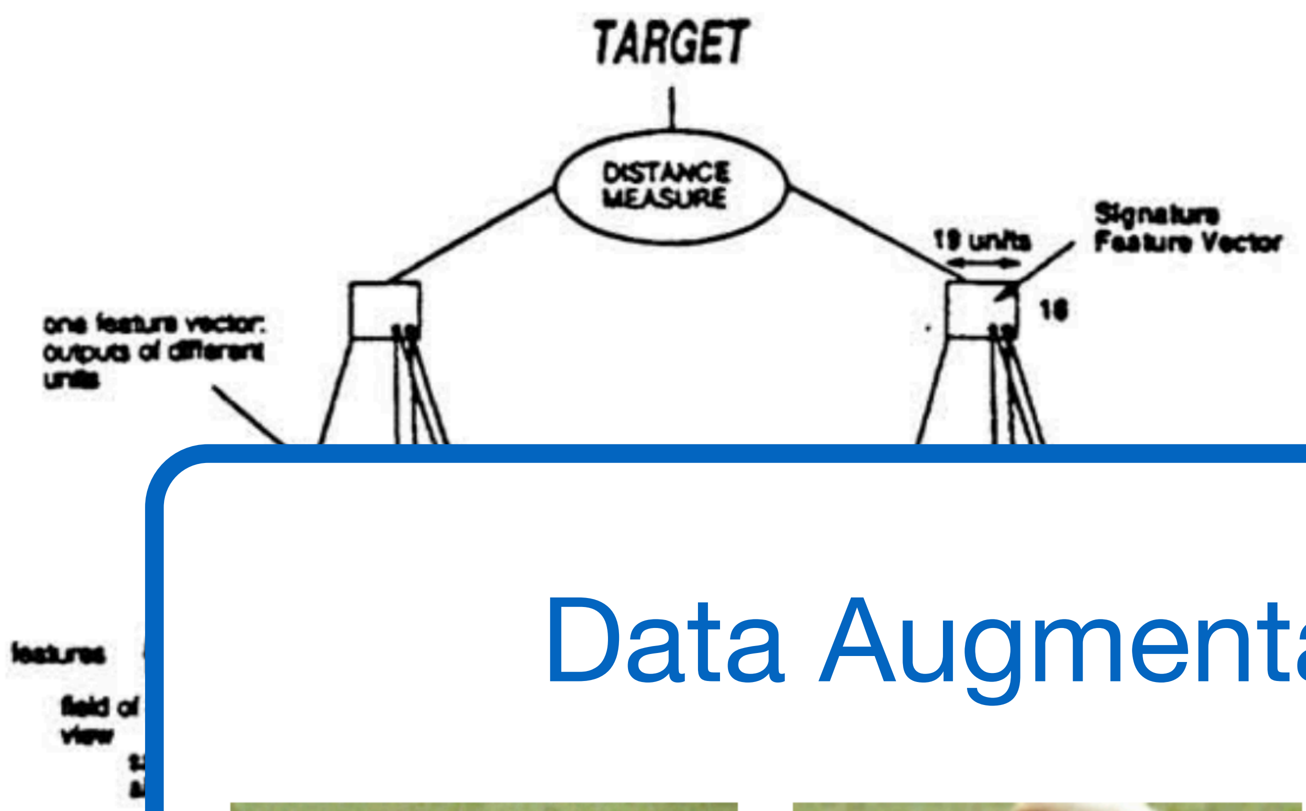


Chen et al. (2020) A Simple Framework for Contrastive Learning of Visual Representations

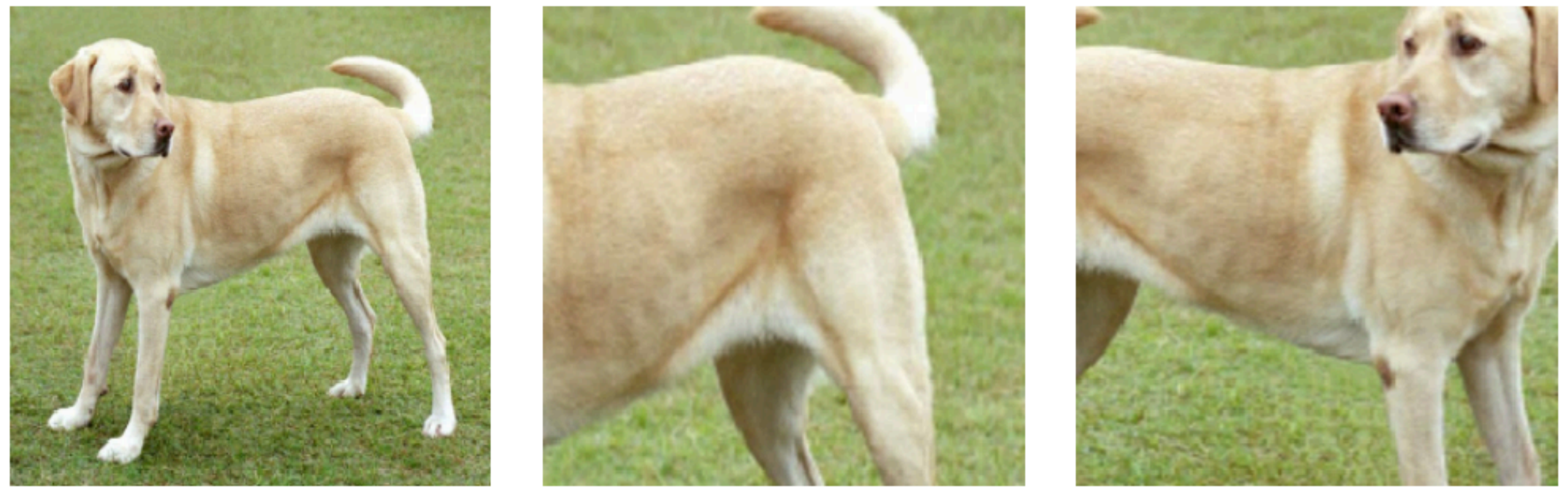
From supervised to unsupervised

1993: supervised → 2020: **un**supervised

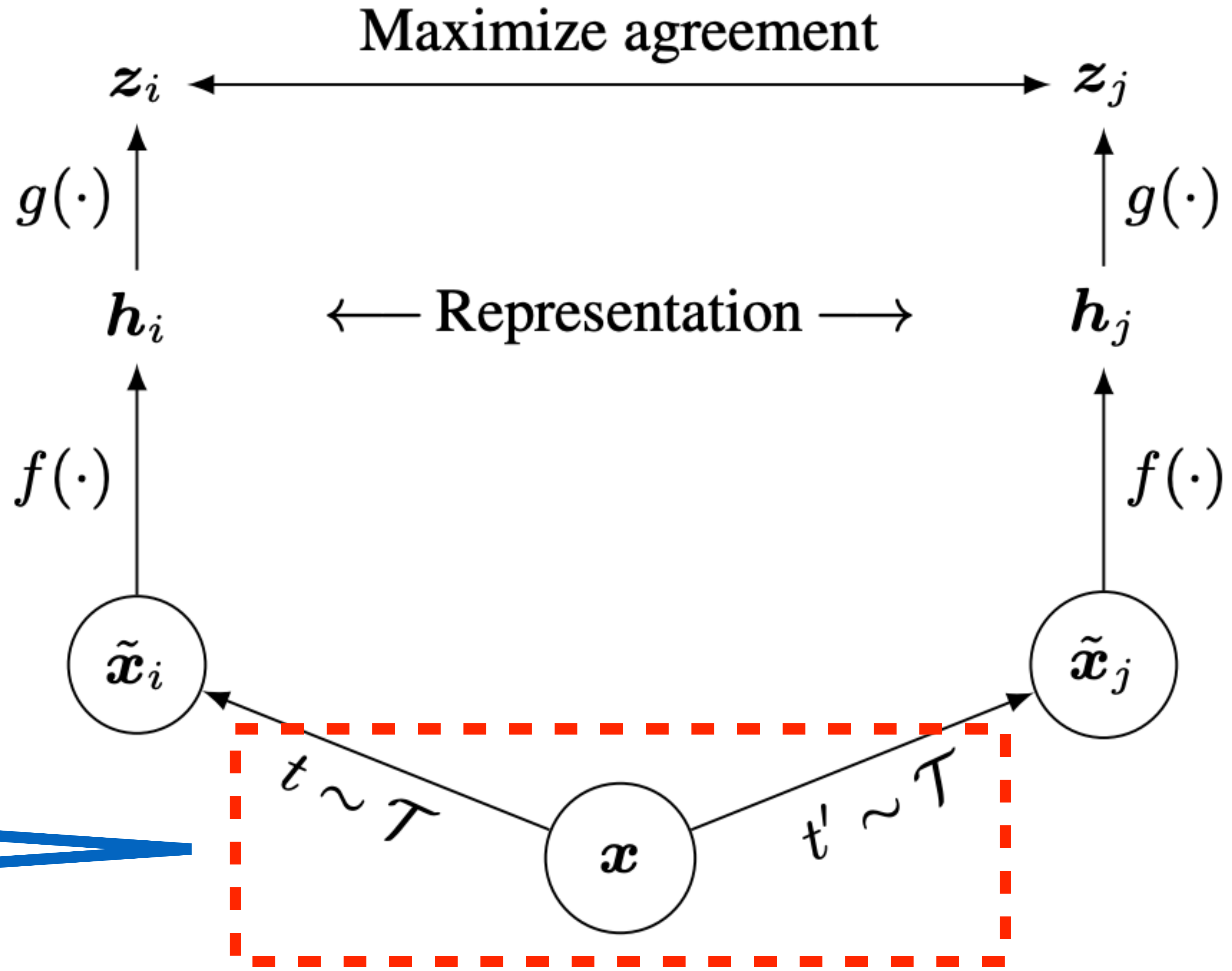
SimCLR [Chen+ 20]



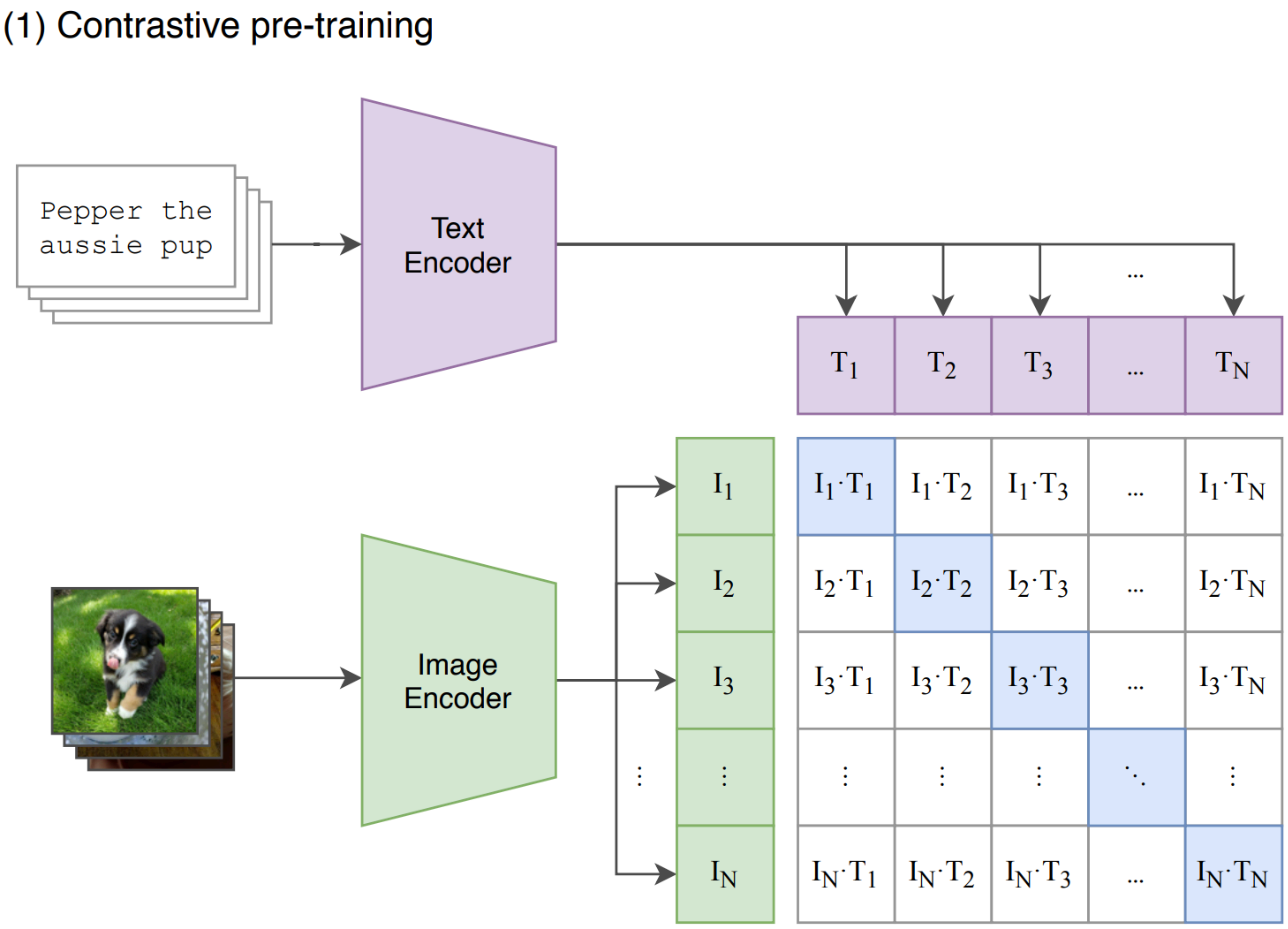
Data Augmentation



(a) Original (b) Crop and resize (c) Crop, resize (and flip)



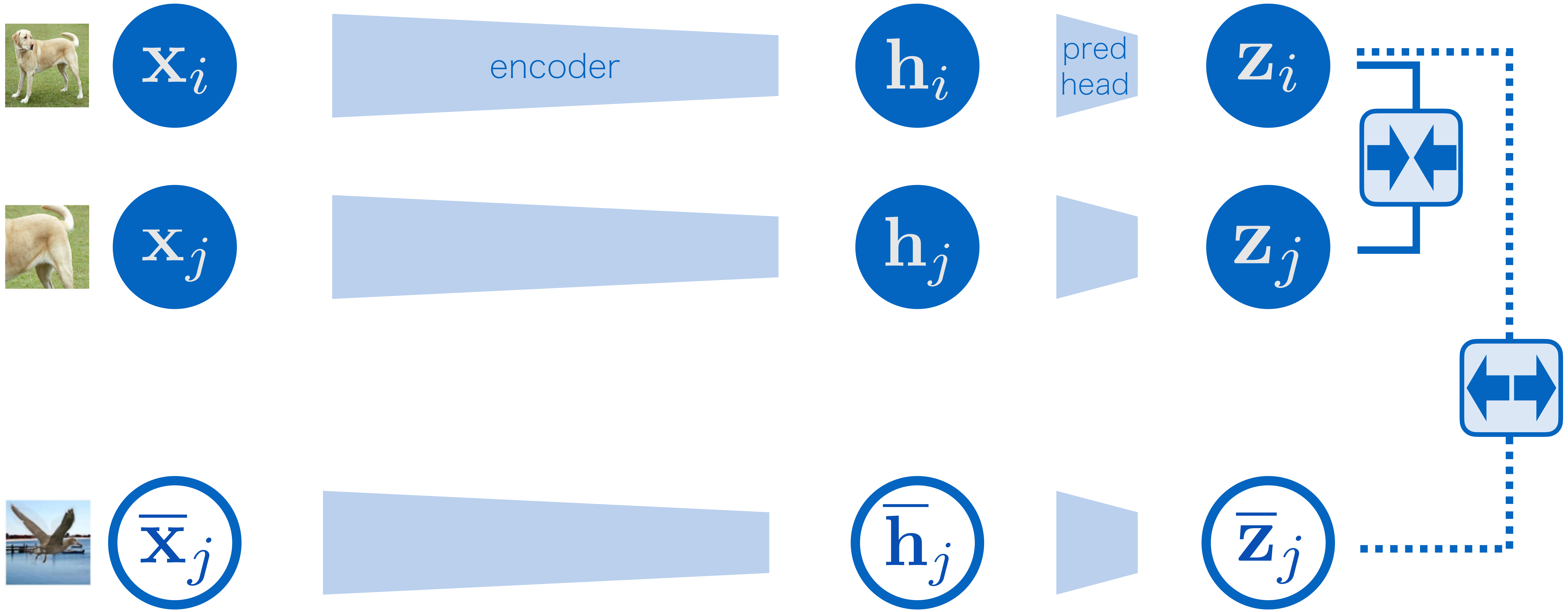
CLIP: multi-view representation learning



Radford et al. (2021) Learning Transferable Visual Models From Natural Language Supervision.

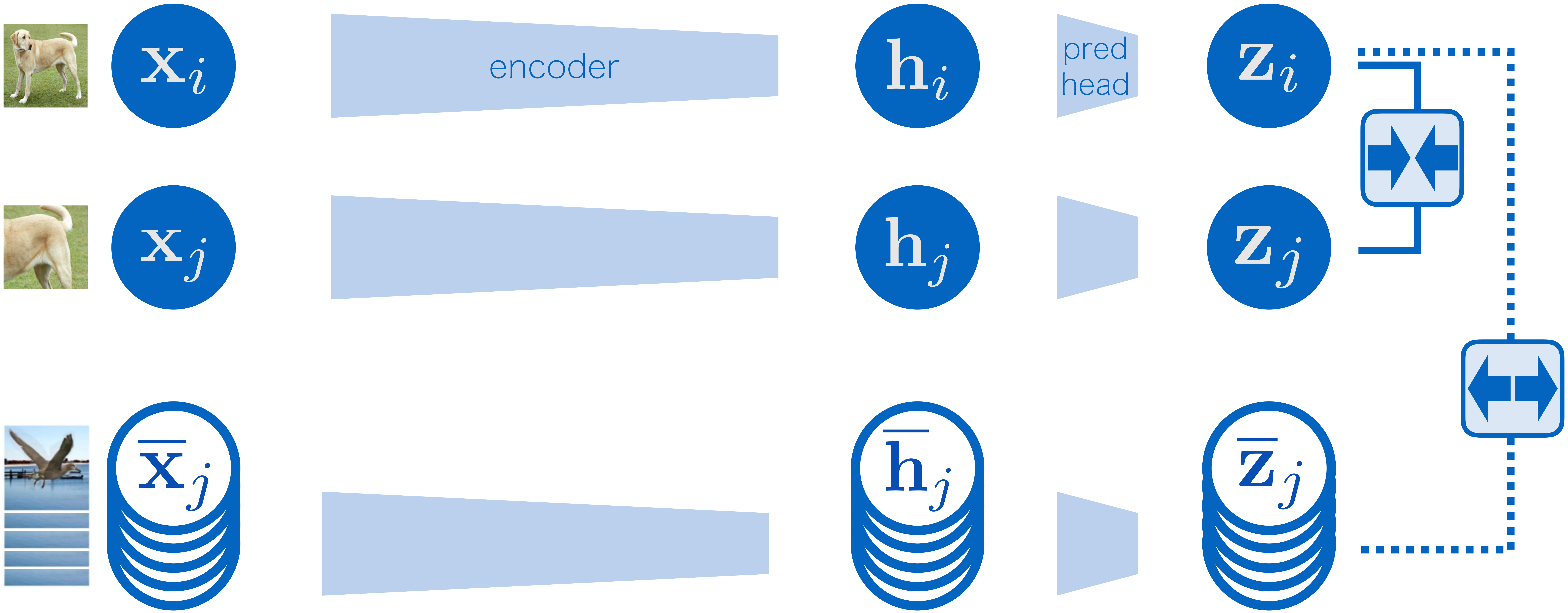
Massive negative sampling

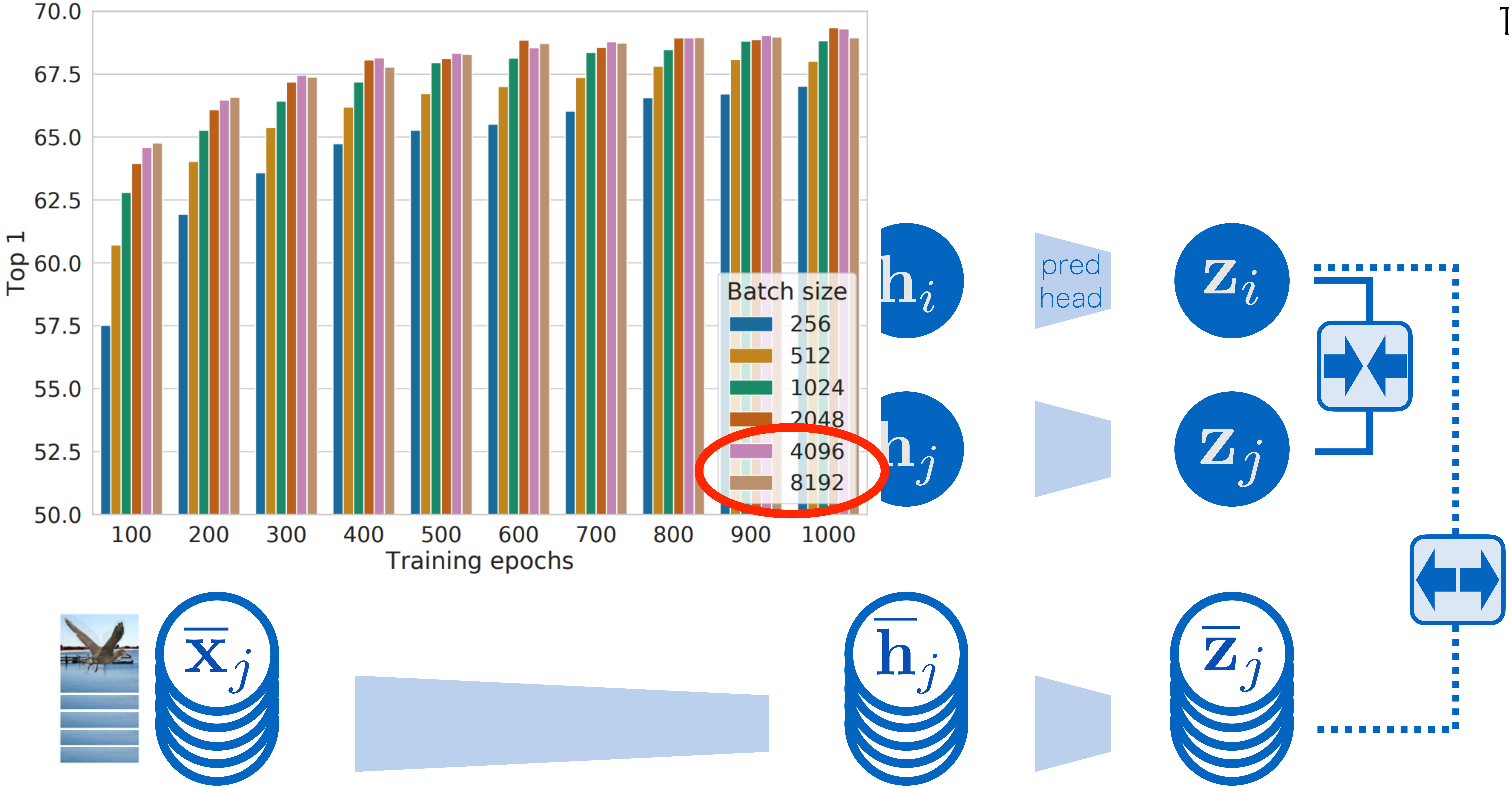
SimCLR [Chen+ 20]



Massive negative sampling

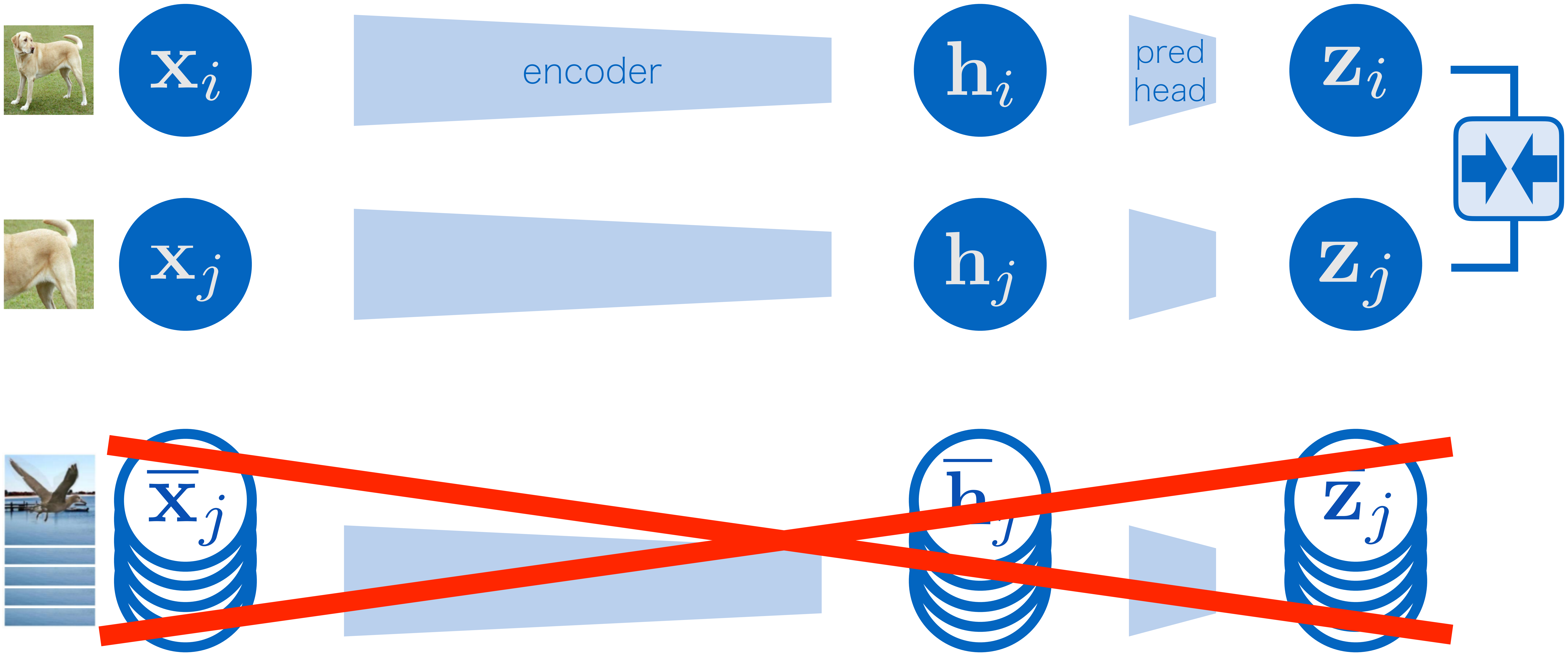
SimCLR [Chen+ 20]




70.0
67.5
65.0
62.5
60.0
57.5
55.0
52.5
50.0
Top 1
100
200
300
400
500
600
700
800
900
1000
Training epochs
Batch size
256
512
1024
2048
4096
8192
$\mathbf{h}_i$
pred
head
$\mathbf{z}_i$
$\mathbf{h}_j$
$\mathbf{z}_j$
$\overline{\mathbf{x}}_j$
$\overline{\mathbf{h}}_j$
$\overline{\mathbf{z}}_j$

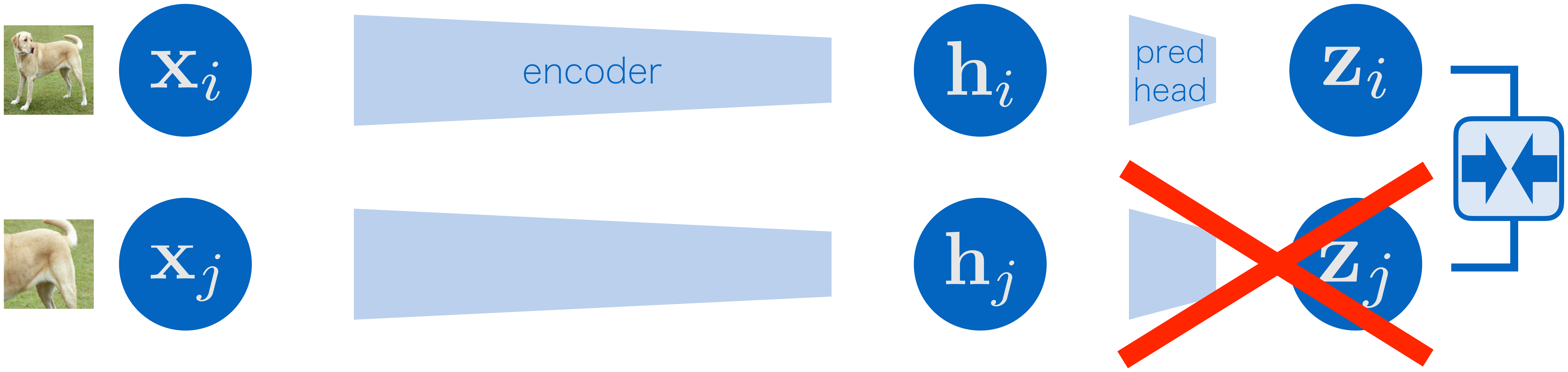
From contrastive to NON-contrastive

SimSiam [Chen-He 21]



From contrastive to NON-contrastive

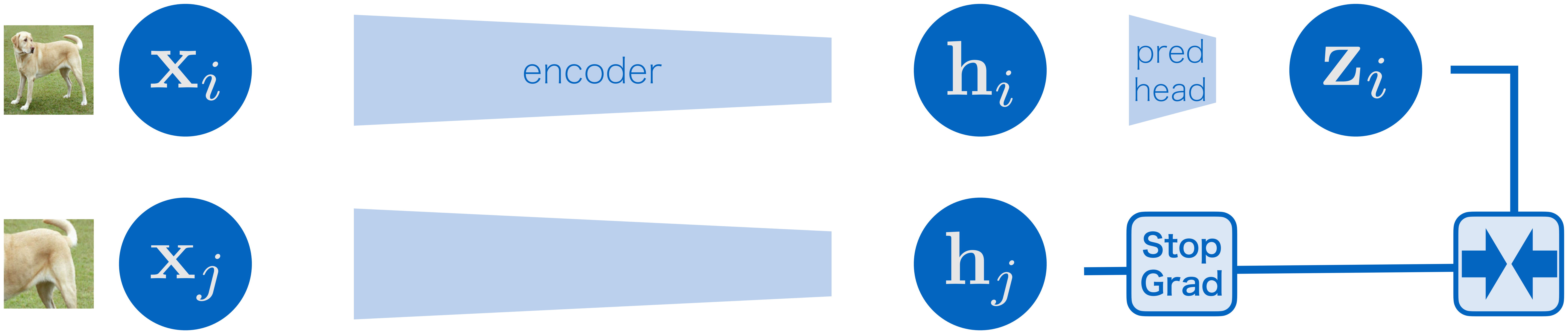
SimSiam [Chen-He 21]



Chen & He (2021) Exploring Simple Siamese Representation Learning.

From contrastive to NON-contrastive

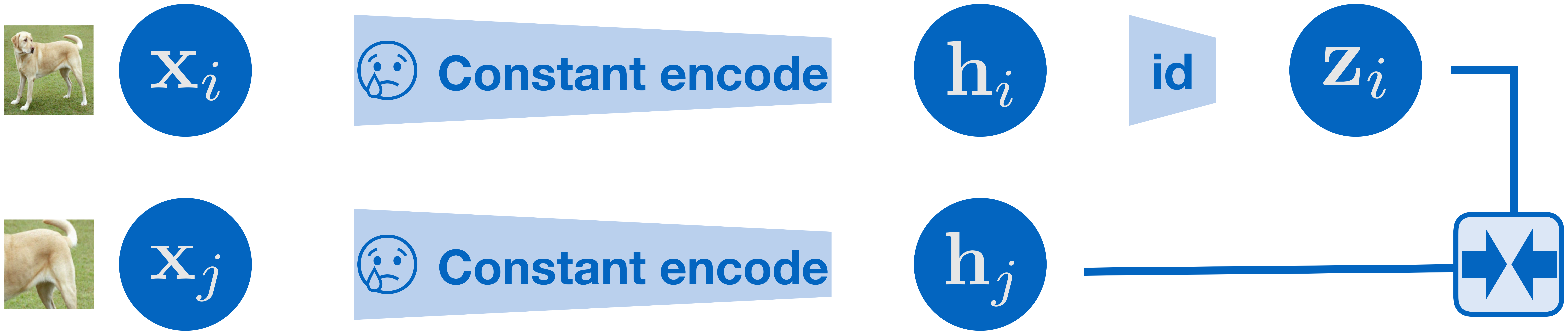
SimSiam [Chen-He 21]



- Data augmentation
- Prediction head (but at anchor side only!)
- Stop gradient

From contrastive to NON-contrastive

SimSiam [Chen-He 21]



- Data augmentation
- Prediction head (but at anchor side only!)
- Stop gradient

How to avoid *constant* encoder?
trivial

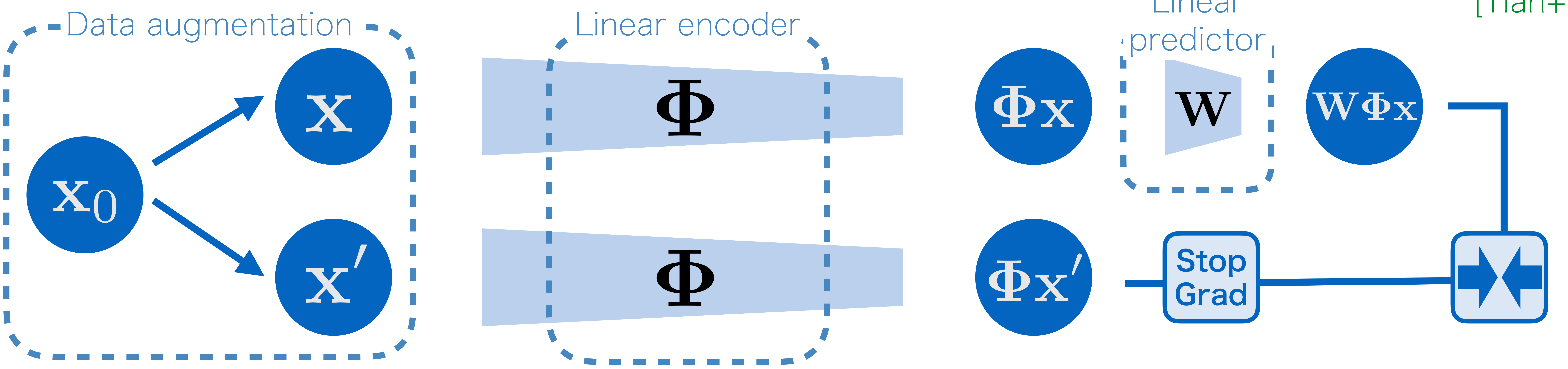
What we can learn from **nonlinear dynamics** and neuroscience

Bao, H. (2023)

Feature Normalization Prevents Collapse of Non-contrastive Learning Dynamics.

Theoretical model of SimSiam

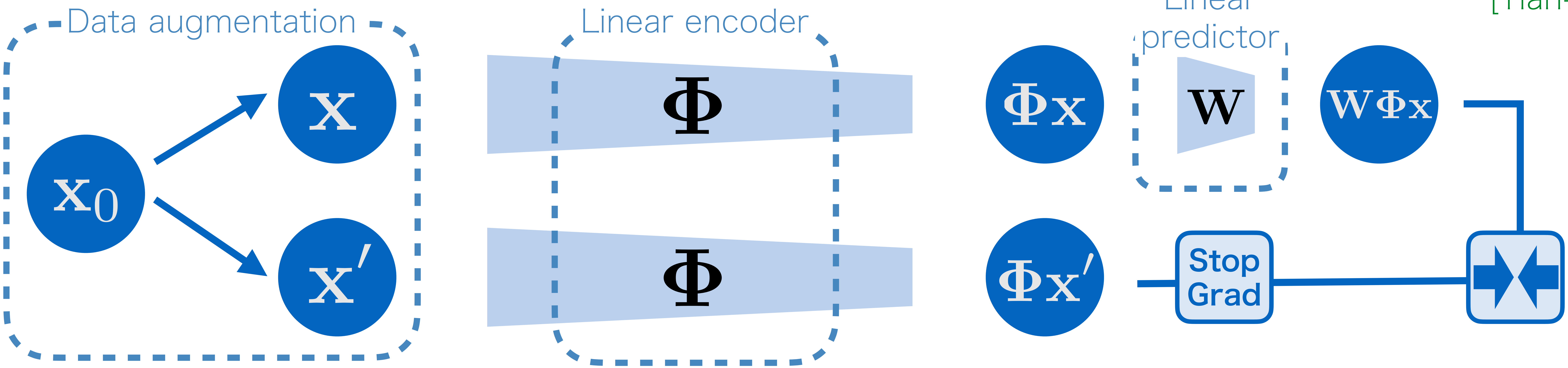
[Tian+ 21]



Tian et al. (2021) Understanding self-supervised learning dynamics without contrastive pairs.

Theoretical model of SimSiam

[Tian+ 21]



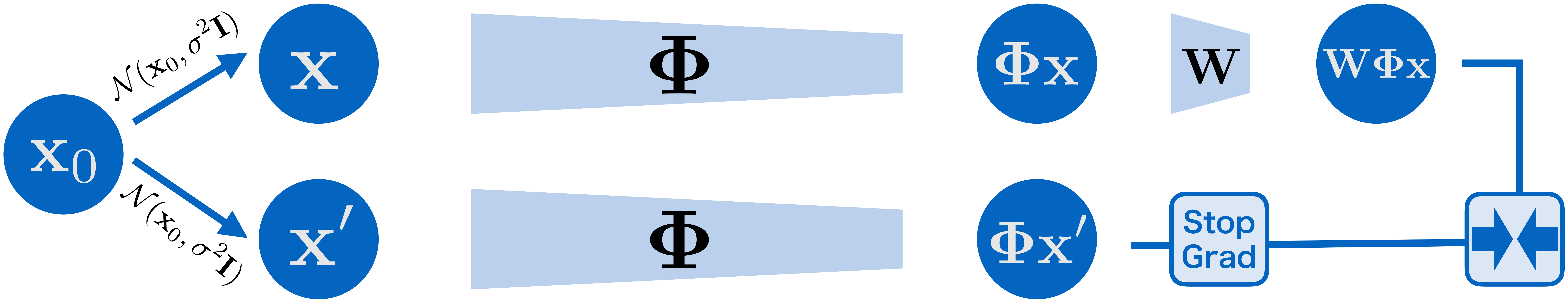
$$\mathbf{x}_0 \sim \mathcal{N}(0, \mathbf{I})$$

$$\mathbf{x}, \mathbf{x}' \sim \mathcal{N}(\mathbf{x}_0, \sigma^2 \mathbf{I})$$

Strength of data aug

Theoretical model of SimSiam

[Tian+ 21]



$$\mathcal{L}(\mathbf{\Phi}, \mathbf{W}) = \frac{1}{2}\mathbb{E}\|\mathbf{W\Phi x} - \mathrm{StopGrad}(\mathbf{\Phi x}')\|_2^2$$

Learning dynamics

[Tian+ 21]

$$\mathcal{L}(\mathbf{\Phi}, \mathbf{W}) = \frac{1}{2}\mathbb{E}\|\mathbf{W}\mathbf{\Phi}\mathbf{x} - \mathrm{StopGrad}(\mathbf{\Phi}\mathbf{x}')\|_2^2$$

Discrete time

gradient descent

weight decay

$$\begin{cases} \mathbf{\Phi}(t+1) = \mathbf{\Phi}(t) - \eta(\nabla_{\mathbf{\Phi}}\mathcal{L} + \rho\mathbf{\Phi}(t)) \\ \mathbf{W}(t+1) = \mathbf{W}(t) - \eta(\nabla_{\mathbf{W}}\mathcal{L} + \rho\mathbf{W}(t)) \end{cases}$$

$\eta \to 0$

Continuous time

gradient flow

$$\begin{cases} \dot{\mathbf{\Phi}} = -\nabla_{\mathbf{\Phi}}\mathcal{L} - \rho\mathbf{\Phi} \\ \dot{\mathbf{W}} = -\nabla_{\mathbf{W}}\mathcal{L} - \rho\mathbf{W} \end{cases}$$

Analysis overview: Eigenvalue decomposition

[Tian+ 21]

Matrix dynamics: not easy to deal with 😢

$$\begin{cases} \dot{\mathbf{\Phi}} = -\nabla_{\mathbf{\Phi}}\mathcal{L} - \rho\mathbf{\Phi} \\ \dot{\mathbf{W}} = -\nabla_{\mathbf{W}}\mathcal{L} - \rho\mathbf{W} \end{cases} \xrightarrow{\mathbf{\Phi}\mathbf{\Phi}^\top \equiv \mathbf{F}} \begin{cases} \dot{\mathbf{F}} = -2(1+\sigma^2)\mathbf{W}^2\mathbf{F} + 2\mathbf{W}\mathbf{F} - 2\rho\mathbf{F} \\ \dot{\mathbf{W}} = -(1+\sigma^2)\mathbf{W}\mathbf{F} + \mathbf{F} - \rho\mathbf{W} \end{cases}$$

$$\downarrow \begin{cases} s: & j\text{-th eigval of } \mathbf{F} \\ p: & j\text{-th eigval of } \mathbf{W} \end{cases}$$

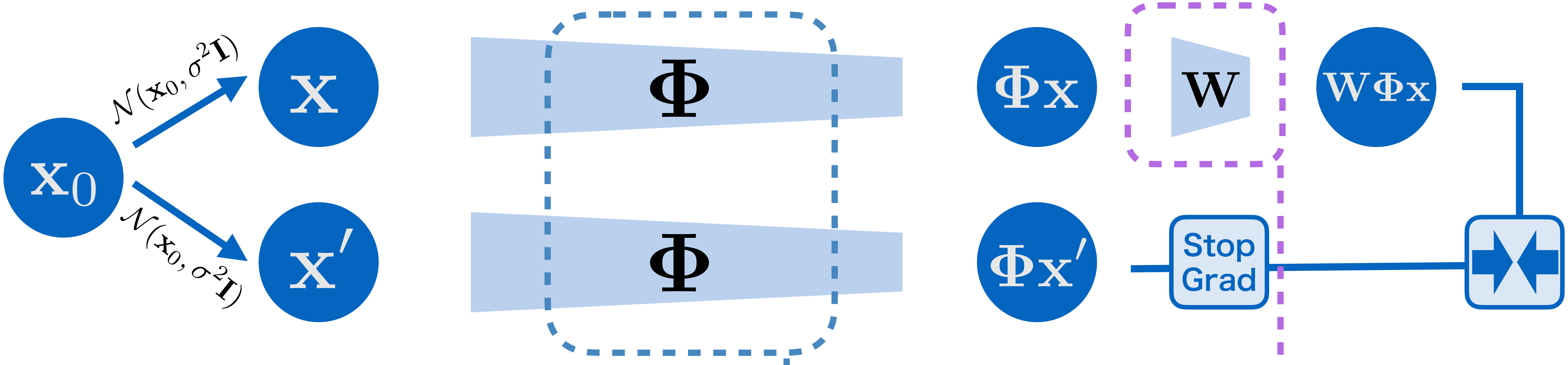
Scalar dynamics:

enabled by eigendecomposition 😁

$$\begin{cases} \dot{s} = -2(1+\sigma^2)p^2 s + 2ps - 2\rho s \\ \dot{p} = -(1+\sigma^2)ps + s - \rho p \end{cases}$$

Recap: gradient flow → decoupled dynamics

[Tian+ 21]



Eigvals of matrices evolves as follows:

$$\begin{cases} \dot{s} = -2(1+\sigma^2)p^2 s + 2ps - 2\rho s \\ \dot{p} = -(1+\sigma^2)ps + s - \rho p \end{cases}$$

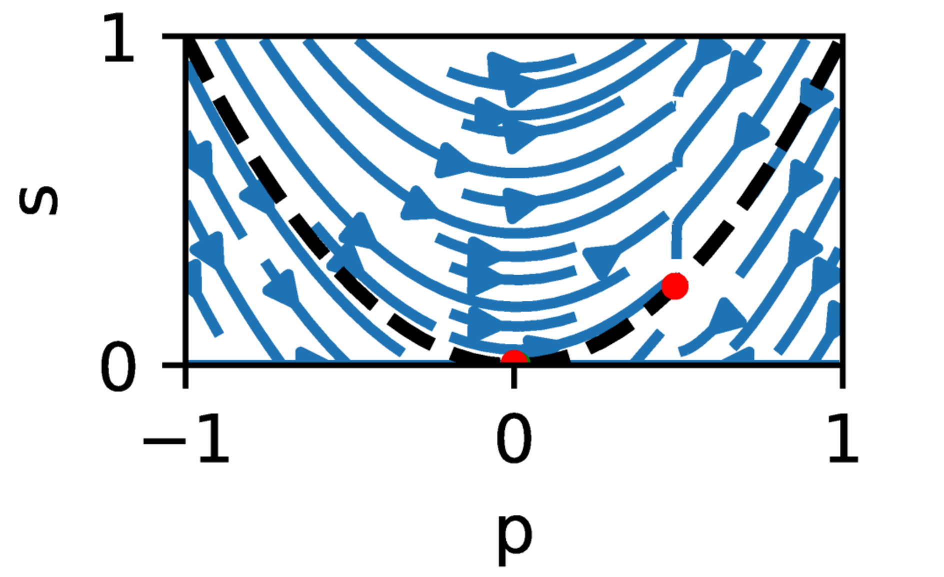
Goal: How to avoid trivial solution?

[Tian+ 21]

Simultaneous ODE

$$\begin{cases} \dot{s} = -2(1+\sigma^2)p^2 s + 2ps - 2\rho s \\ \dot{p} = -(1+\sigma^2)ps + s - \rho p \end{cases}$$

Adiabatic elimination



Eigval ODE of predictor

$$\dot{p} = p^2\{1 - (1+\sigma^2)p\} - \rho p$$

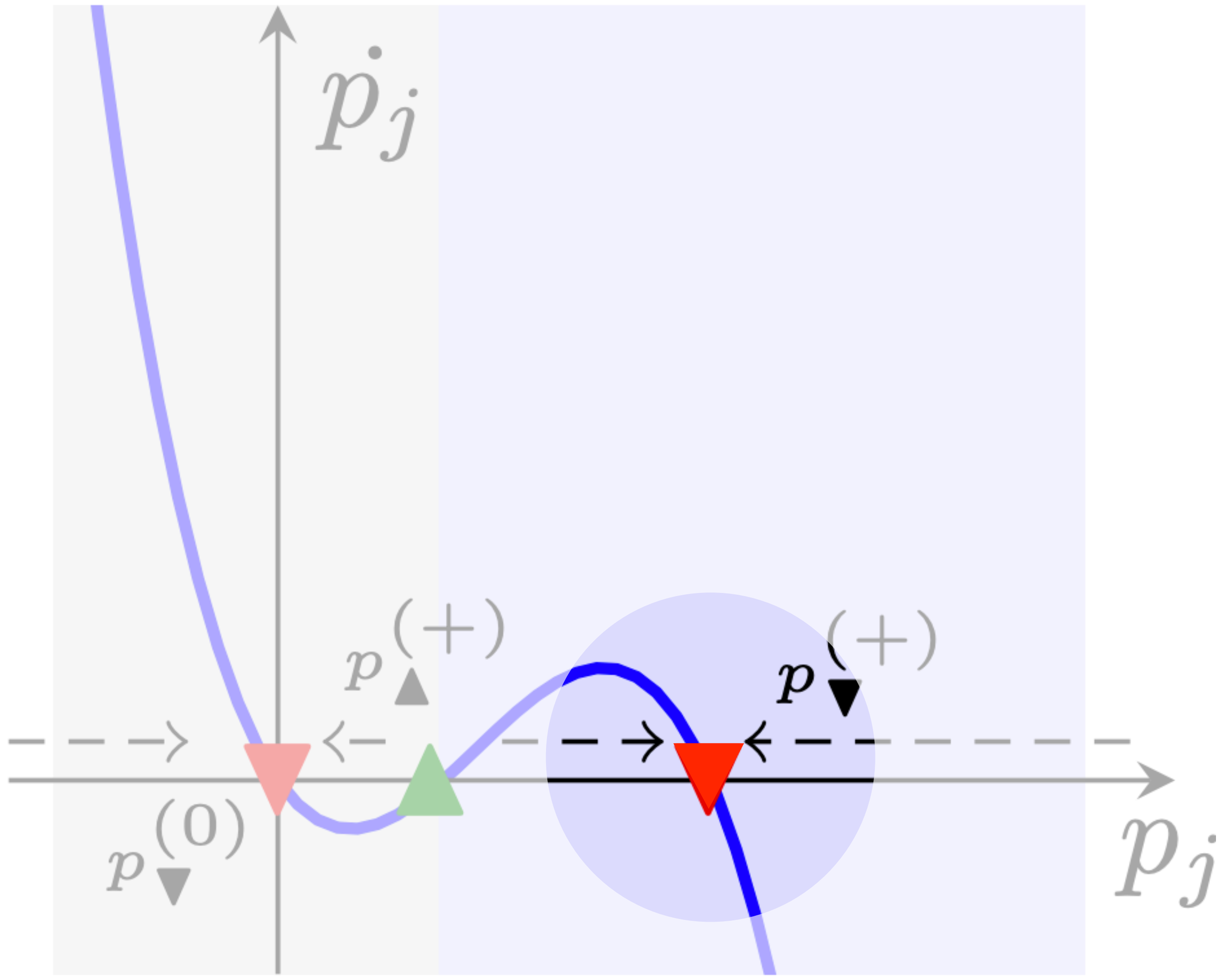
- Two params: σ^2 (data aug) & ρ (weight decay)
- **Q.** How to avoid constant predictor?

↕

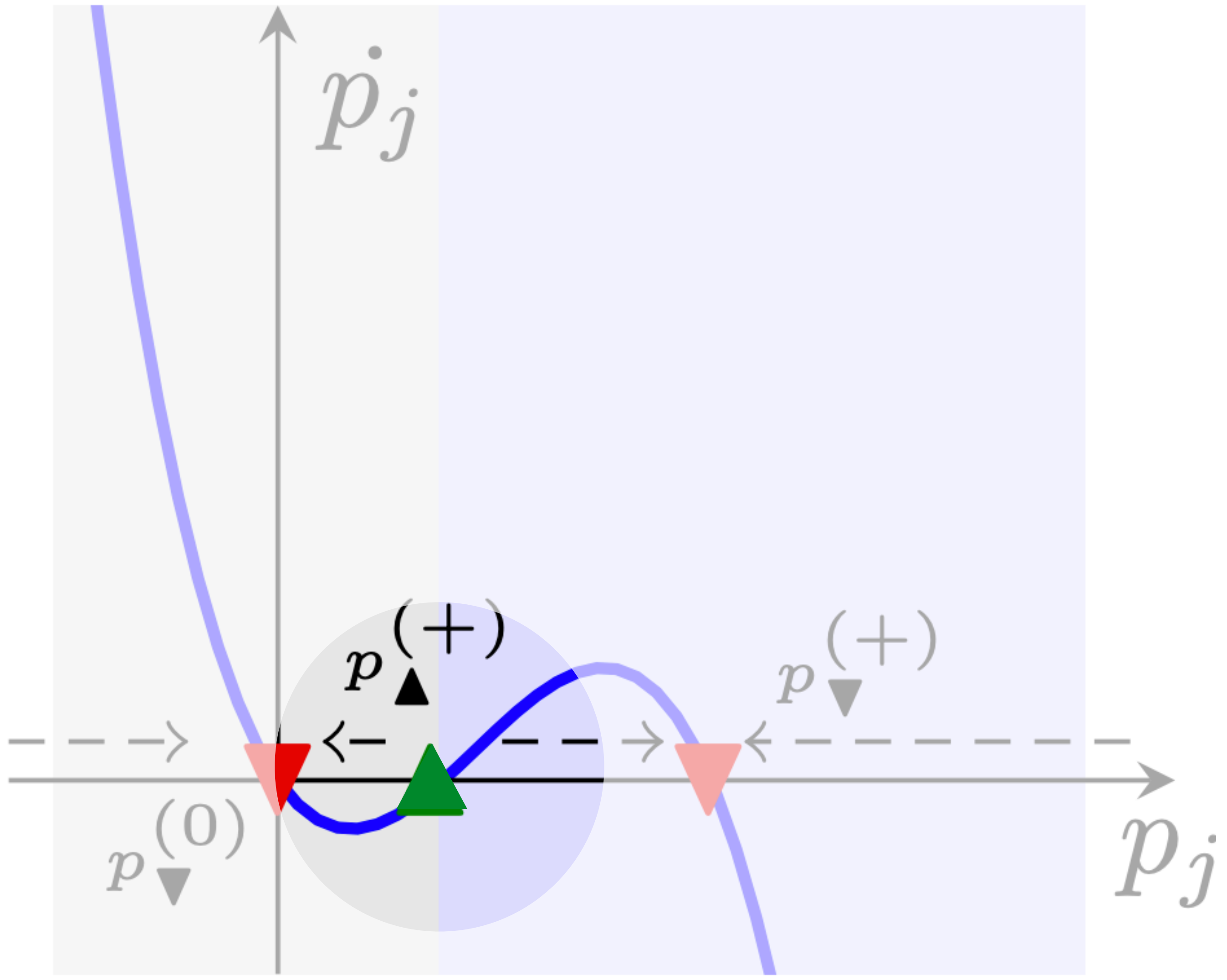
- **Q.** How to avoid $p = 0$?

Quick pre-requisite

- Stability analysis of ODE $\dot{p} = f(p)$
- $\dot{p} = 0$ is equilibrium (but can be unstable)
- If $f(p) < 0$: **stable**
- If $f(p) > 0$: **unstable**



because $f(p)$ head for the equilibrium locally



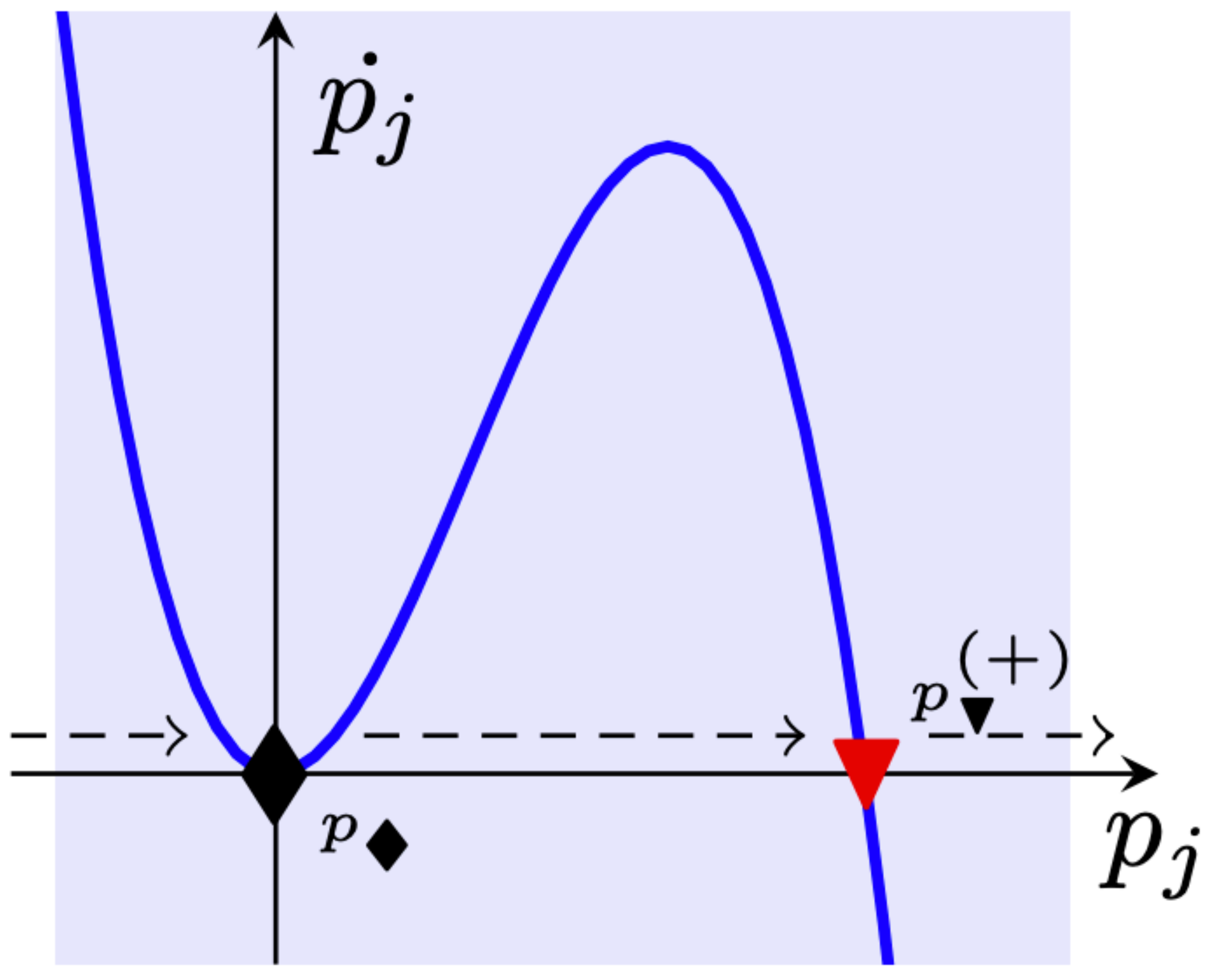
Bifurcation: too strong weight decay collapses 😢

[Tian+ 21]

Eigval ODE of projector

$$\dot{p} = p^2\{1 - (1+\sigma^2)p\} - \rho p$$

Case (a): $\rho = 0$



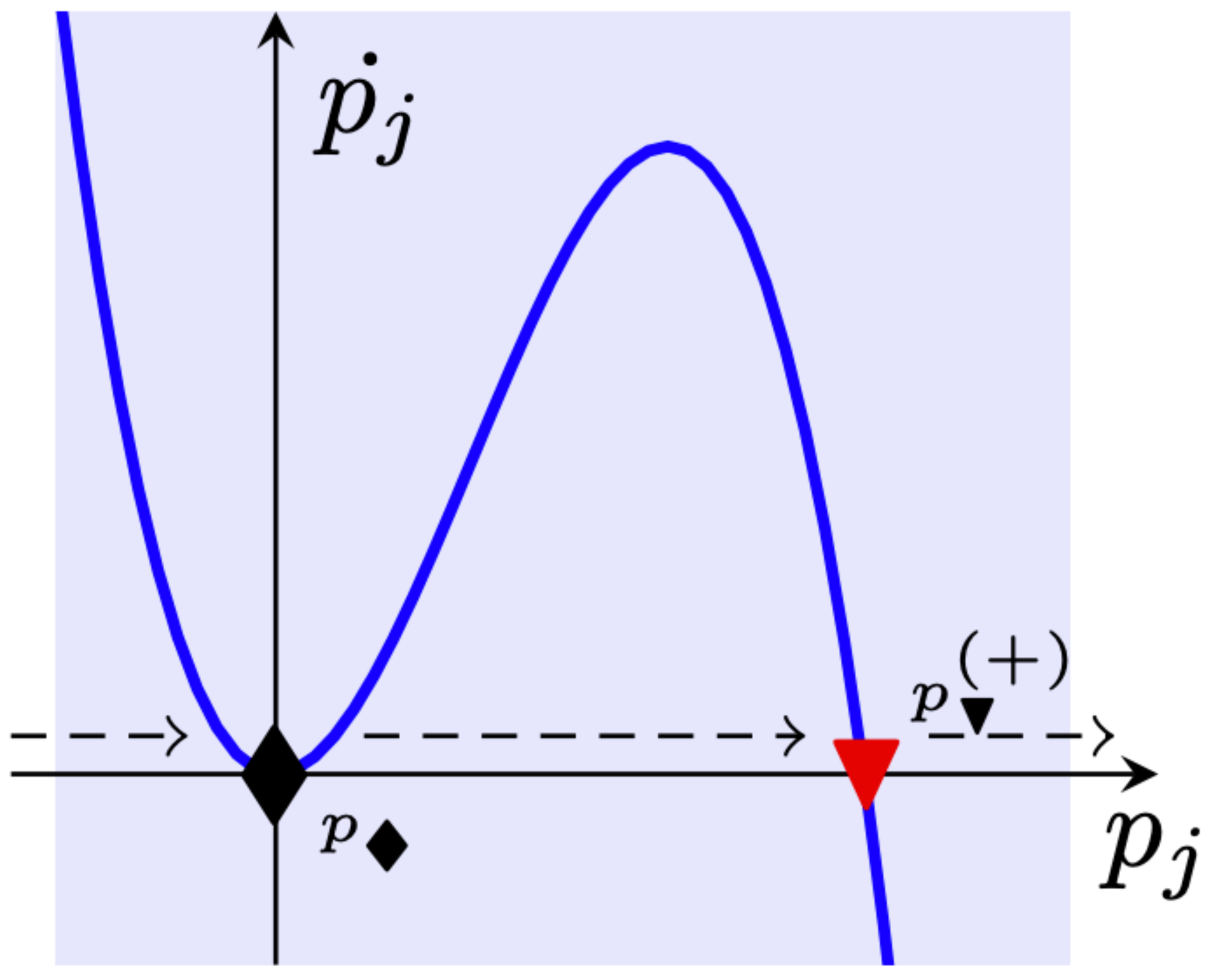
Bifurcation: too strong weight decay collapses 😢

[Tian+ 21]

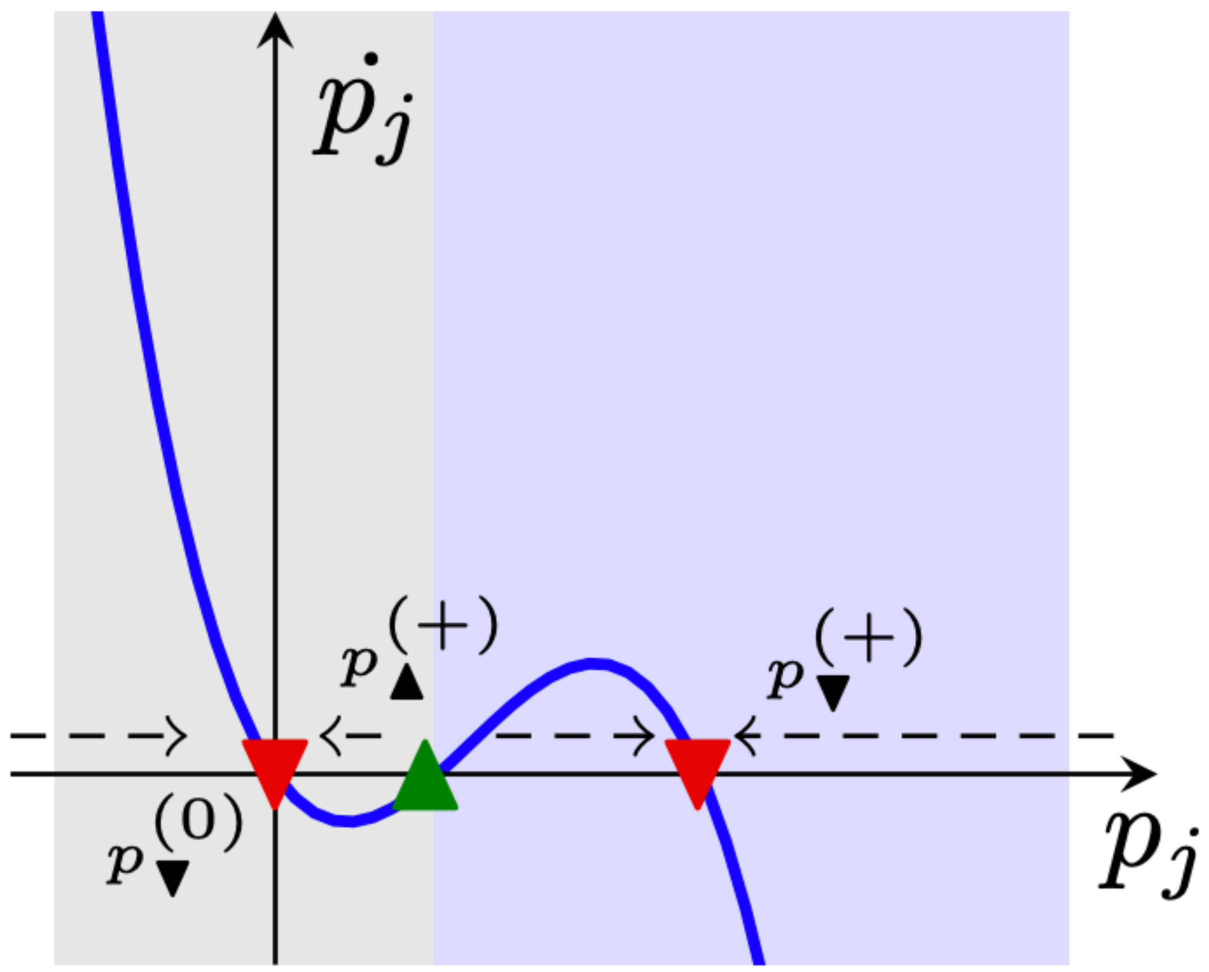
Eigval ODE of projector

$$\dot{p} = p^2\{1 - (1 + \sigma^2)p\} - \rho p$$

Case (a): $\rho = 0$



Case (b): $\rho < \frac{1}{4(1+\sigma^2)}$

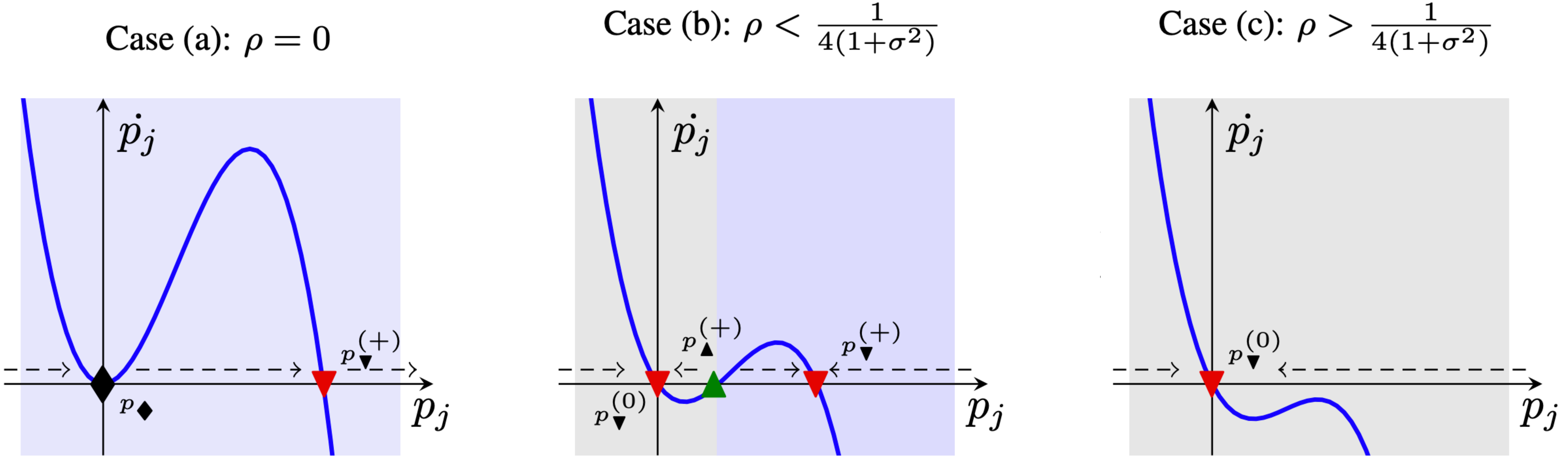


Bifurcation: too strong weight decay collapses 😢

[Tian+ 21]

Eigval ODE of projector

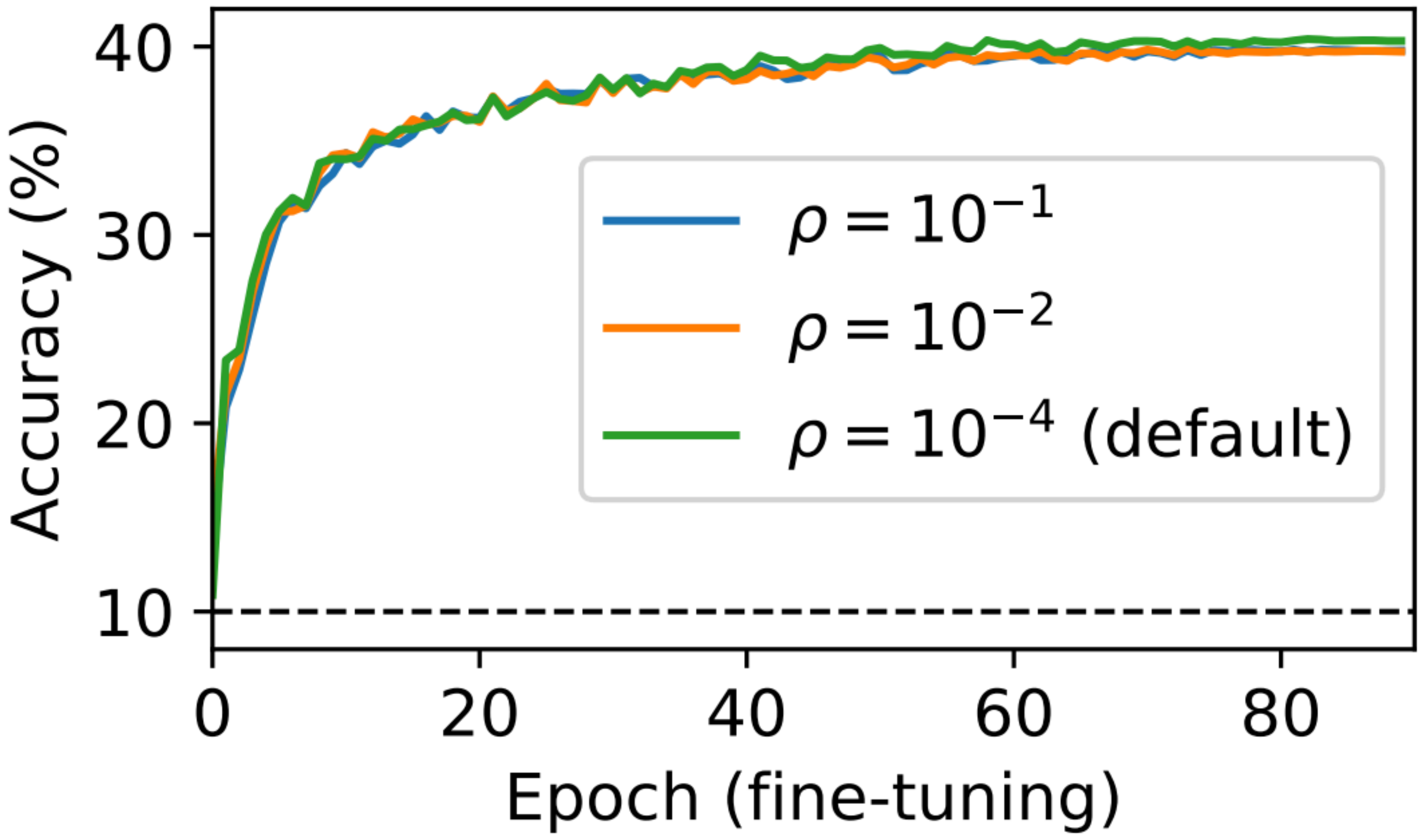
$$\dot{p} = p^2\{1 - (1+\sigma^2)p\} - \rho p$$



strong weight decay:
trivial solution $p = 0$ only

But is this really happening?

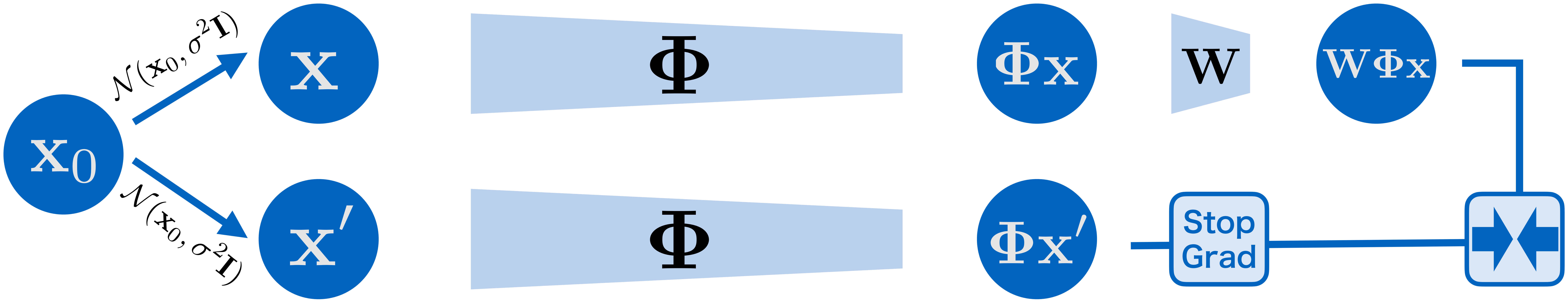
- Pilot study: SimSiam on CIFAR-10
  - evaluation: linear probing accuracy
- [Chen-He 21] Let's use small enough WD!
- [Bao 23] Intensifying WD keeps working
- 🔑 small learning rate
  (to enter gradient flow regime)



larger WD ρ still works

= accuracy does not breaks down

What differs from practice?



[Tian+ 21]

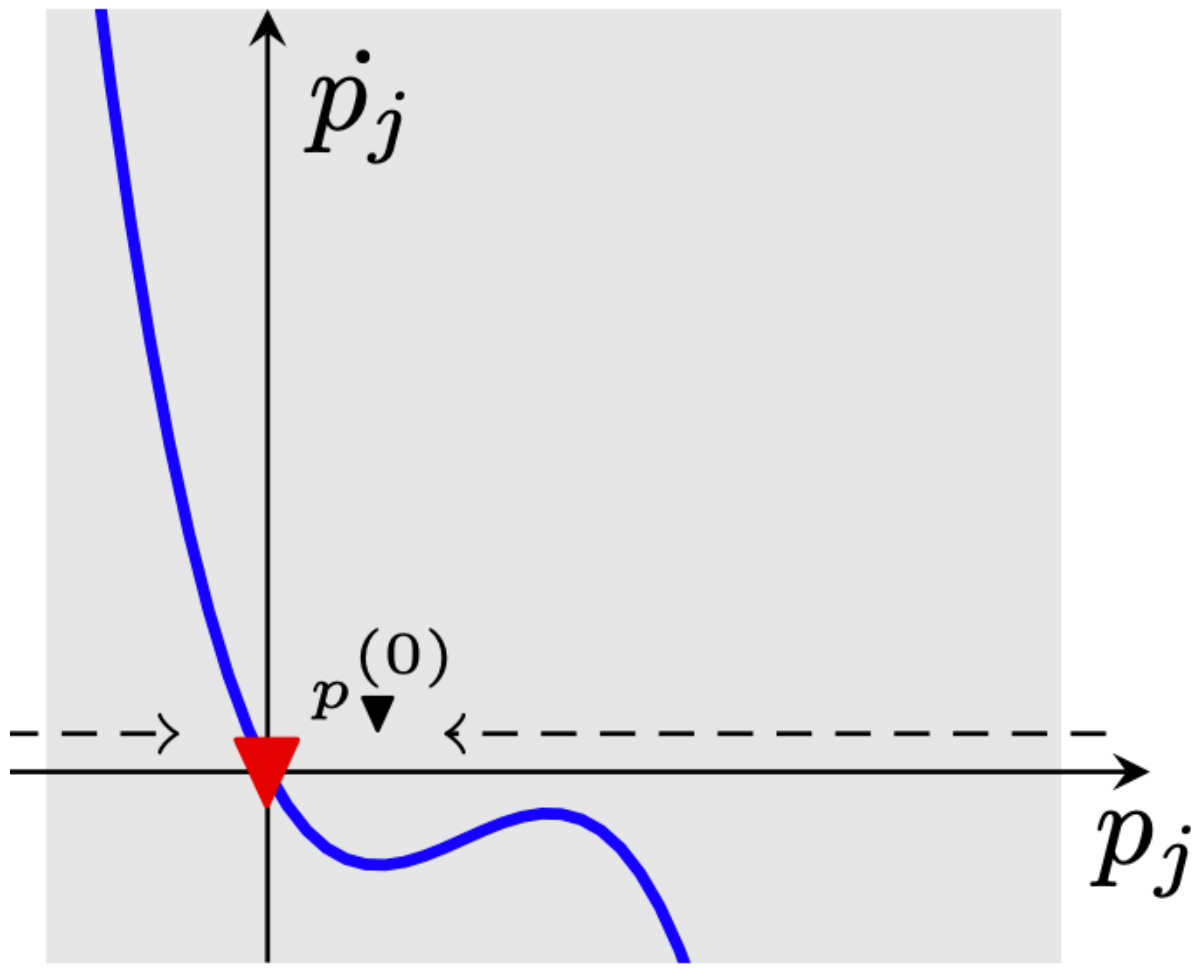
$$\mathcal{L}(\mathbf{\Phi}, \mathbf{W}) = \frac{1}{2}\mathbb{E}\|\mathbf{W\Phi x} - \mathrm{StopGrad}(\mathbf{\Phi x}')\|_2^2$$

SimSiam impl

$$\mathcal{L}(\mathbf{\Phi}, \mathbf{W}) = \mathbb{E}\left[-\frac{\langle \mathbf{W\Phi x}, \mathrm{StopGrad}(\mathbf{\Phi x}')\rangle}{\|\mathbf{W\Phi x}\|\|\mathrm{StopGrad}(\mathbf{\Phi x}')\|}\right]$$

Cosine loss may prevent collapse

- If collapsing ($p = 0$), predictor goes to zero $\mathbf{W} = \mathbf{O}$, blowing up cosine loss



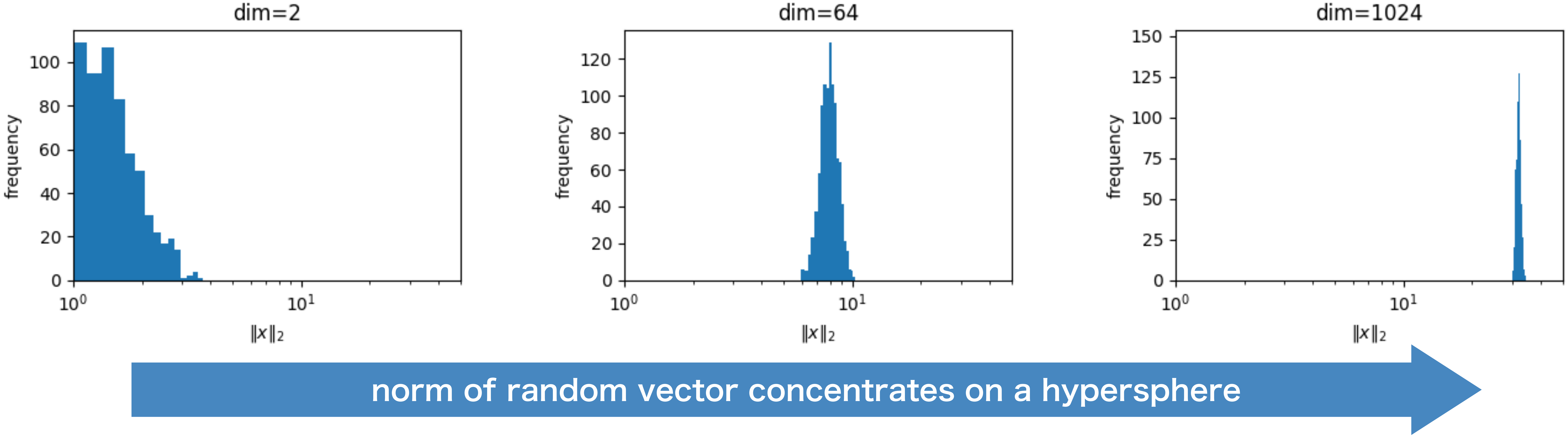
$$\mathcal{L}(\boldsymbol{\Phi}, \mathbf{W}) = \mathbb{E}\left[-\frac{\langle \mathbf{W}\boldsymbol{\Phi}\mathbf{x}, \mathrm{StopGrad}(\boldsymbol{\Phi}\mathbf{x}')\rangle}{\|\mathbf{W}\boldsymbol{\Phi}\mathbf{x}\| \|\mathrm{StopGrad}(\boldsymbol{\Phi}\mathbf{x}')\|}\right]$$

🤔 **What does the cosine-loss dynamics look like?**

Challenges of cosine loss: normalization

$$\mathcal{L}(\mathbf{\Phi}, \mathbf{W}) = \mathbb{E}\left[-\frac{\langle \mathbf{W\Phi x}, \mathrm{StopGrad}(\mathbf{\Phi x}')\rangle}{\|\mathbf{W\Phi x}\|\|\mathrm{StopGrad}(\mathbf{\Phi x}')\|}\right]$$

- Taking derivative wrt normalizer makes gradient complicated
- Solution: **high-dimensional limit**



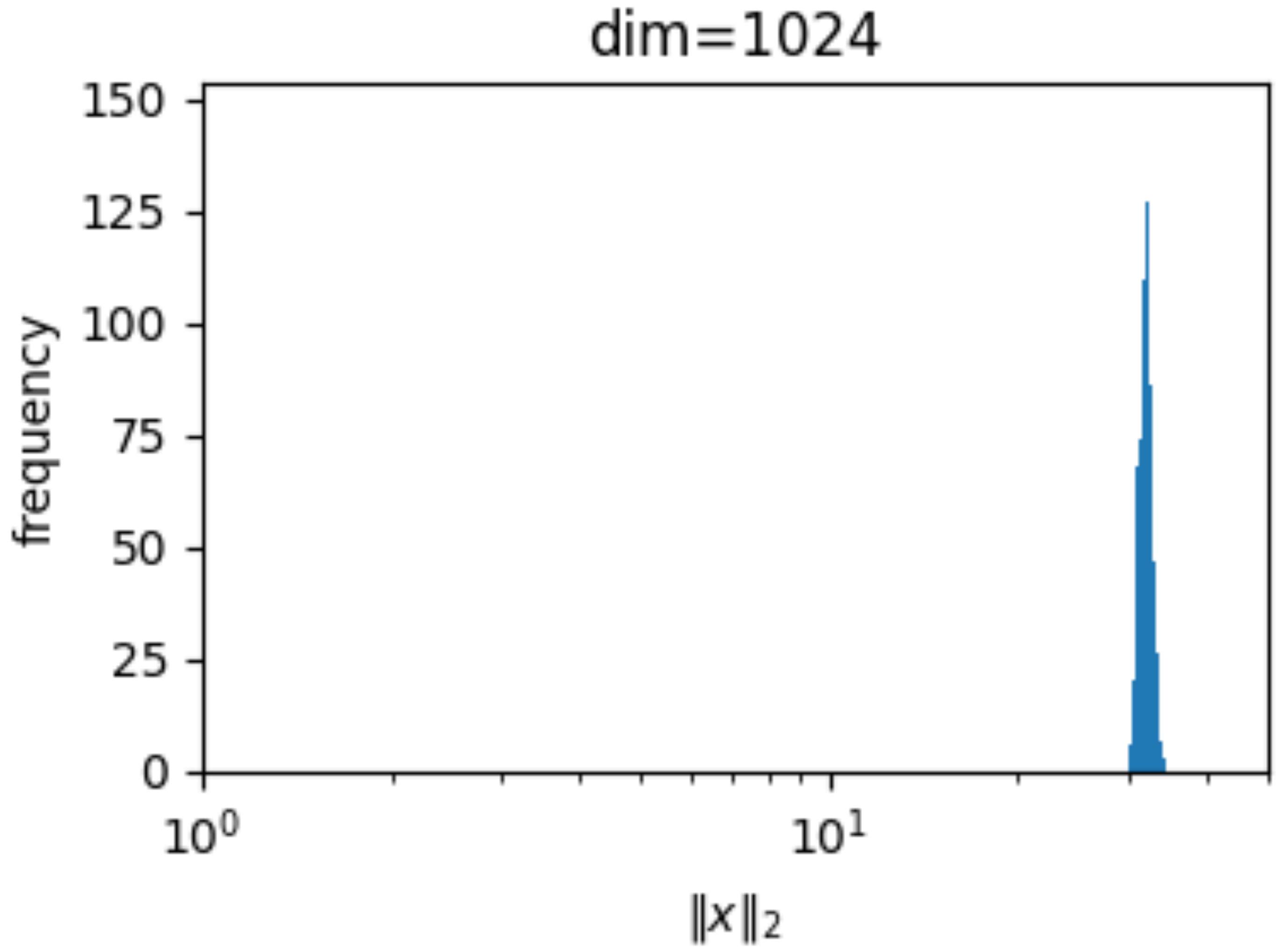
Challenges of cosine loss: normalization

$$\mathcal{L}(\mathbf{\Phi}, \mathbf{W}) = \mathbb{E}\left[-\frac{\langle \mathbf{W}\mathbf{\Phi}\mathbf{x}, \mathrm{StopGrad}(\mathbf{\Phi}\mathbf{x}')\rangle}{\|\mathbf{W}\mathbf{\Phi}\mathbf{x}\| \|\mathrm{StopGrad}(\mathbf{\Phi}\mathbf{x}')\|}\right]$$

- Taking derivative wrt normalizer makes gradient complicated
- Solution: **high-dimensional limit**

[Chen-He 21]

- *Prediction MLP*. The prediction MLP (h) has BN applied to its hidden fc layers. Its output fc does not have BN (ablation in Sec. 4.4) or ReLU. This MLP has 2 layers. The dimension of h's input and output (z and p) is $d = 2048$, and h's hidden layer's dimension is 512, making h a bottleneck structure (ablation in supplement).



Approx $\|\mathbf{W}\mathbf{\Phi}\mathbf{x}\| = \mathrm{const.}$

$$\dot{\mathbf{\Phi}} = -\nabla_{\mathbf{\Phi}}\mathcal{L} - \rho\mathbf{\Phi}$$

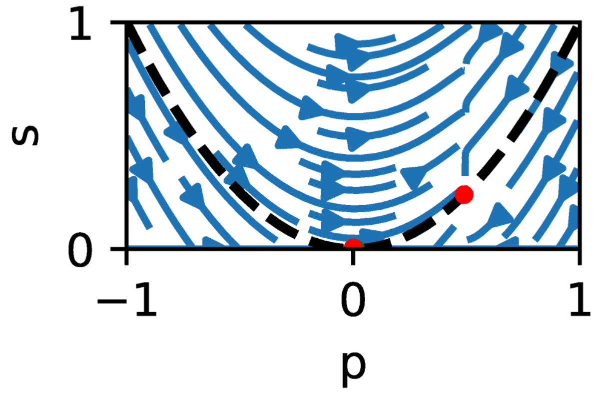
$$\dot{\mathbf{W}} = -\nabla_{\mathbf{W}}\mathcal{L} - \rho\mathbf{W}$$

L2 loss [Tian+ 21]

$$\dot{s} = -2(1+\sigma^2)p^2 s + 2ps - 2\rho s$$

$$\dot{p} = -(1+\sigma^2)ps + s - \rho p$$

Adiabatic elimination



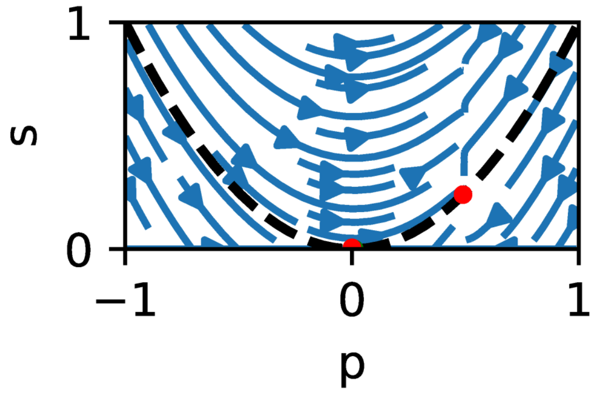
$$\dot{p} = p^2\{1 - (1+\sigma^2)p\} - \rho p$$

Cosine loss

$$\dot{s}_j = -\frac{2}{(1+\sigma^2)N_\Phi N_\Psi}\left(\frac{2}{N_\Psi^2}s_j^2 p_j^3 + N_\times s_j p_j^2 - s_j p_j\right) - 2\rho s_j.$$

$$\dot{p}_j = -\frac{1}{(1+\sigma^2)N_\Phi N_\Psi}\left(\frac{2}{N_\Psi^2}s_j^2 p_j^2 + N_\times s_j p_j - s_j\right) - \rho p_j,$$

Adiabatic elimination

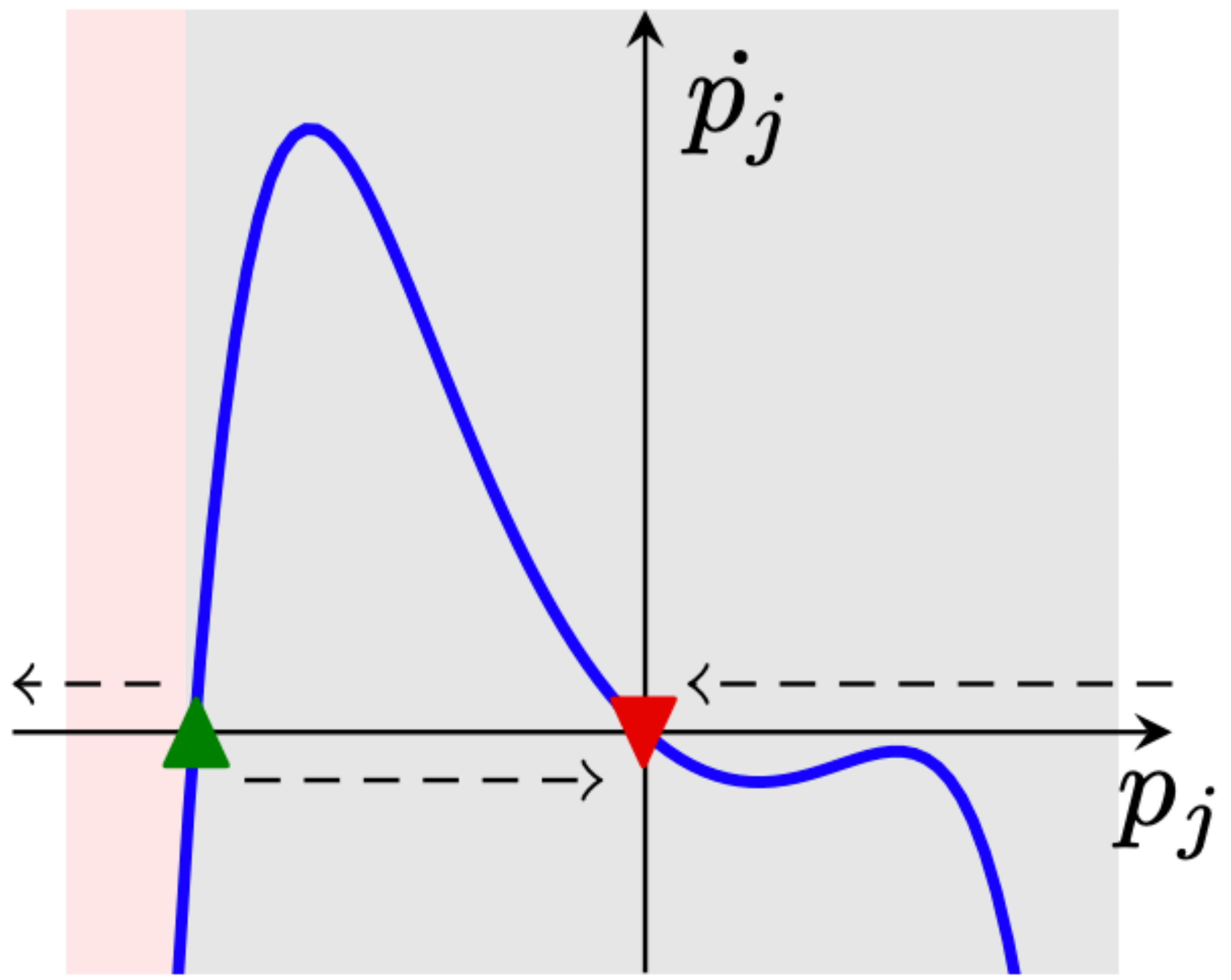


$$\dot{p} = -\frac{2C_1 p^6 + C_2 p^3 - C_3 p^2}{1+\sigma^2} - \rho p$$

Bifurcation: collapsed solution is not stable 😁

Eigval ODE of projector (C_i depends on $\|\mathbf{W}\mathbf{\Phi}\mathbf{x}\|$)

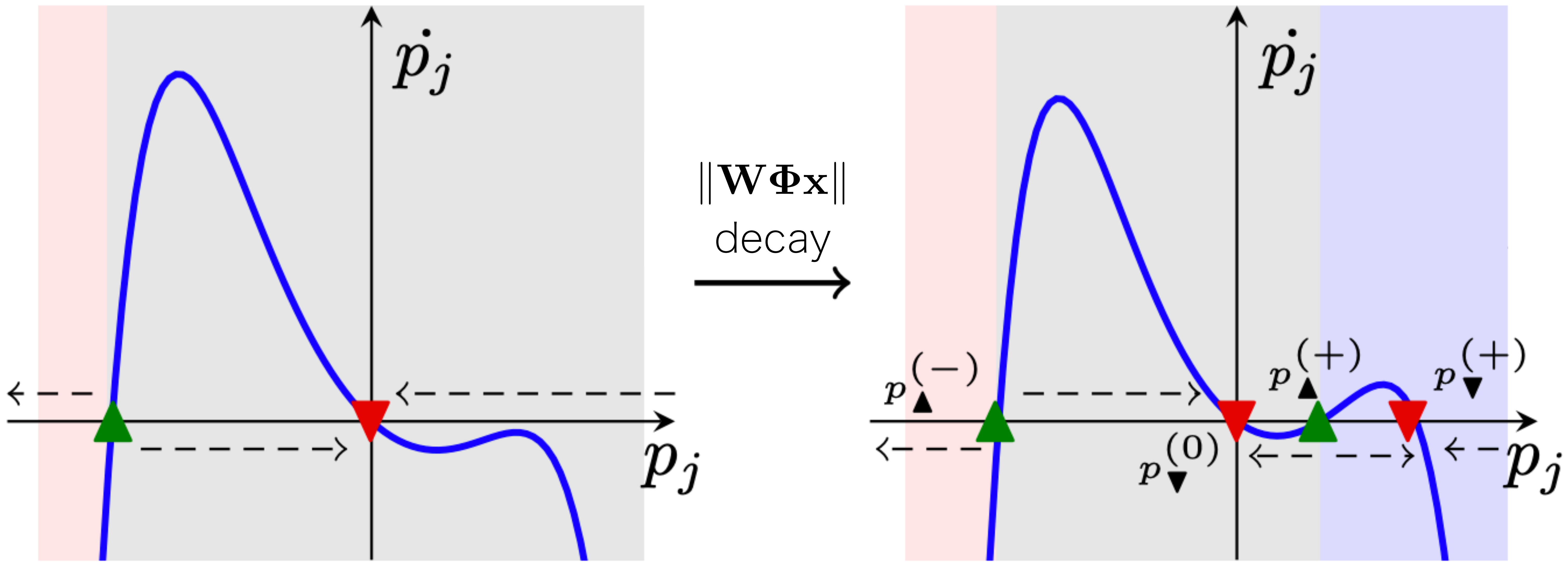
$$\dot{p} = -\frac{2C_1p^6 + C_2p^3 - C_3p^2}{1+\sigma^2} - \rho p$$



Bifurcation: collapsed solution is not stable 😁

Eigval ODE of projector (C_i depends on $\|\mathbf{W\Phi x}\|$)

$$\dot{p} = -\frac{2C_1p^6 + C_2p^3 - C_3p^2}{1+\sigma^2} - \rho p$$

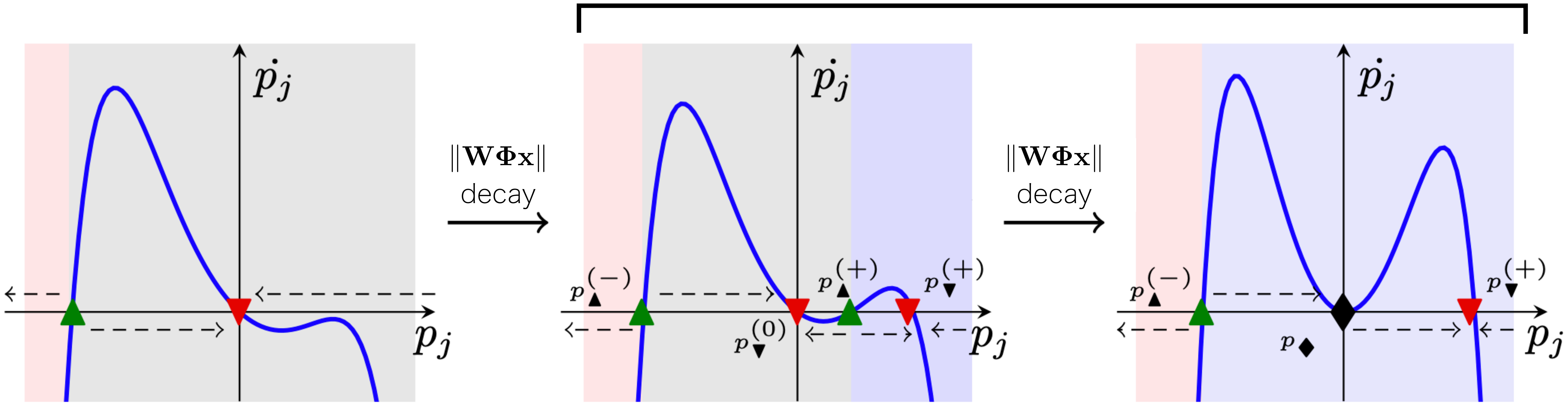


Bifurcation: collapsed solution is not stable 😁

Eigval ODE of projector (C_i depends on $\|\mathbf{W\Phi x}\|$)

$$\dot{p} = -\frac{2C_1p^6 + C_2p^3 - C_3p^2}{1+\sigma^2} - \rho p$$

non-trivial solution exists



collapse $p = 0$ is **saddle**!

Part 1: What we learn from nonlinear dynamics

- 😁 Dynamics analysis provides **stability analysis** beyond analyzing loss minimizer solely
  - Why StopGrad? Why encoder-predictor? etc.
- 😁 Difference loss function may yield more **adaptivity** during optimization

$$\dot{p} = p^2\{1 - (1+\sigma^2)p\} - \rho p$$

L2 dynamics

$$\dot{p} = -\frac{2C_1 p^6 + C_2 p^3 - C_3 p^2}{1+\sigma^2} - \rho p$$

cosine dynamics

- 😢 What we don't answer: **feature learning**
  - since analytical solution to ODE typically requires strong Gaussianity assumptions

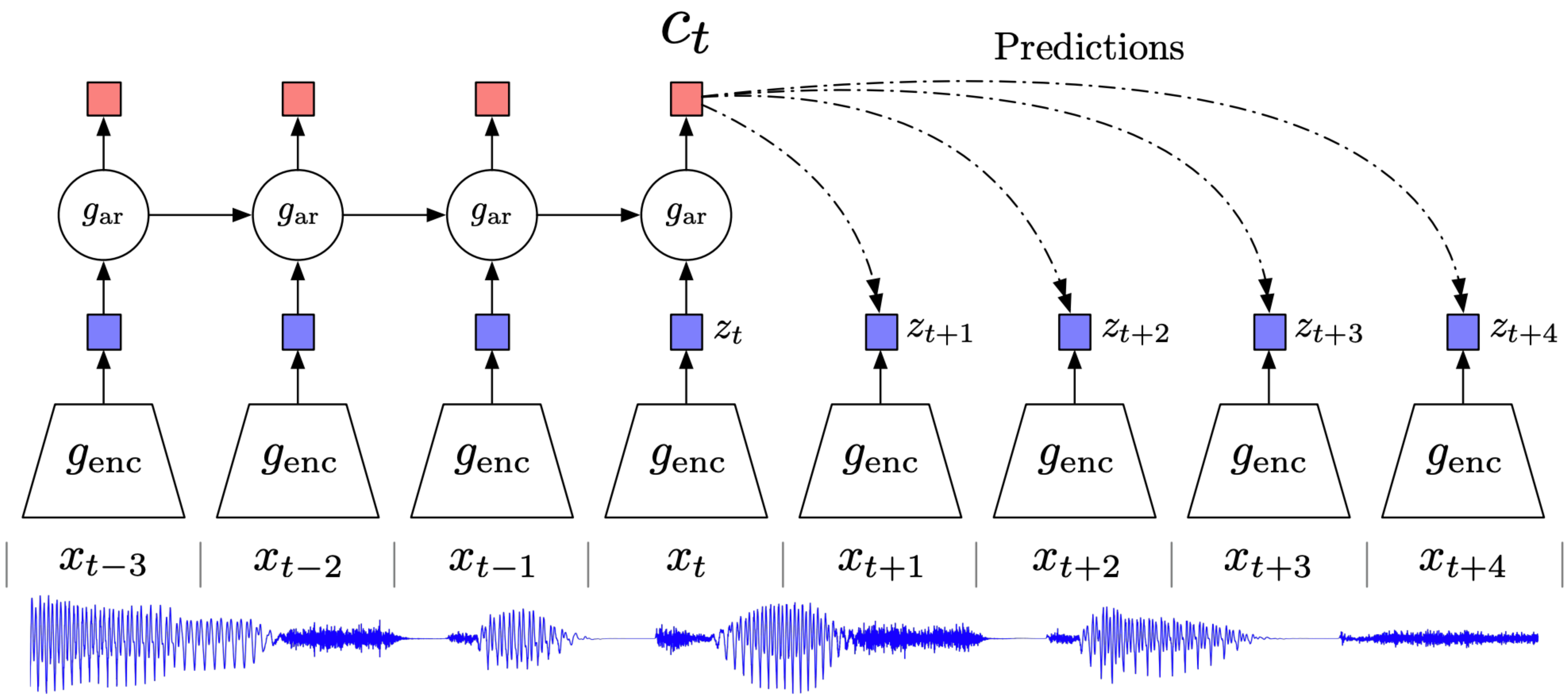
What we can learn from nonlinear dynamics and **neuroscience**

Ishikawa, S.*, Yamada, M.*, **Bao, H.**, & Takezawa, Y. (ICLR2025)
PhiNets: Brain-inspired Non-contrastive Learning Based on Temporal Prediction Hypothesis.

Predictive coding

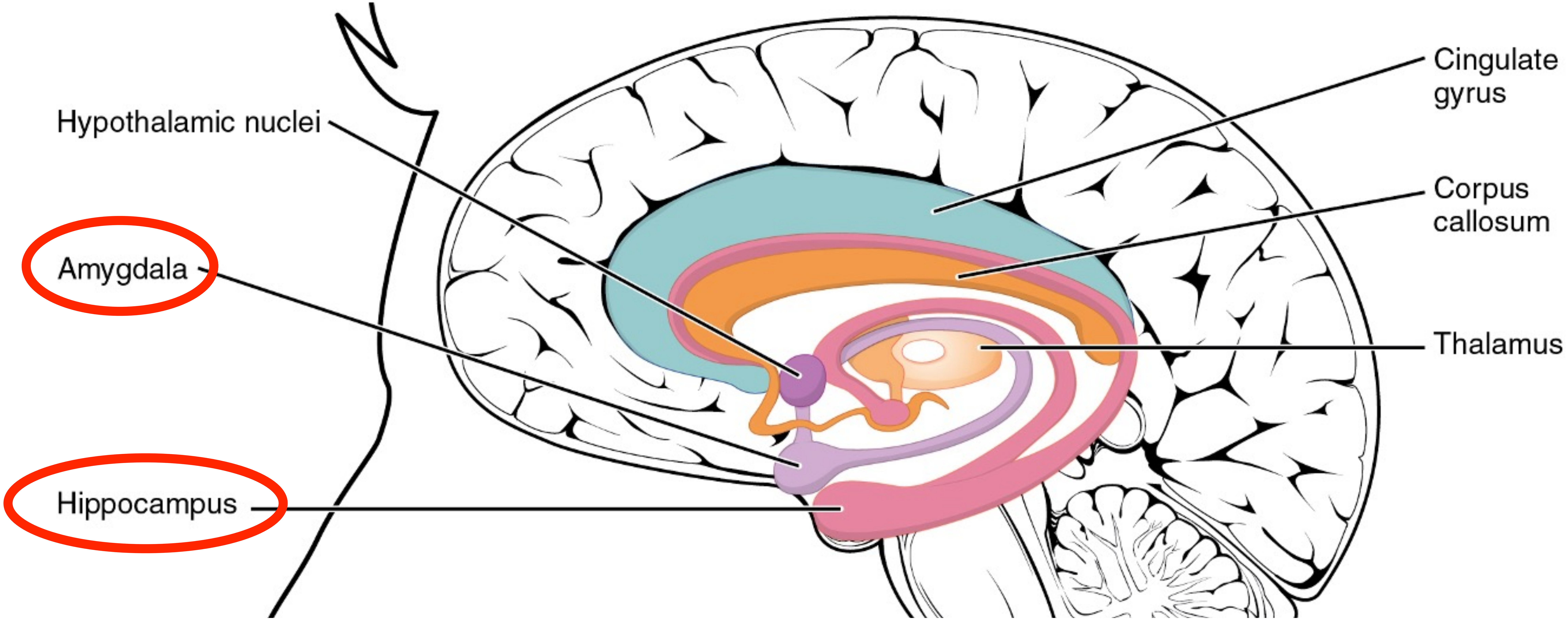
- Brain predicts a future/neighboring input signal
  - ❖ dopamine is secreted if prediction makes a mistake

Contrastive predictive coding [van den Oord+ 18]



Predictive coding

- Brain predicts a future/neighboring input signal **at various level**



https://en.wikipedia.org/wiki/Hippocampus

Neocortex and hippocampus

Hippocampus

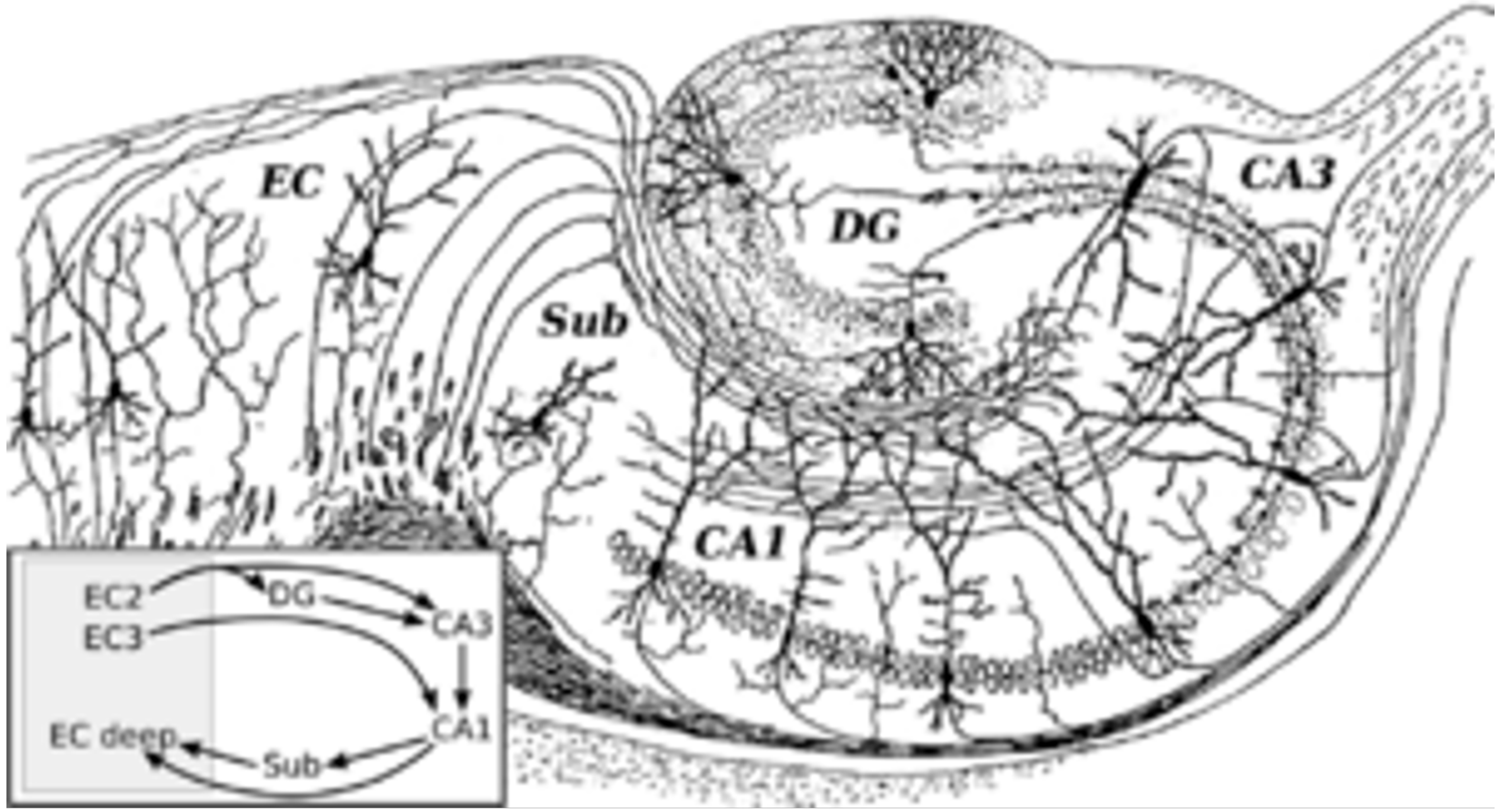
Short-term memory

external stimuli

Neocortex

Long-term memory

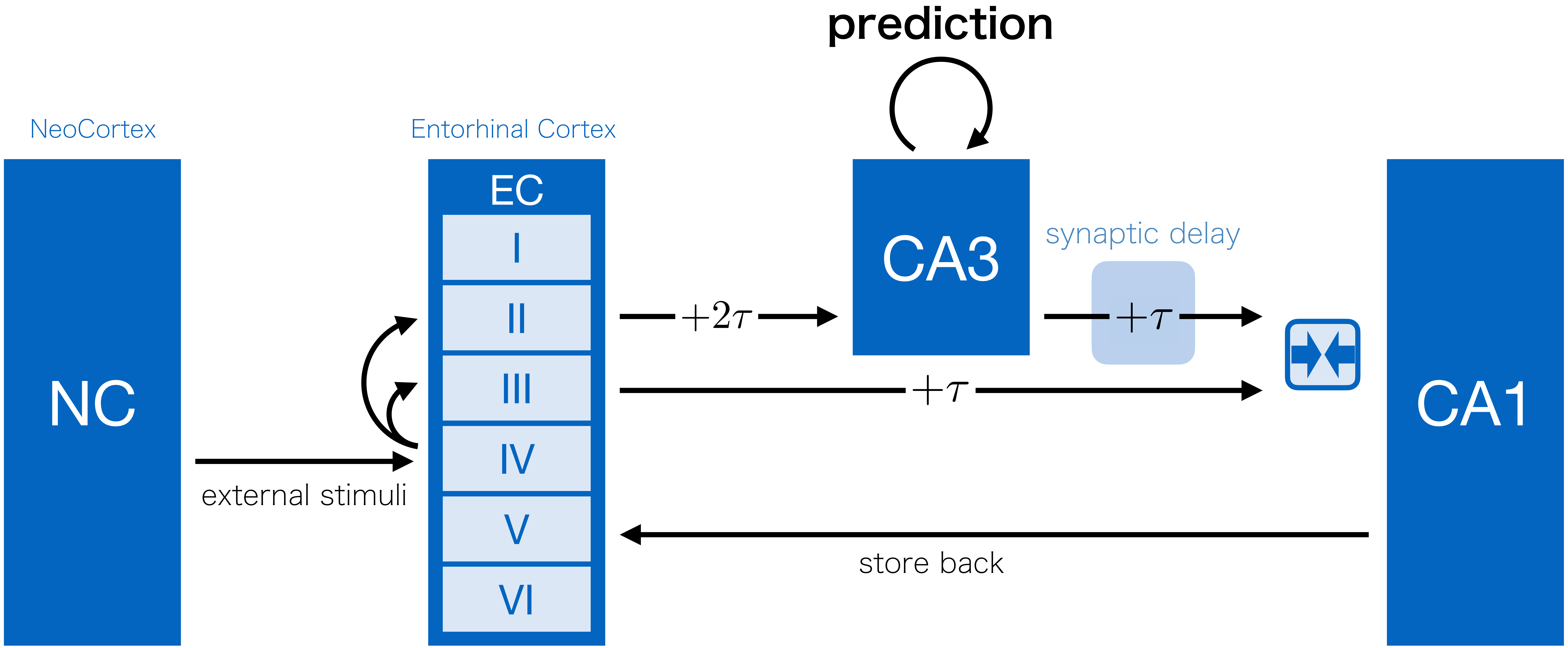
store memory



https://en.wikipedia.org/wiki/Hippocampus

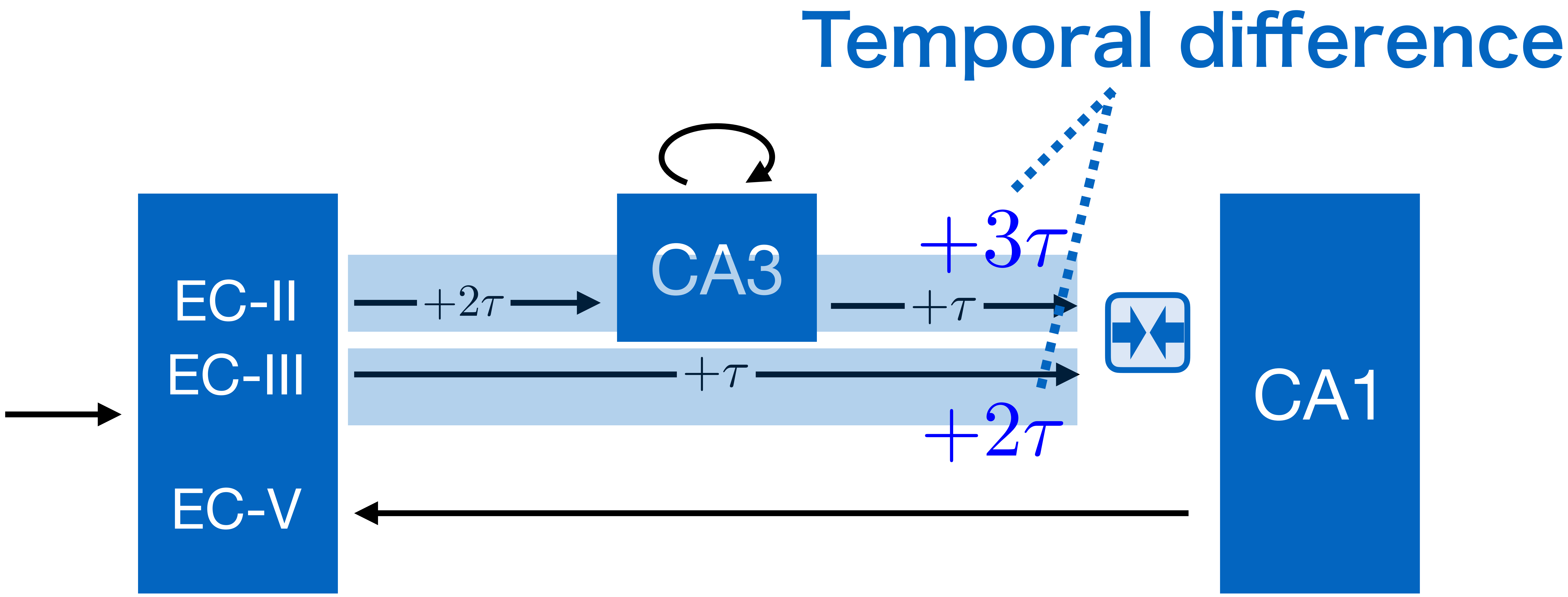
Hippocampus as a self-supervised learning model

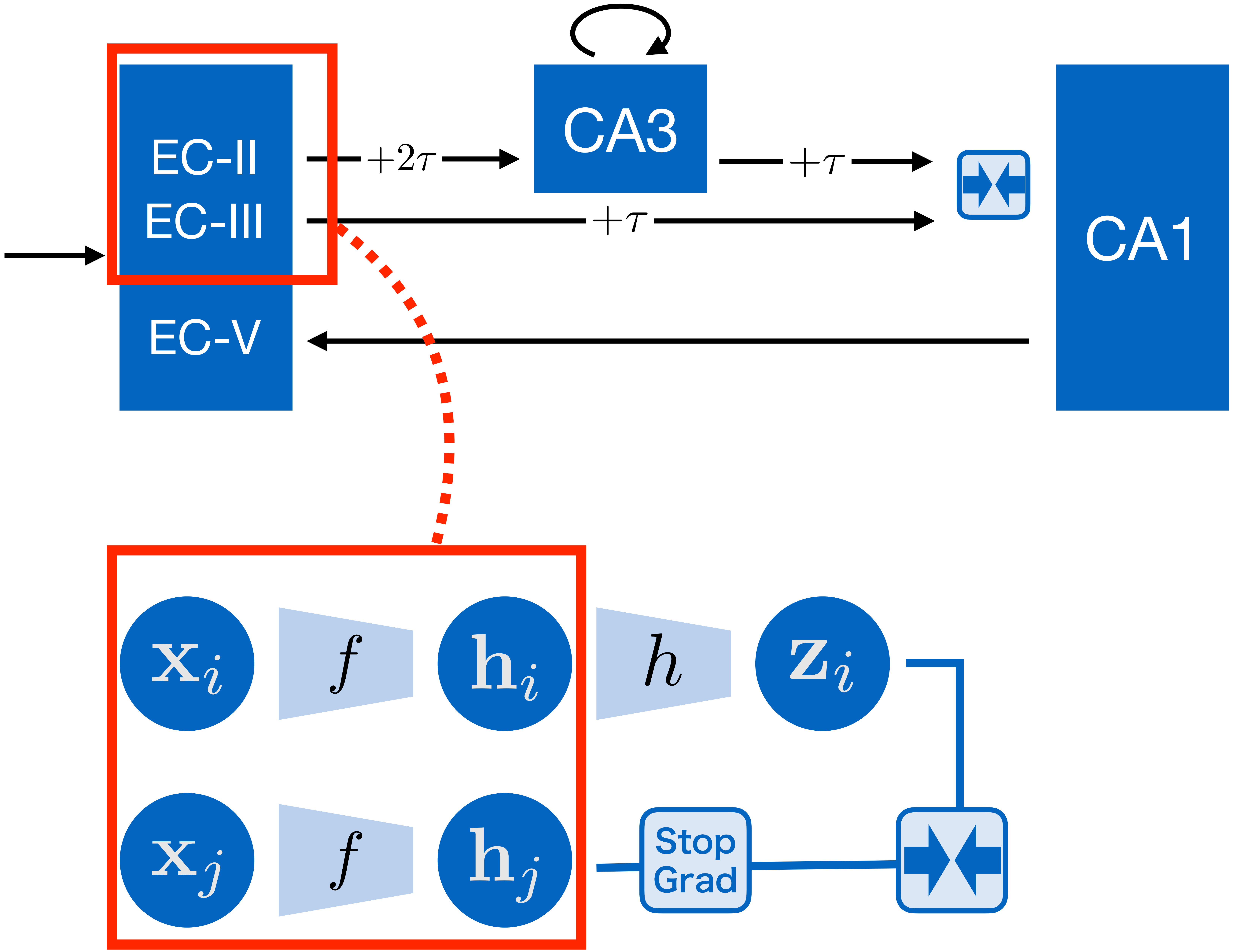
[Chen+ 24]

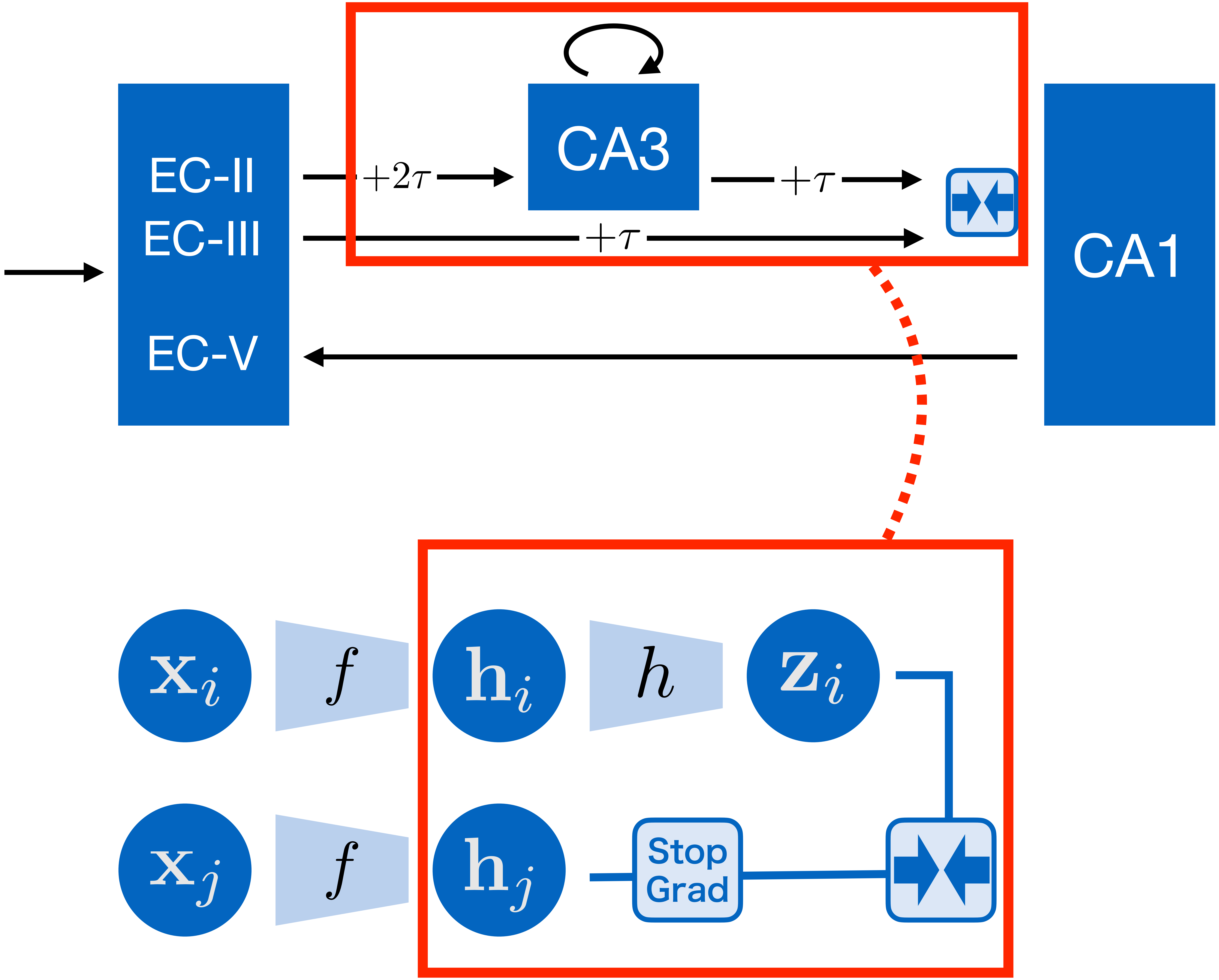


Transmission delay b/w CA3 and CA1 forms a self-supervised feedback
⇒ with prediction, neural activity is replicated more accurately

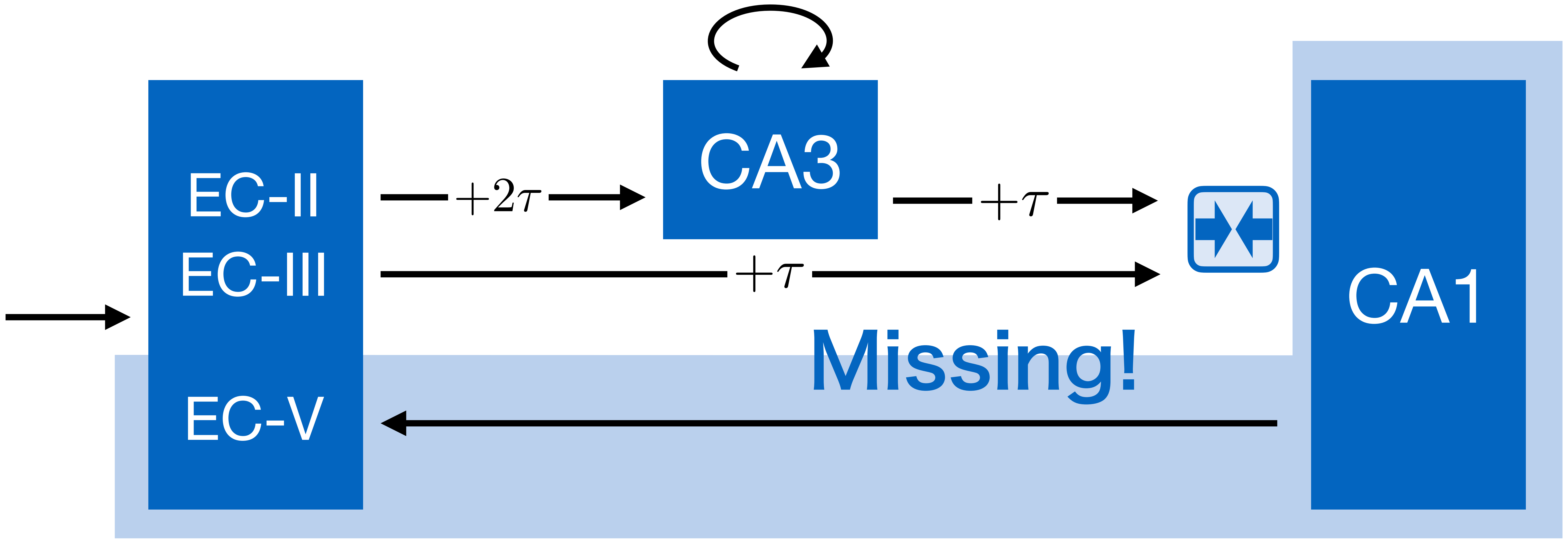
Temporal prediction hypothesis

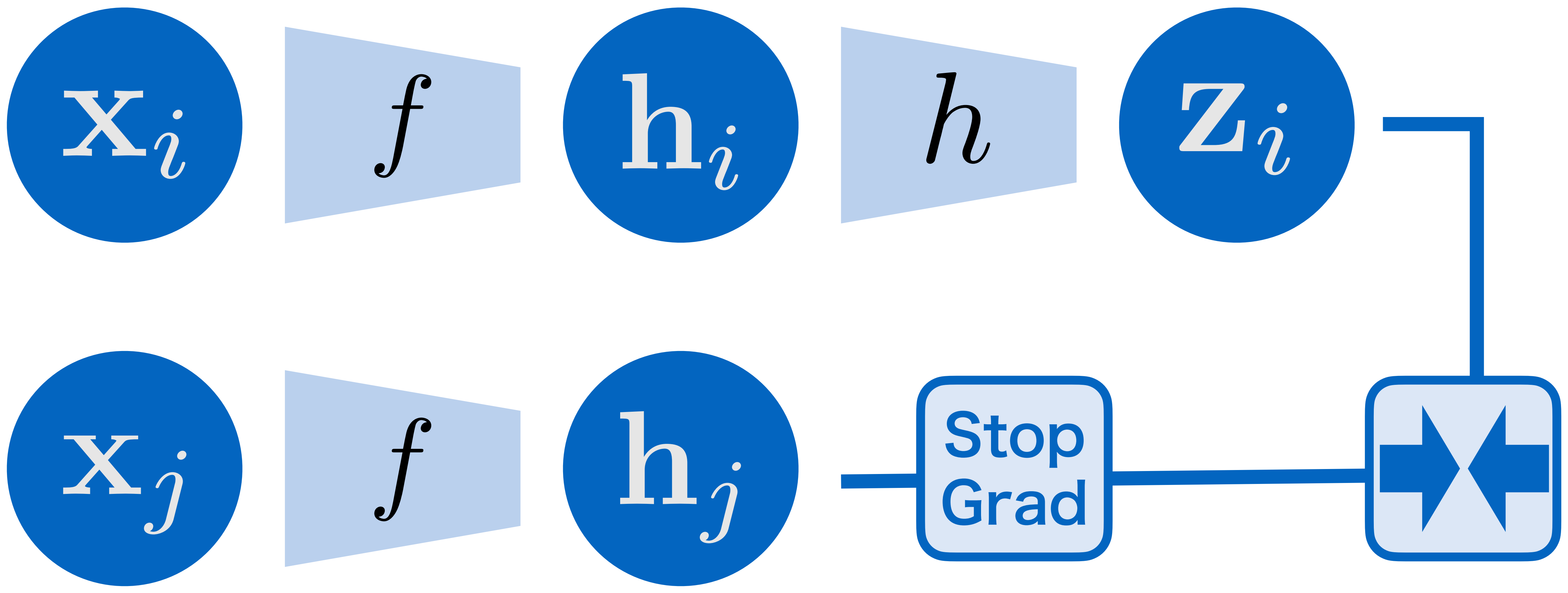


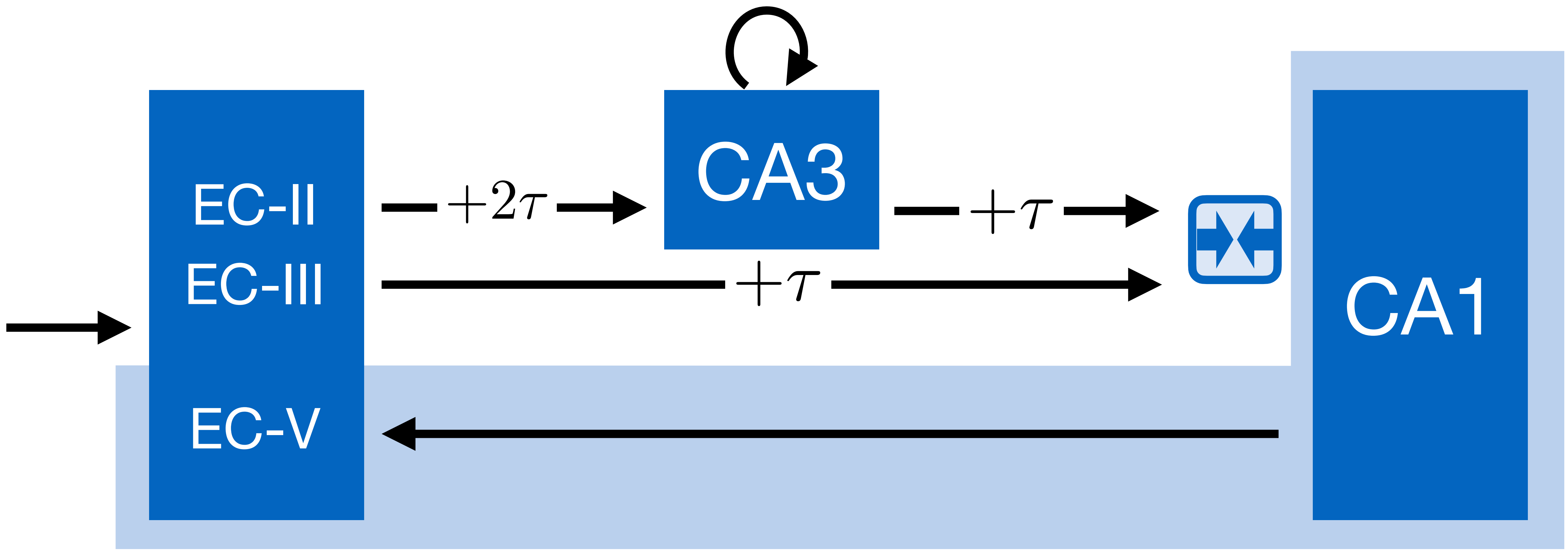

EC-II
EC-III
EC-V
$+2\tau$
CA3
$+\tau$
$+\tau$
CA1
$\mathbf{x}_i$
f
$\mathbf{h}_i$
h
$\mathbf{z}_i$
$\mathbf{x}_j$
f
$\mathbf{h}_j$
Stop Grad

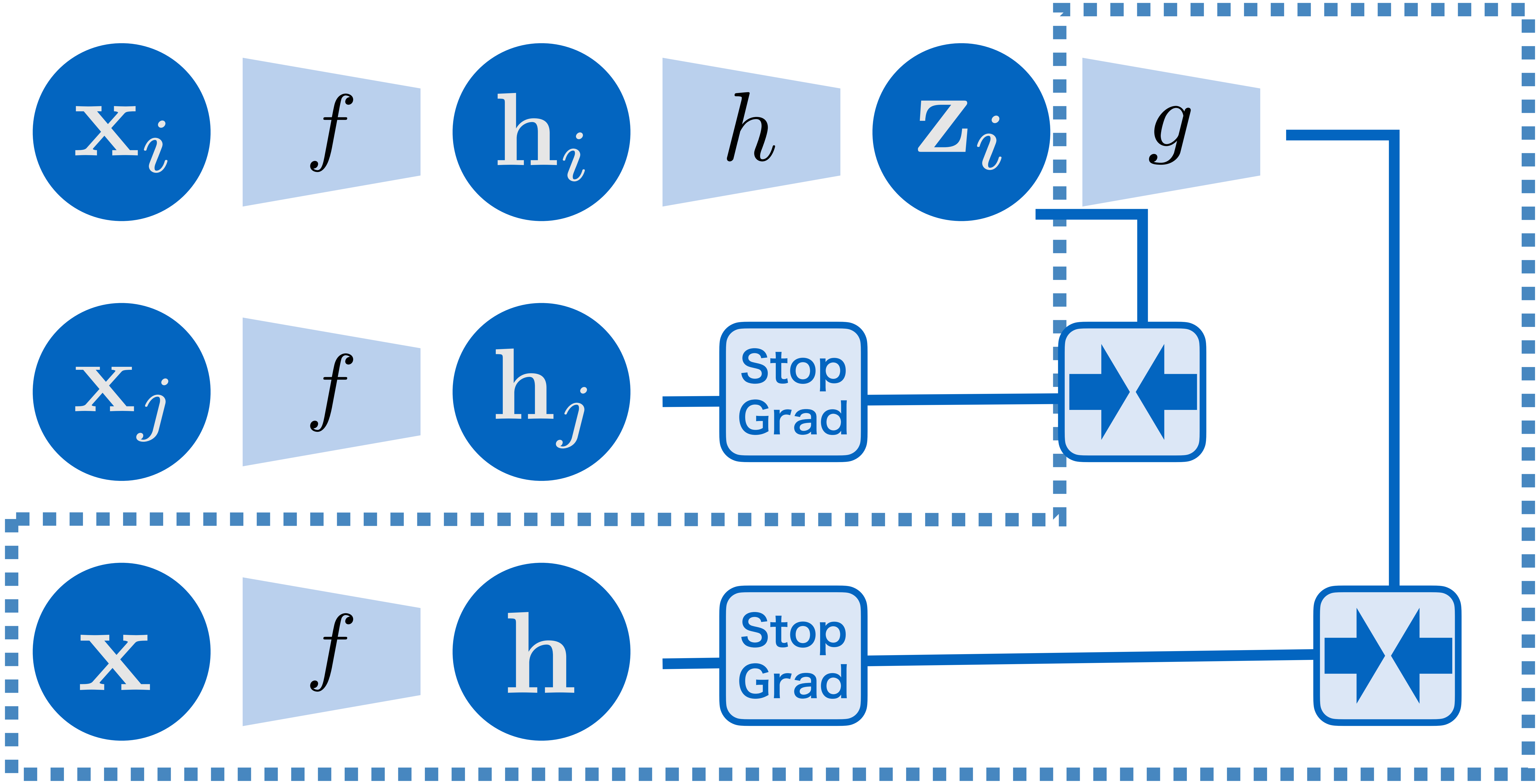


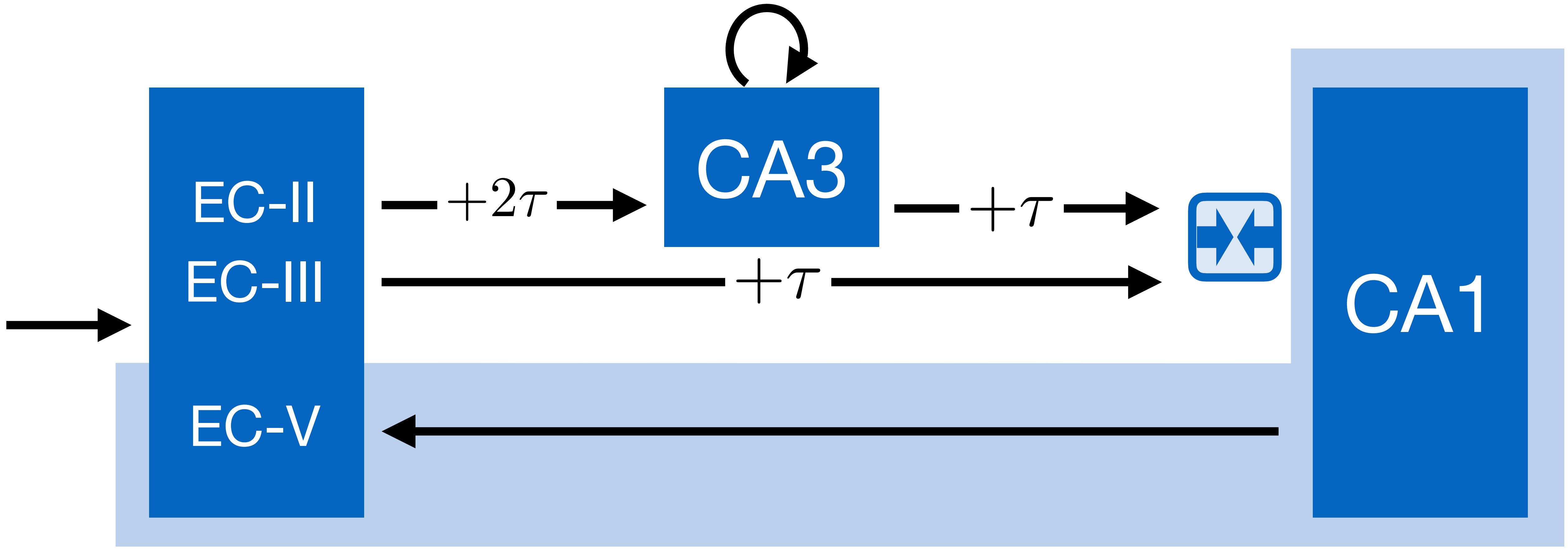
EC-II
EC-III
EC-V
$+2\tau$
CA3
$+\tau$
$+\tau$
CA1
$\mathbf{x}_i$
f
$\mathbf{h}_i$
h
$\mathbf{z}_i$
$\mathbf{x}_j$
f
$\mathbf{h}_j$
Stop Grad

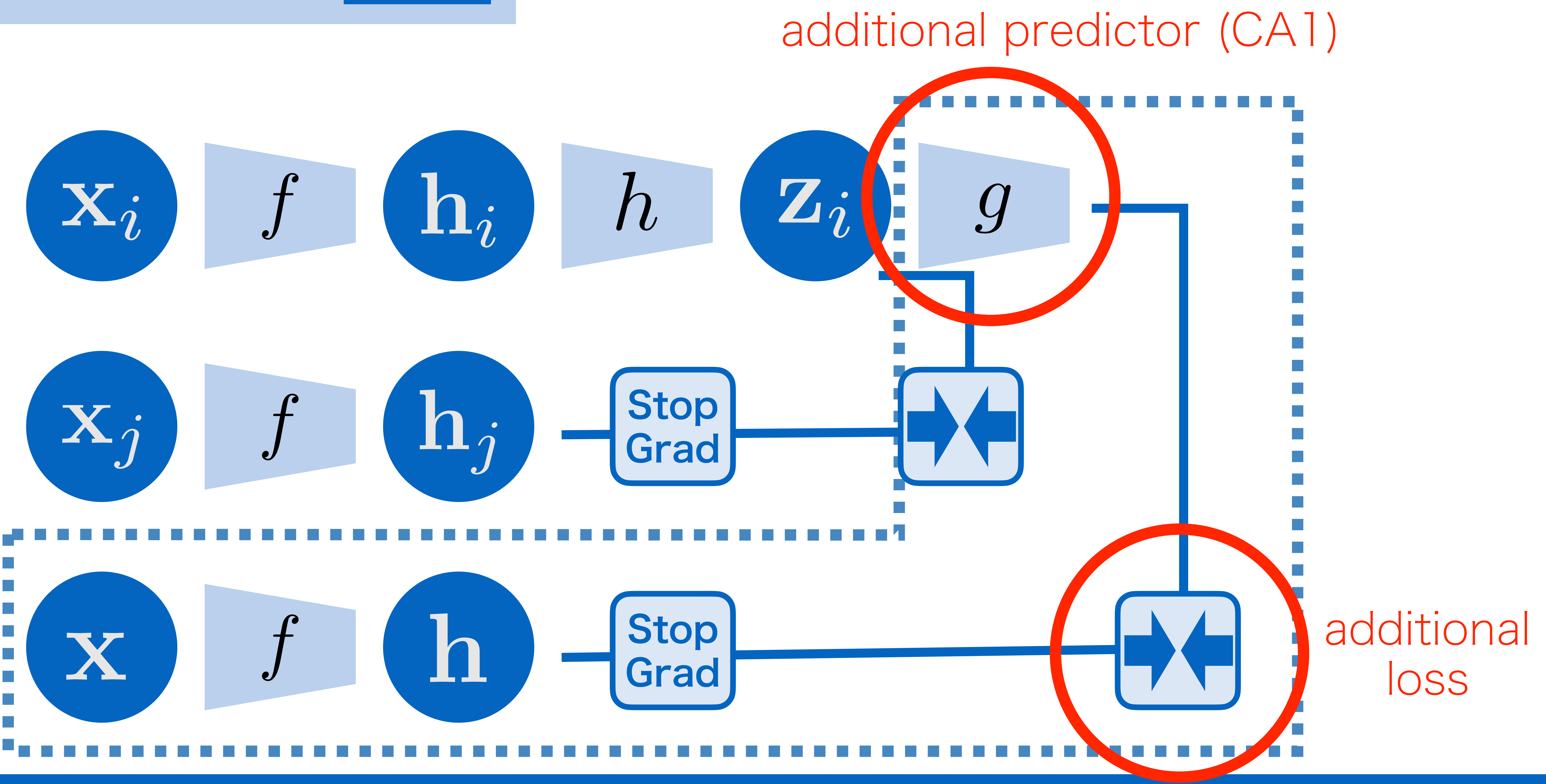

EC-II
EC-III
EC-V
+2τ
CA3
+τ
+τ
CA1
Missing!


$\mathbf{x}_i$
f
$\mathbf{h}_i$
h
$\mathbf{z}_i$
$\mathbf{x}_j$
f
$\mathbf{h}_j$
Stop Grad

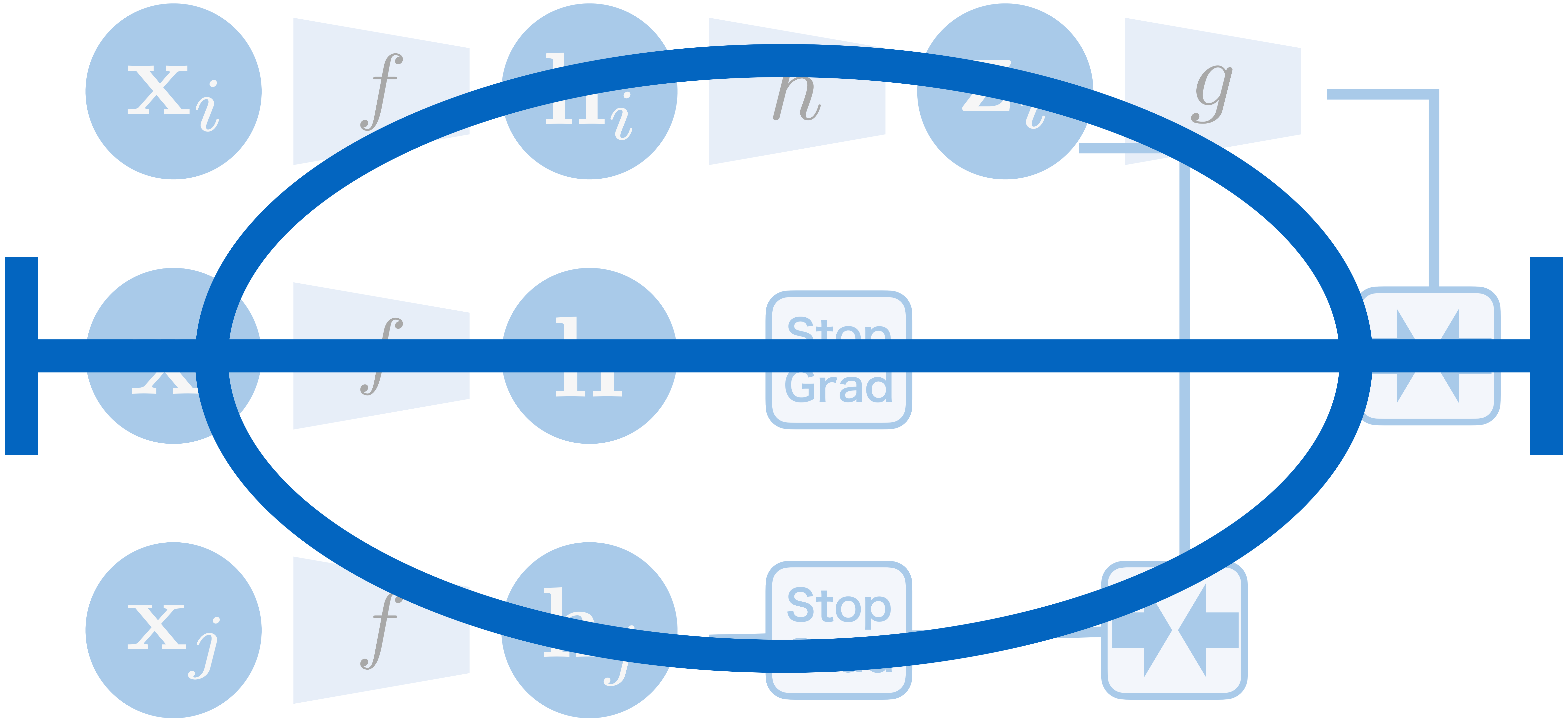

EC-II
EC-III
EC-V
$+2\tau$
CA3
$+\tau$
$+\tau$
CA1


$\mathbf{x}_i$
f
$\mathbf{h}_i$
h
$\mathbf{z}_i$
g
$\mathbf{x}_j$
f
$\mathbf{h}_j$
Stop Grad
$\mathbf{x}$
f
$\mathbf{h}$
Stop Grad

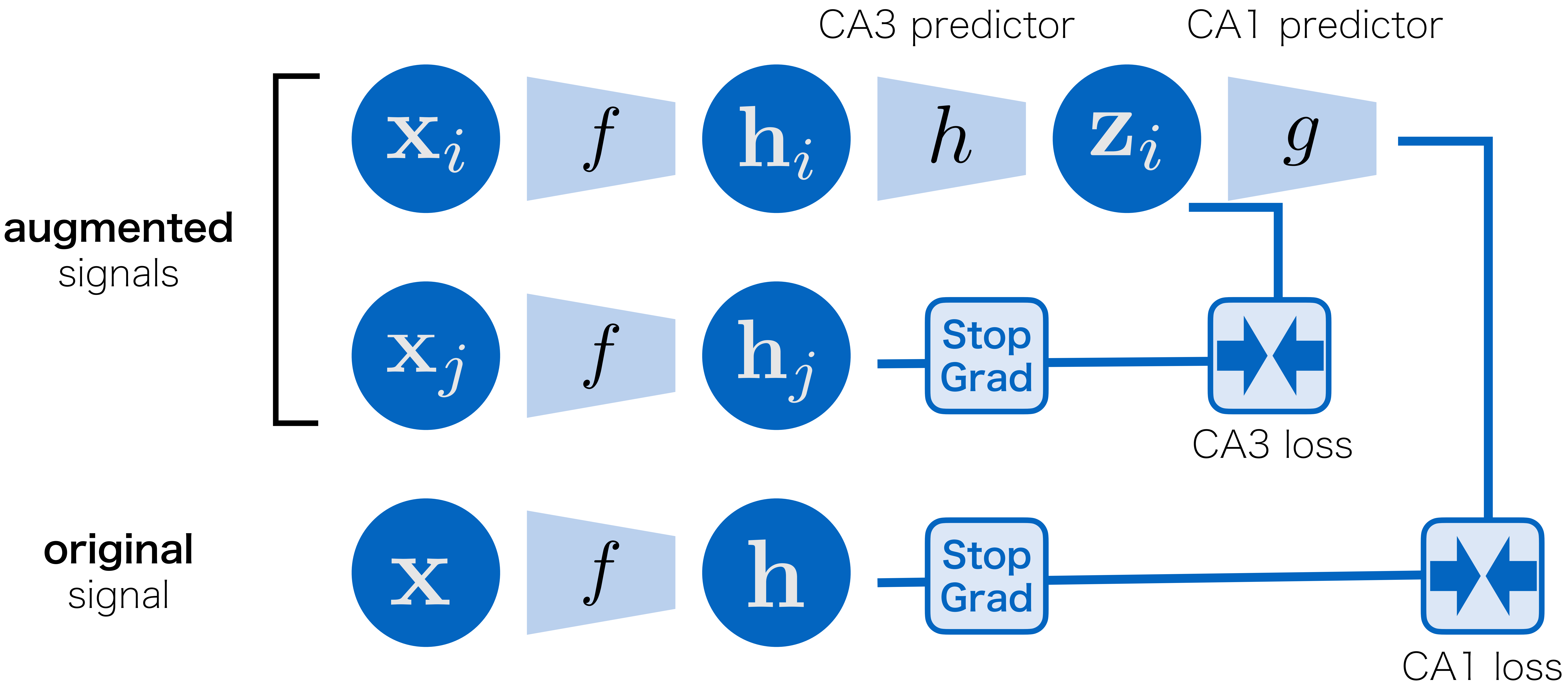

EC-II
EC-III
EC-V
CA3
CA1
$+2\tau$
$+\tau$
$+\tau$


additional predictor (CA1)
$\mathbf{x}_i$
f
$\mathbf{h}_i$
h
$\mathbf{z}_i$
g
$\mathbf{x}_j$
f
$\mathbf{h}_j$
Stop Grad
$\mathbf{x}$
f
$\mathbf{h}$
Stop Grad
additional loss

Φ-Net



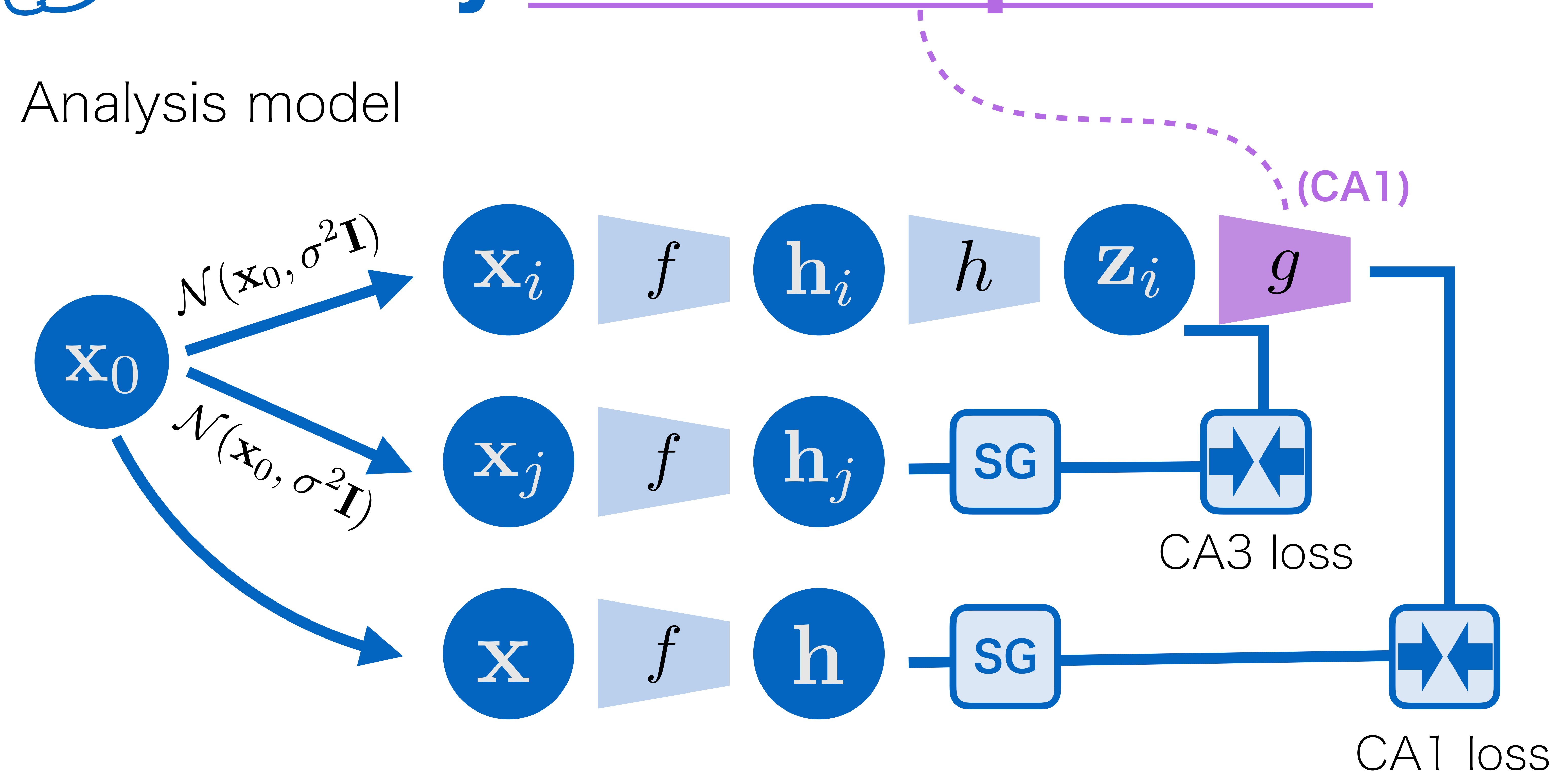
Implementation details



- Encoder f is shared
- All layers are optimized by backprop simultaneously

But why additional predictor?

Analysis model



$$\mathcal{L}(\mathbf{W}_f, \mathbf{W}_g, \mathbf{W}_h) = \frac{1}{2}\mathbb{E}\left[\underbrace{\|\mathbf{W}_h\mathbf{W}_f\mathbf{x}_1 - \mathrm{SG}(\mathbf{W}_f\mathbf{x}_2)\|^2}_{\text{CA3 loss}} + \underbrace{\|\mathbf{W}_g\mathbf{W}_h\mathbf{W}_f\mathbf{x}_1 - \mathrm{SG}(\mathbf{W}_f\mathbf{x})\|^2}_{\text{CA1 loss}}\right]$$

Disclaimer: cosine loss is not considered for simplicity

🤔 But why additional predictor?

$$\mathcal{L}(\mathbf{W}_f, \mathbf{W}_g, \mathbf{W}_h) = \frac{1}{2}\mathbb{E}\left[\|\mathbf{W}_h\mathbf{W}_f\mathbf{x}_1 - \mathrm{SG}(\mathbf{W}_f\mathbf{x}_2)\|^2 + \|\mathbf{W}_g\mathbf{W}_h\mathbf{W}_f\mathbf{x}_1 - \mathrm{SG}(\mathbf{W}_f\mathbf{x})\|^2\right]$$

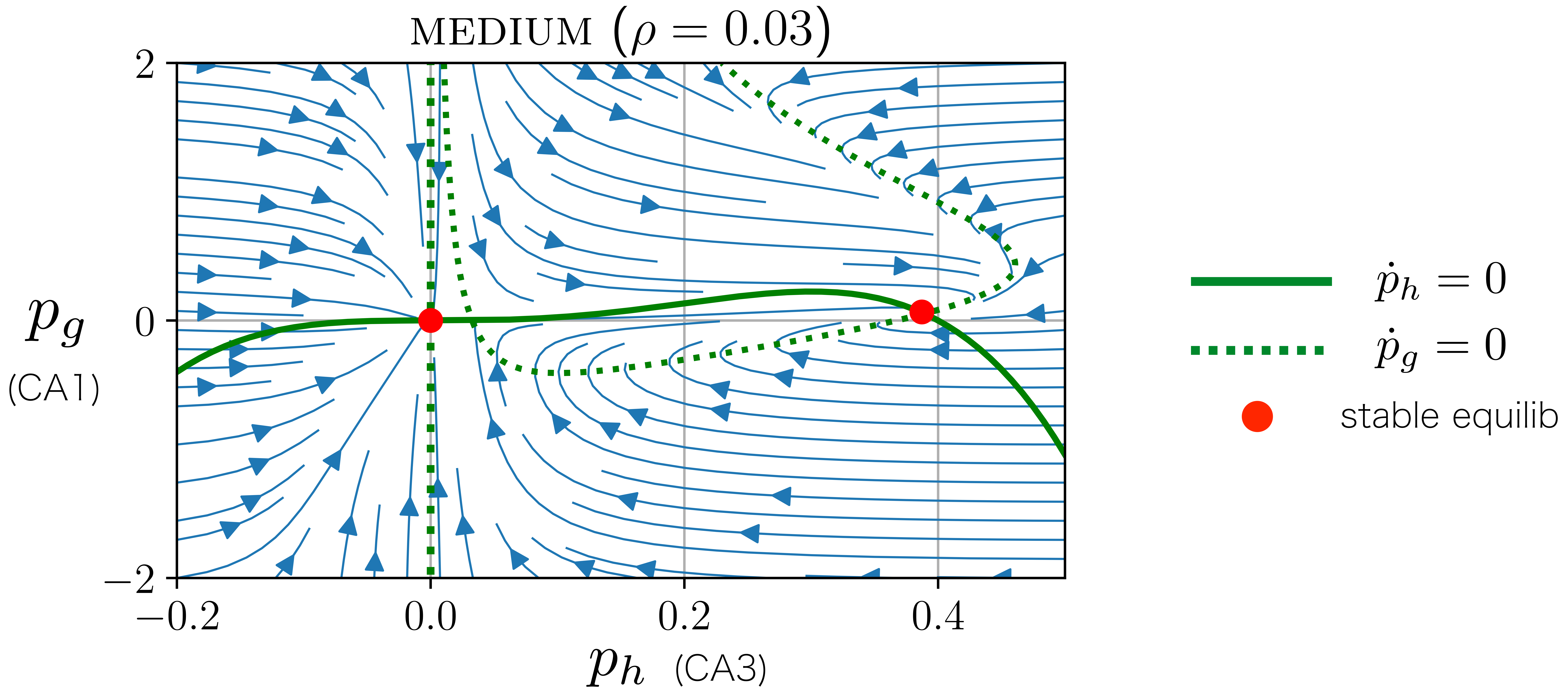
↓ eigendecomposition
adiabatic elimination

$$\begin{cases} \dot{p}_h = \{(1+p_g) - (1+\sigma^2)(1+p_g^2)p_h\}p_h^2 - \rho p_h & \text{(CA3 predictor)} \\ \dot{p}_g = \{1 - (1+\sigma^2)p_h\}p_h^3 - \rho p_g & \text{(CA1 predictor)} \end{cases}$$

cf. SimSiam dynamics

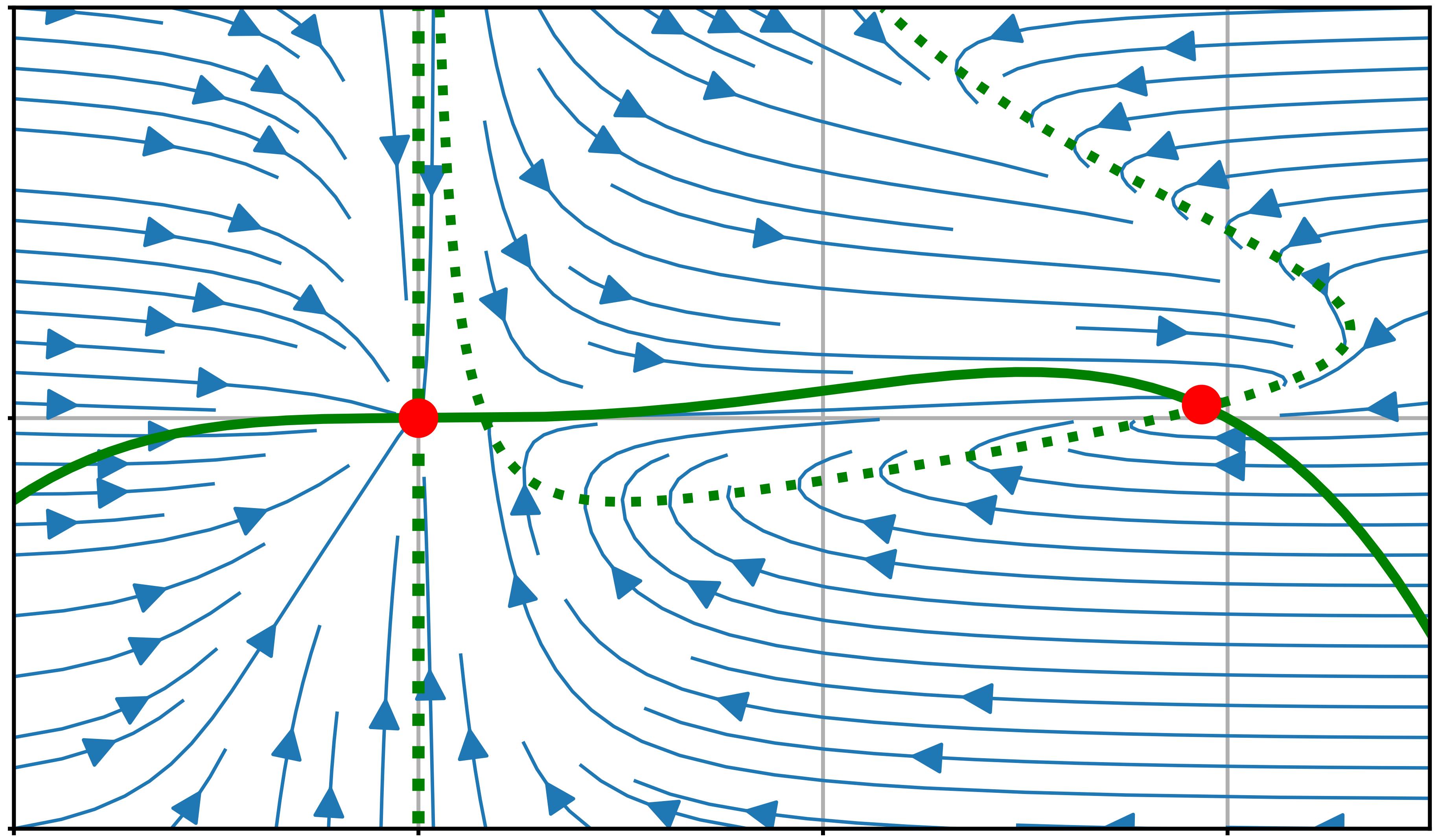
$$\dot{p} = p^2\{1 - (1+\sigma^2)p\} - \rho p \quad \text{(CA3 predictor)}$$

PhiNet dynamics (2D)

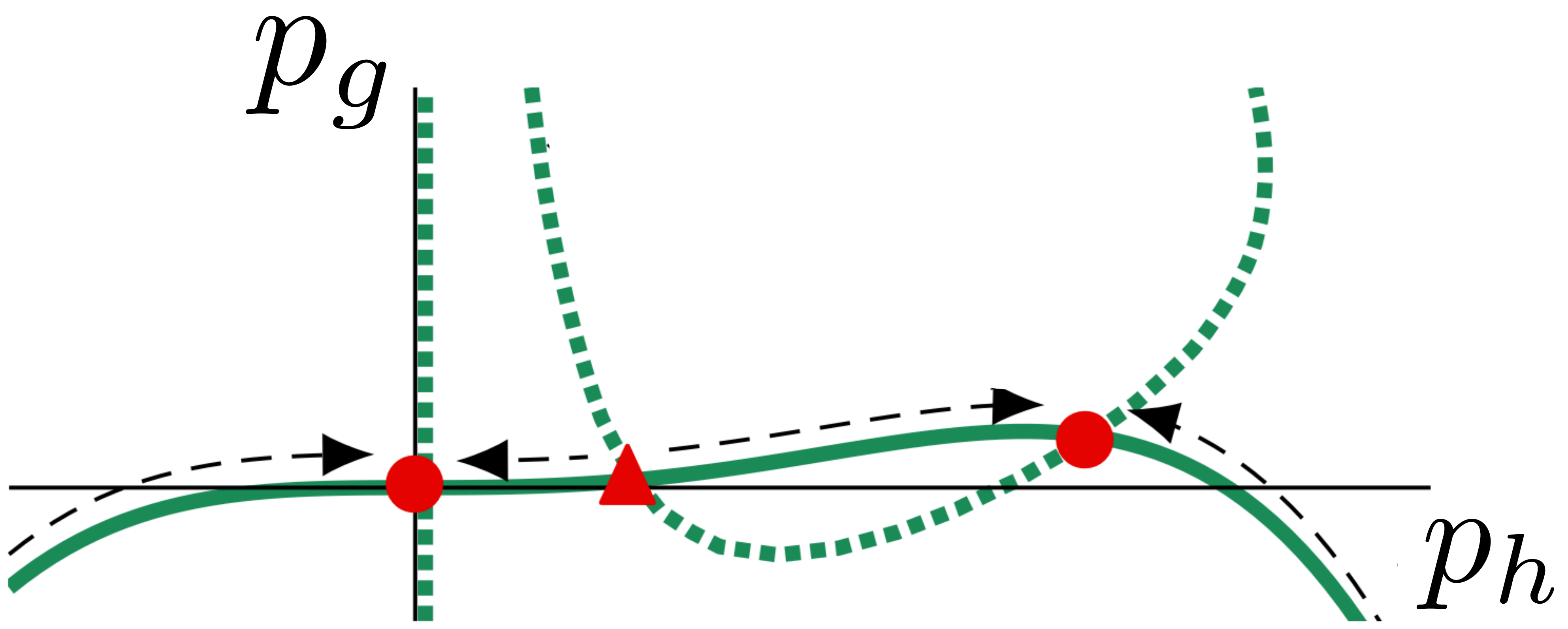


PhiNet (2D) vs SimSiam (1D)

PhiNet dynamics

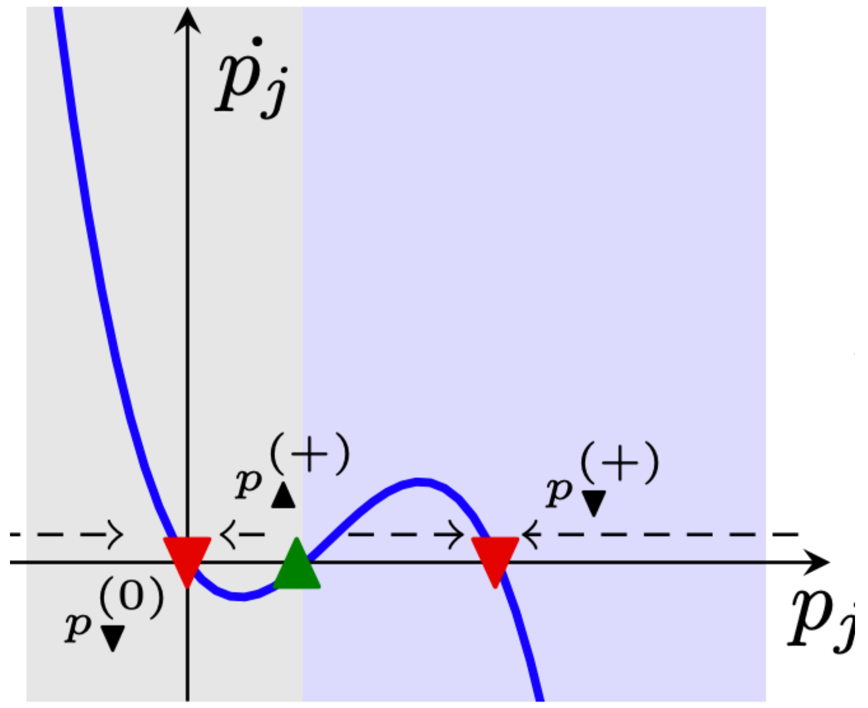


extract nullclines →

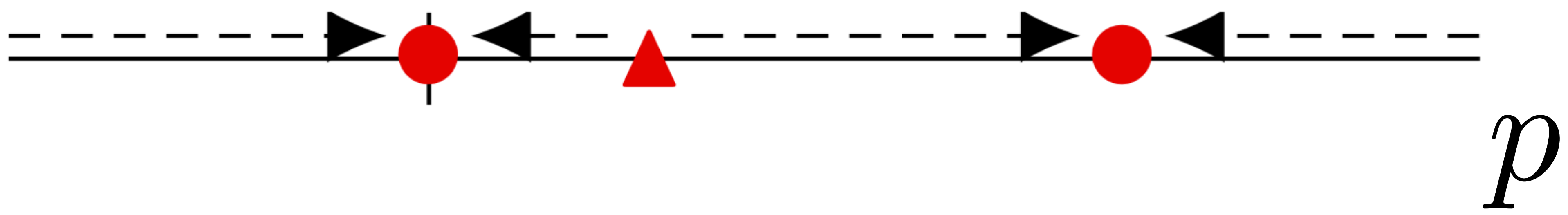


topologically conjugate

SimSiam dynamics

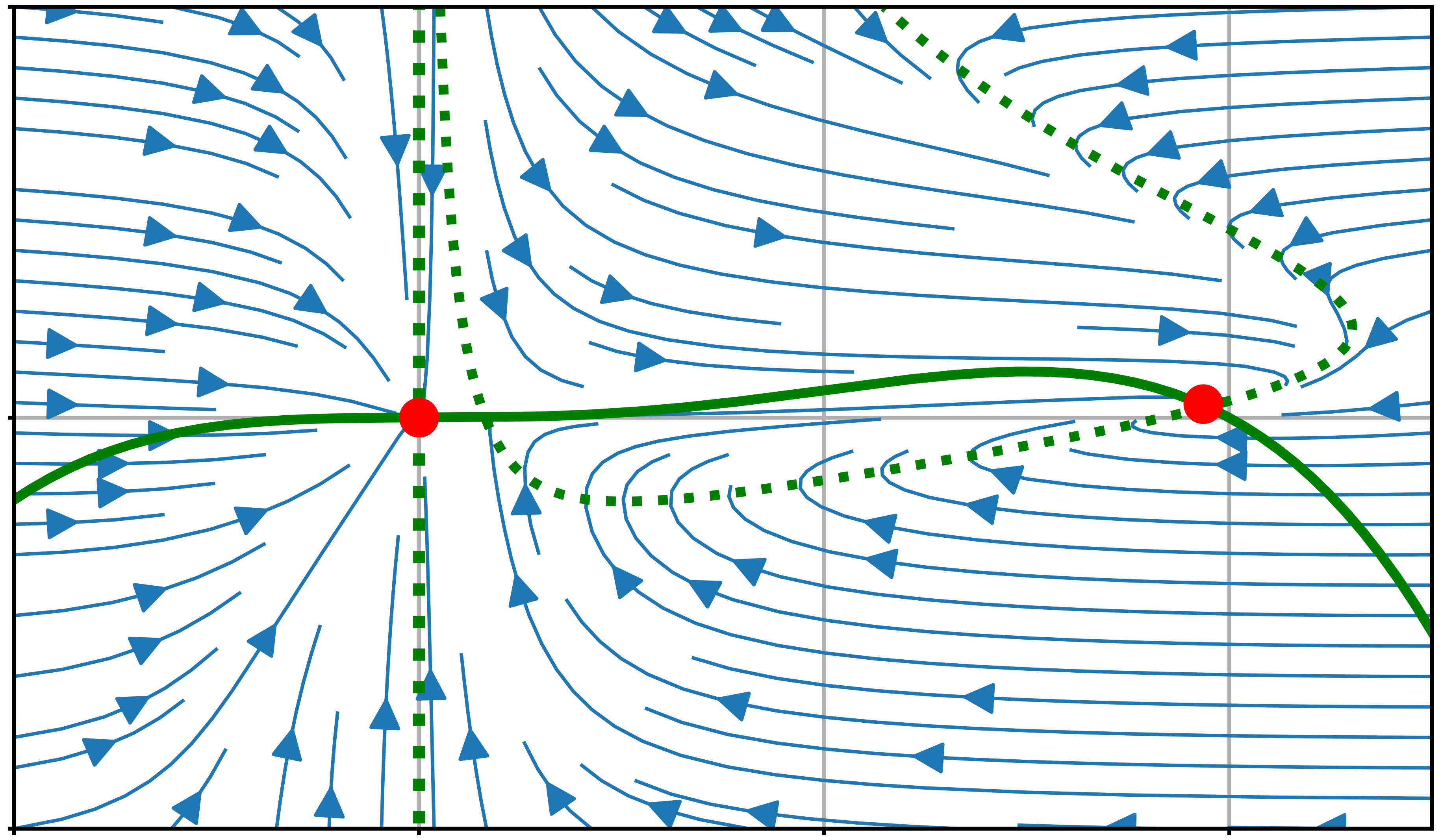


extract nullclines →

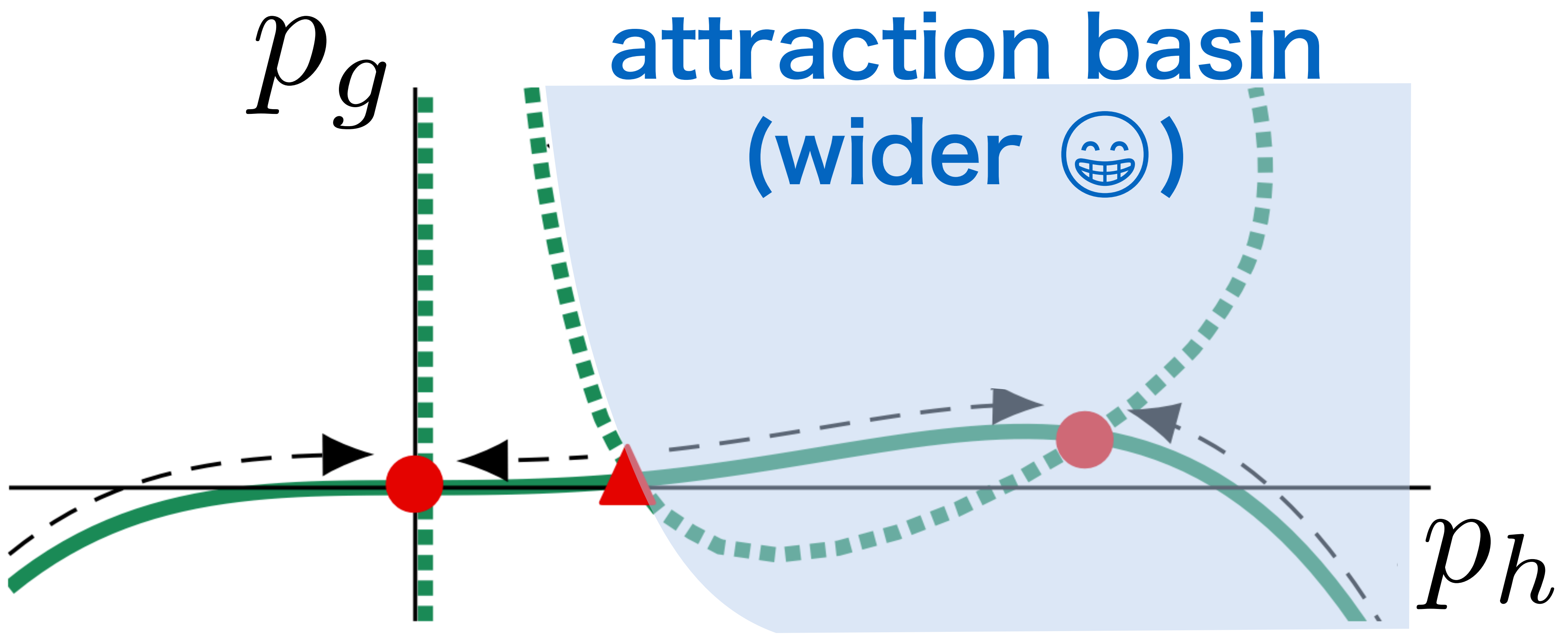


PhiNet (2D) vs SimSiam (1D)

PhiNet dynamics

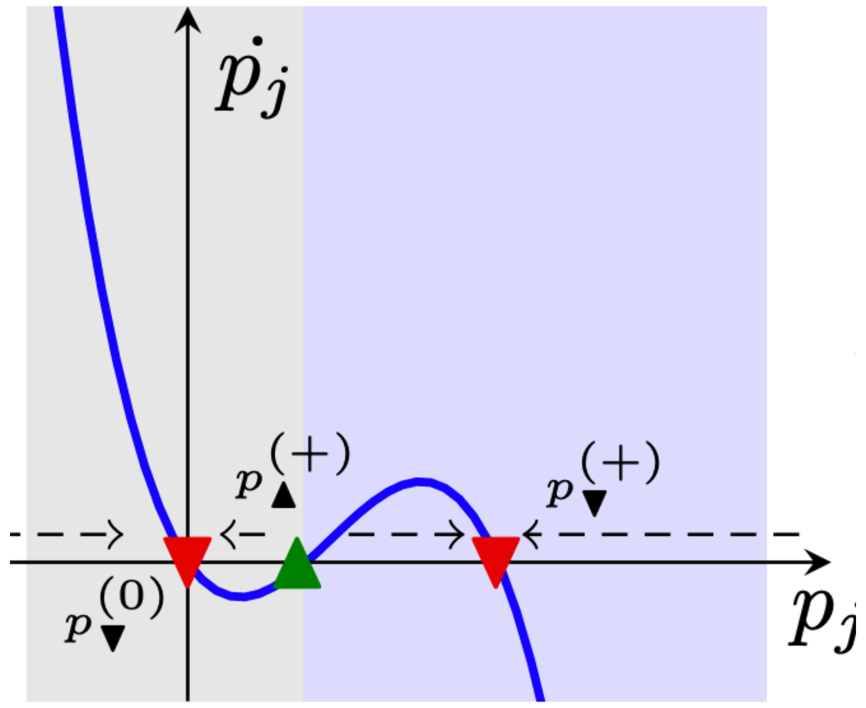


extract nullclines →

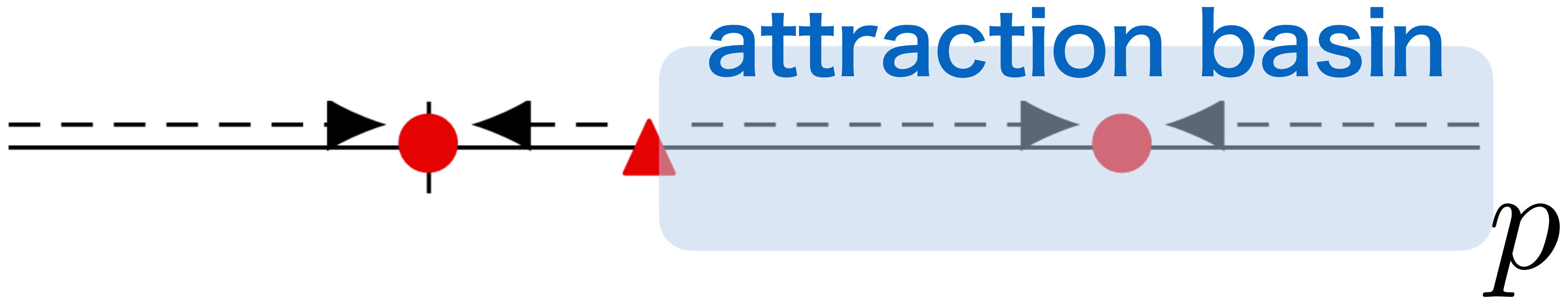


↕ topologically conjugate

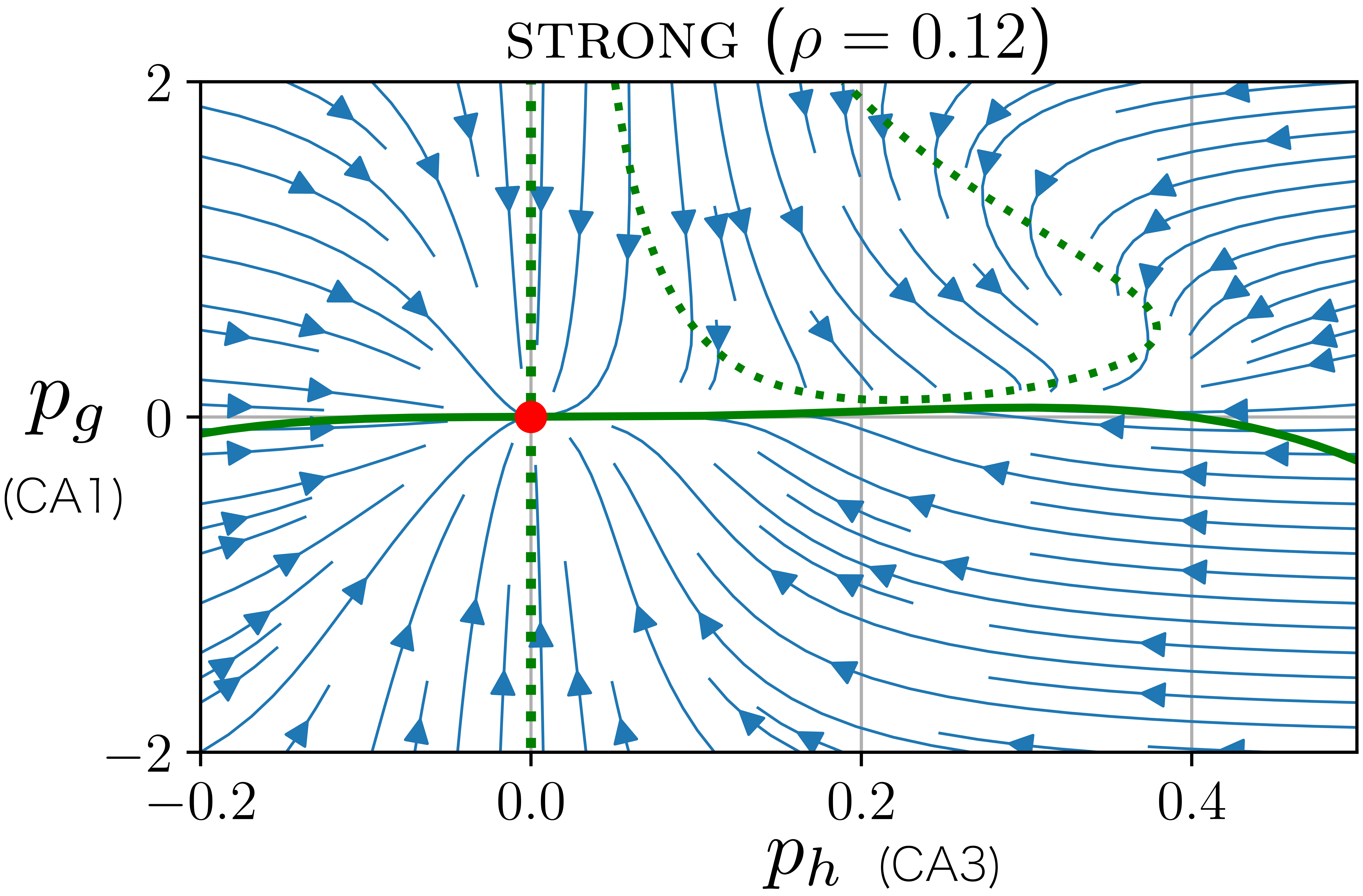
SimSiam dynamics



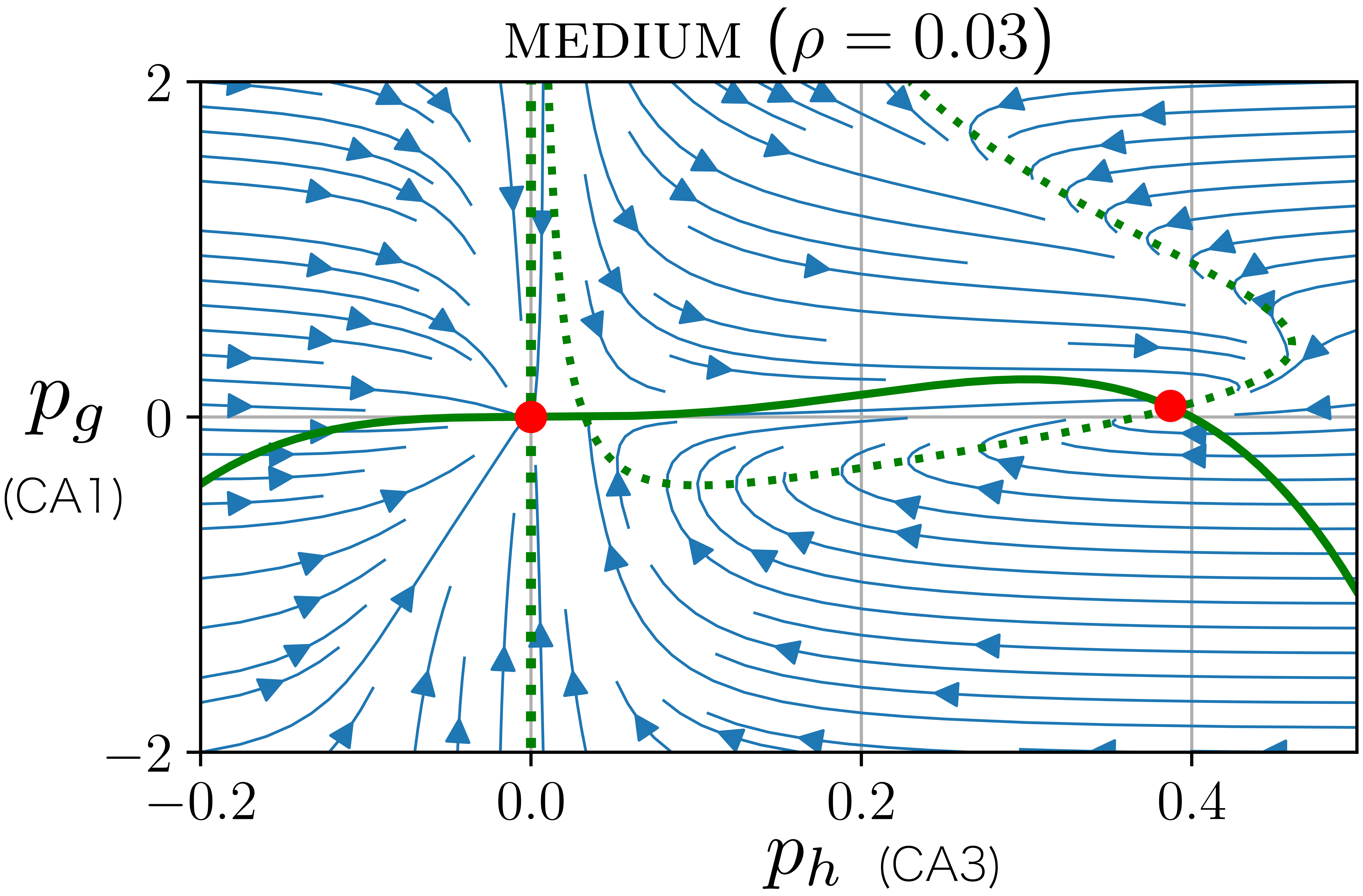
extract nullclines →



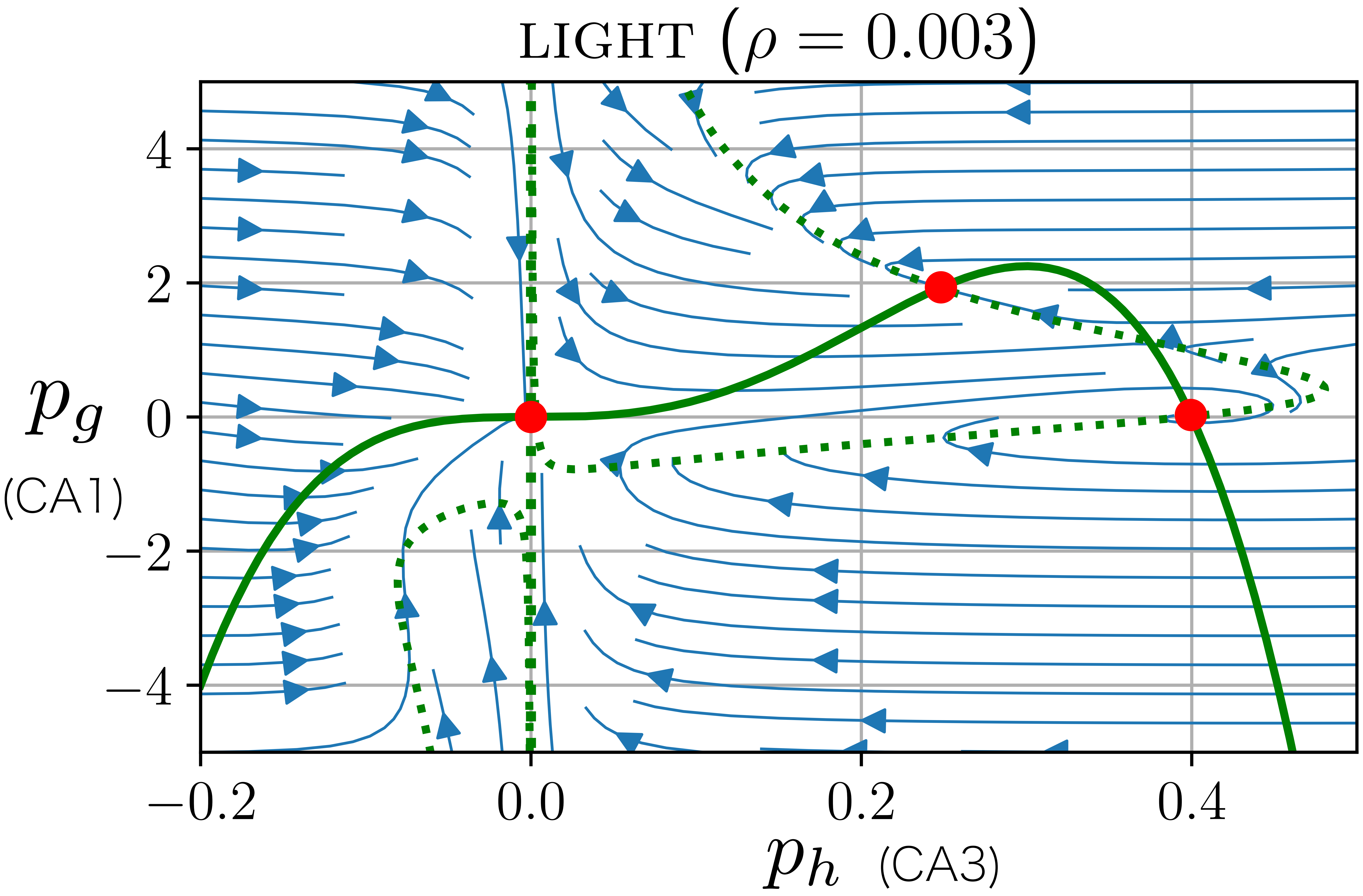
PhiNet dynamics with different weight decay



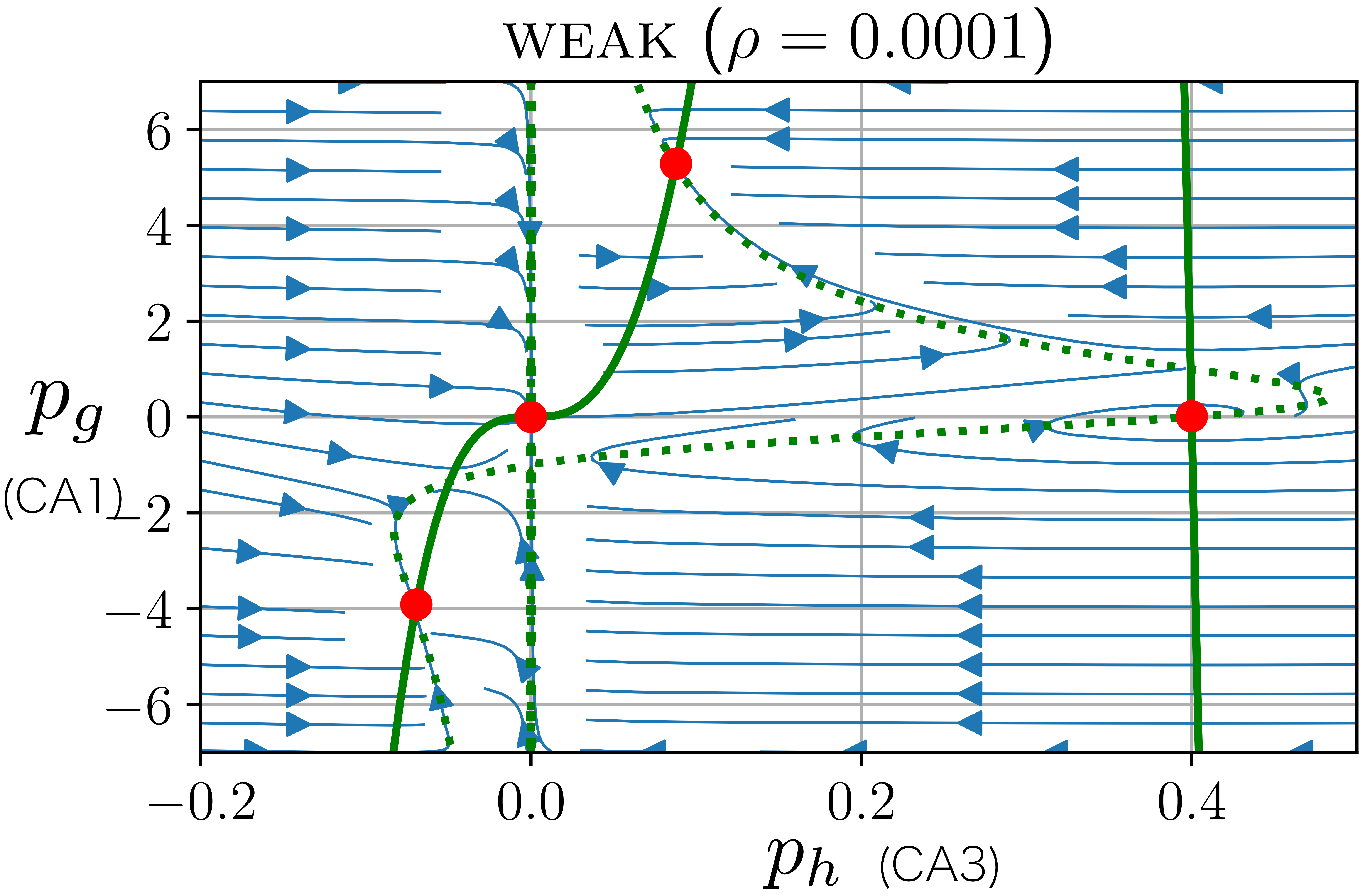
PhiNet dynamics with different weight decay



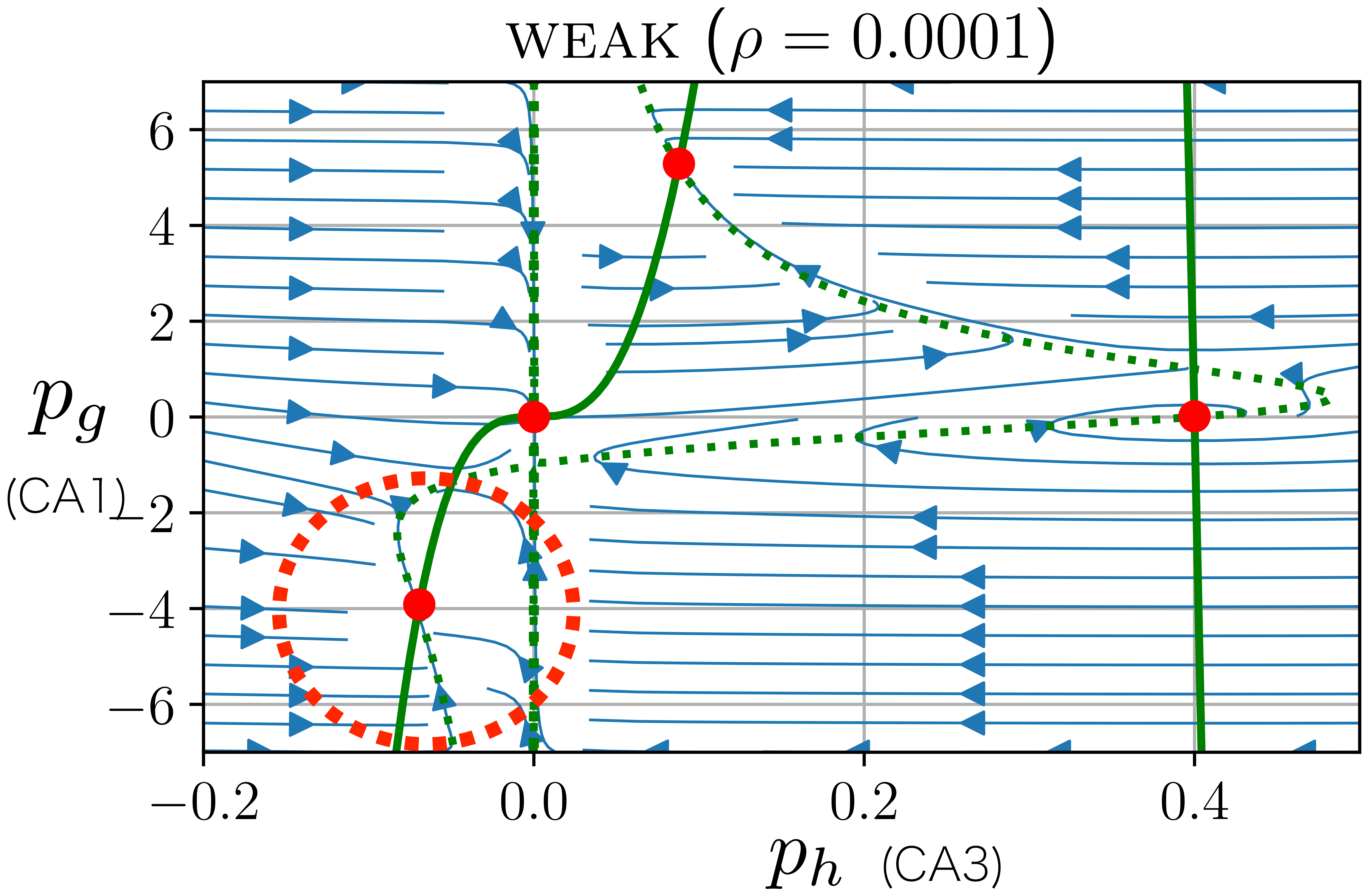
PhiNet dynamics with different weight decay



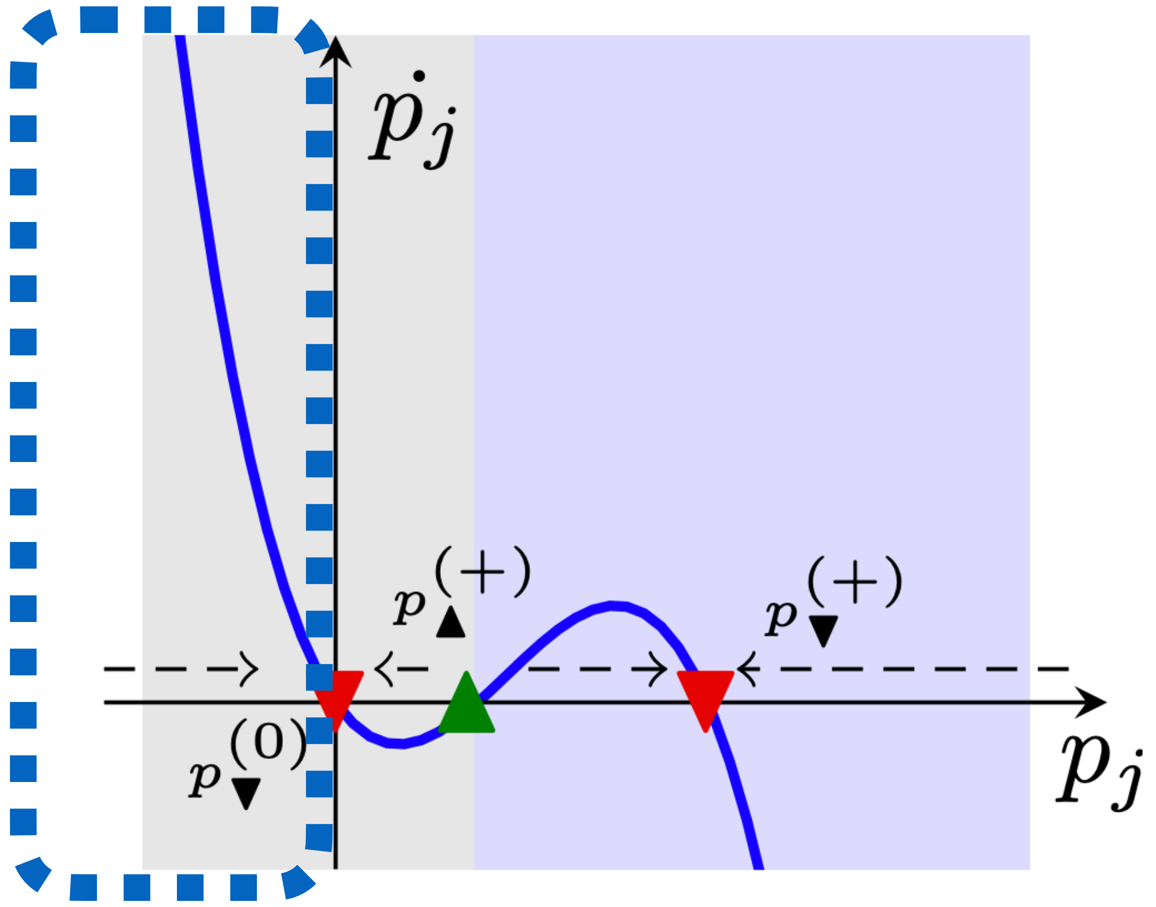
PhiNet dynamics with different weight decay



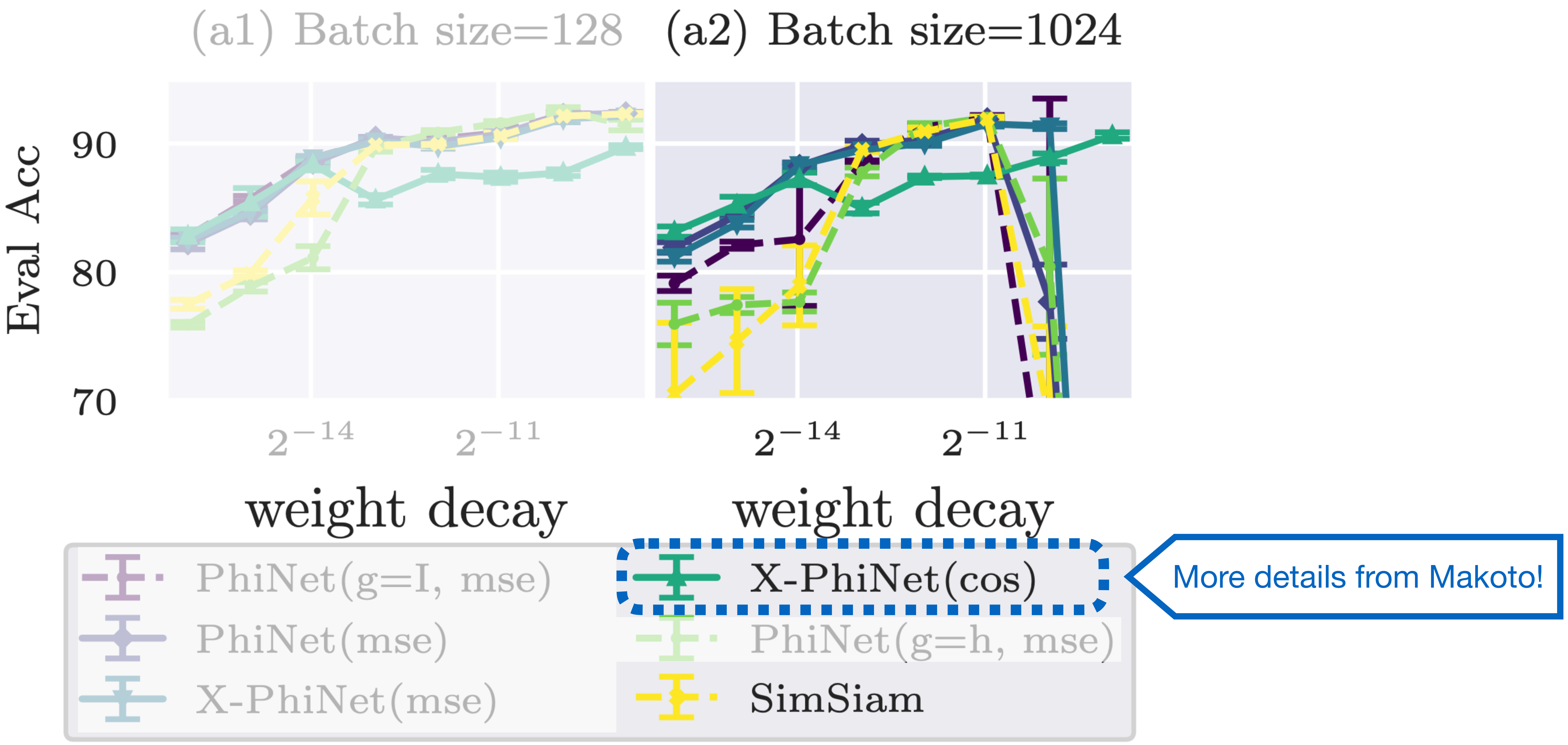
Negatively initialized eigval can converge non-trivially



SimSiam cannot avoid collapse if negatively initialized



Enhanced stability wrt weight decay



Dataset: CIFAR-10 / Evaluation: kNN accuracy

Summary

Interaction bw ML, nonlinear dynamics, neuroscience