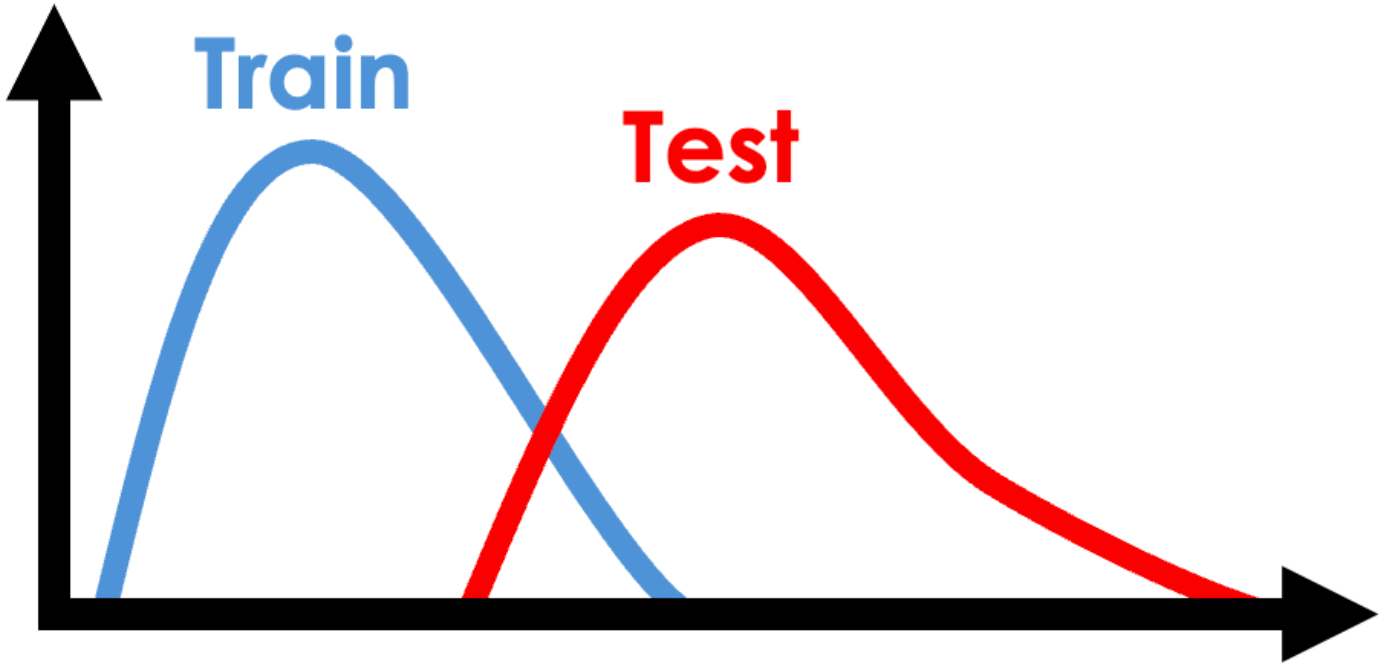
Logical Extrapolation on Mazes with Recurrent and Implicit Networks

AAAI-26

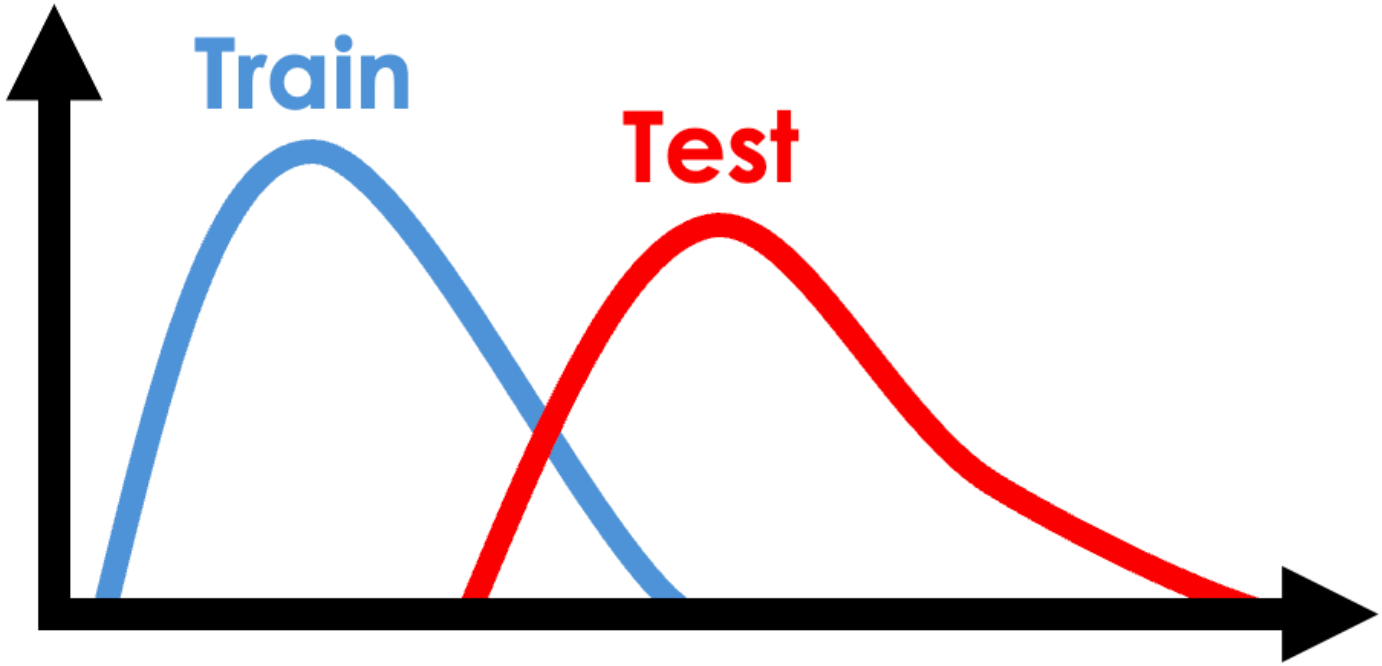
Brandon Knutson, Amandin Chyba Rabeendran, Michael Ivanitskiy, Jordan Pettyjohn, Cecilia Diniz Behn, Samy Wu Fung, Daniel McKenzie

COLORADO SCHOOL OF MINES

What is logical extrapolation?

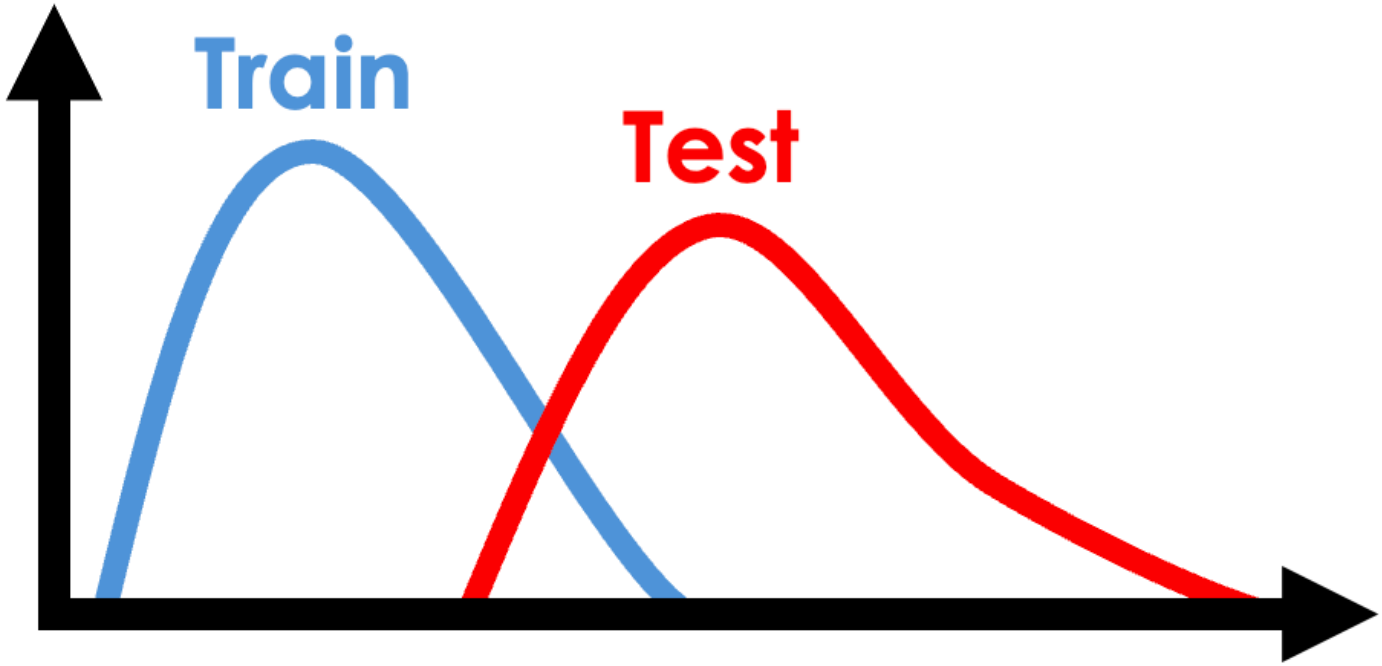


What is logical extrapolation?



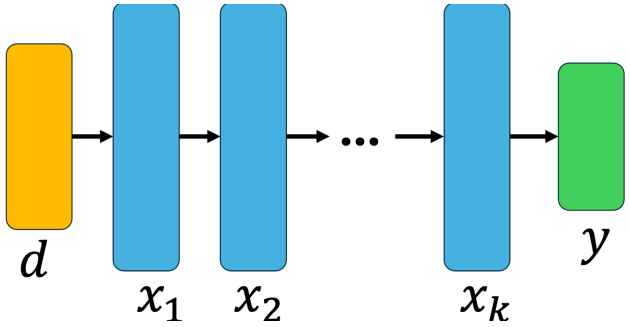
- **Out-of-distribution (OOD) generalization** is the ability for a model to perform well on test data distributed differently from training data.

What is logical extrapolation?

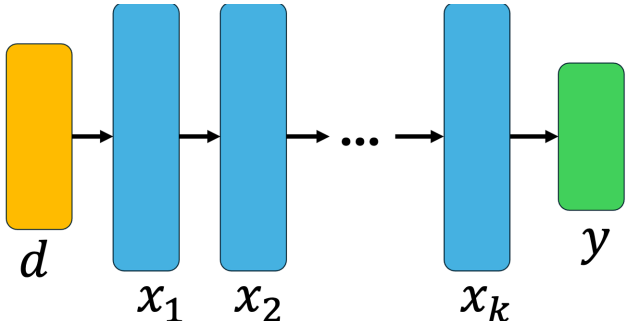


- **Out-of-distribution (OOD) generalization** is the ability for a model to perform well on test data distributed differently from training data.
- **Logical extrapolation** is OOD generalization to *harder* problems.

Recurrent and implicit neural networks



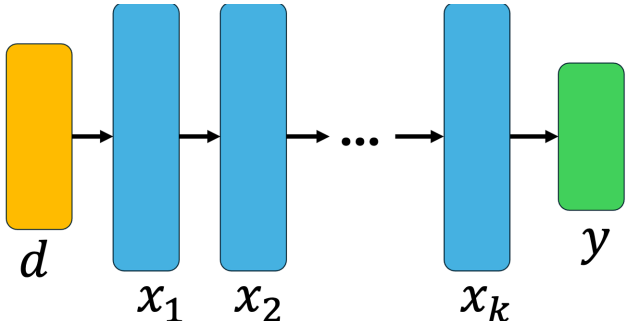
Recurrent and implicit neural networks



Recurrent Neural Network (RNN)

Choose x_0, depth k
for $j = 1, \ldots, k$ **do**
  $x_j \leftarrow F(x_{j-1}, \theta, d)$
end for
return x_k

Recurrent and implicit neural networks



Recurrent Neural Network (RNN)

Choose x_0, depth k
for $j = 1, \ldots, k$ **do**
  $x_j \leftarrow F(x_{j-1}, \theta, d)$
end for
return x_k

Implicit Neural Network (INN)

Choose x_0, max depth K, tolerance ε
for $k = 1, \ldots, K$ **do**
  $x_k \leftarrow \text{Solver}(x_{k-1}, \theta, d)$
  if $\|x_k - x_{k-1}\| < \varepsilon$ **then break**
end for
return x_k $(x_k \approx F(x_k, \theta, d))$

Research questions

Prior work

Can RNNs/INNs perform logical extrapolation?

Prior work

Can RNNs/INNs perform logical extrapolation?

- Yes: (Anil et al., 2022; Bansal et al., 2022)

Research questions

Prior work

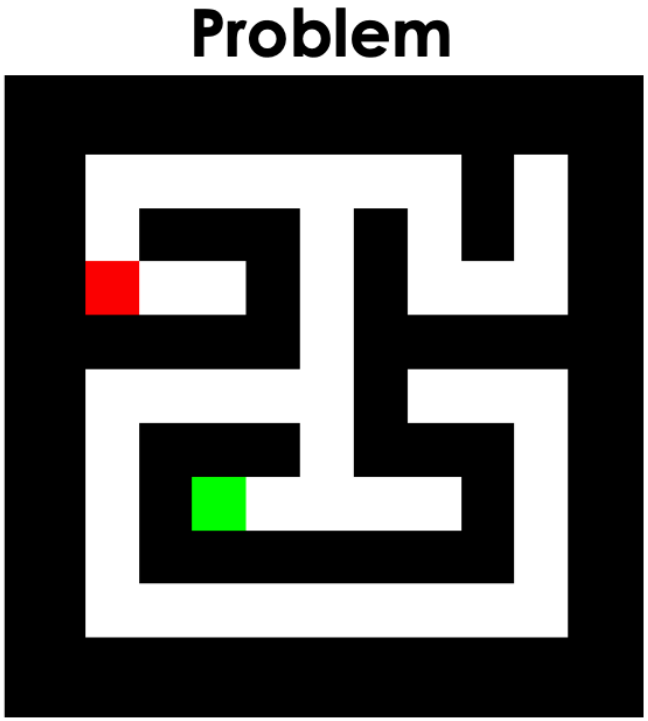
Can RNNs/INNs perform logical extrapolation?

- Yes: (Anil et al., 2022; Bansal et al., 2022)

This work

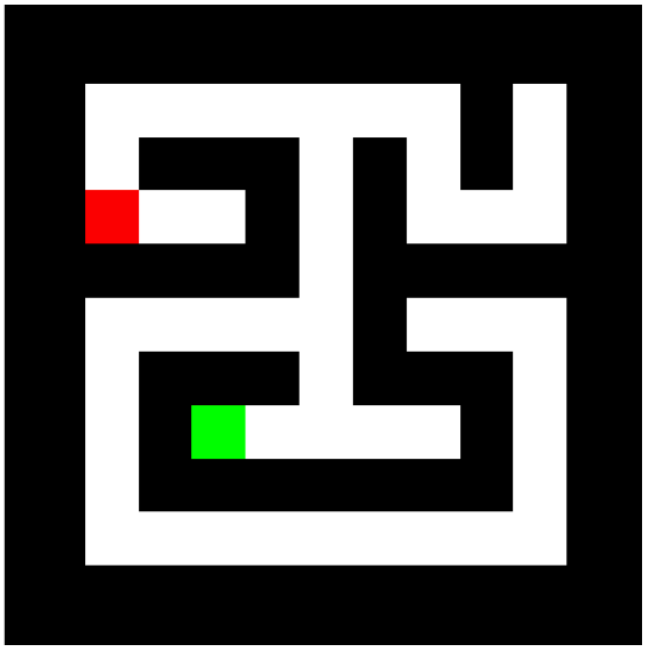
(1) **Do RNN/INNs learn a known algorithm?**
(2) **What do iterates converge to, and does it matter?**
(3) **How does training diversity affect extrapolation?**

Maze solving

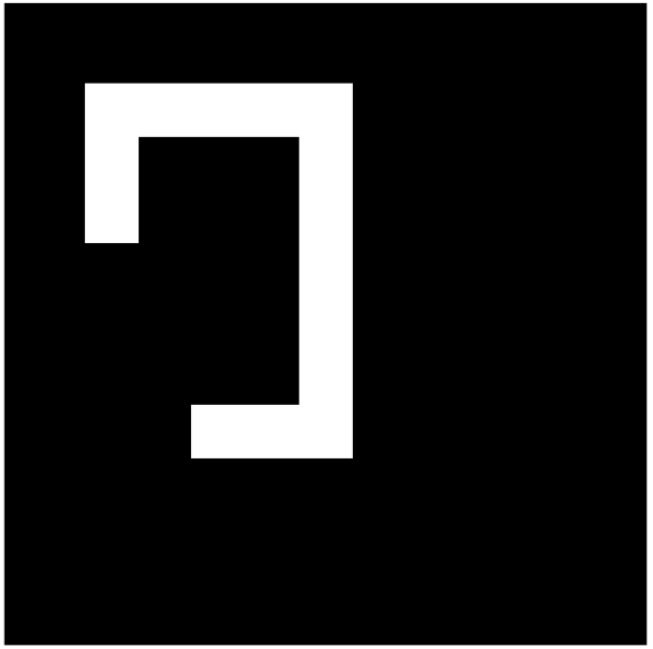


Maze solving

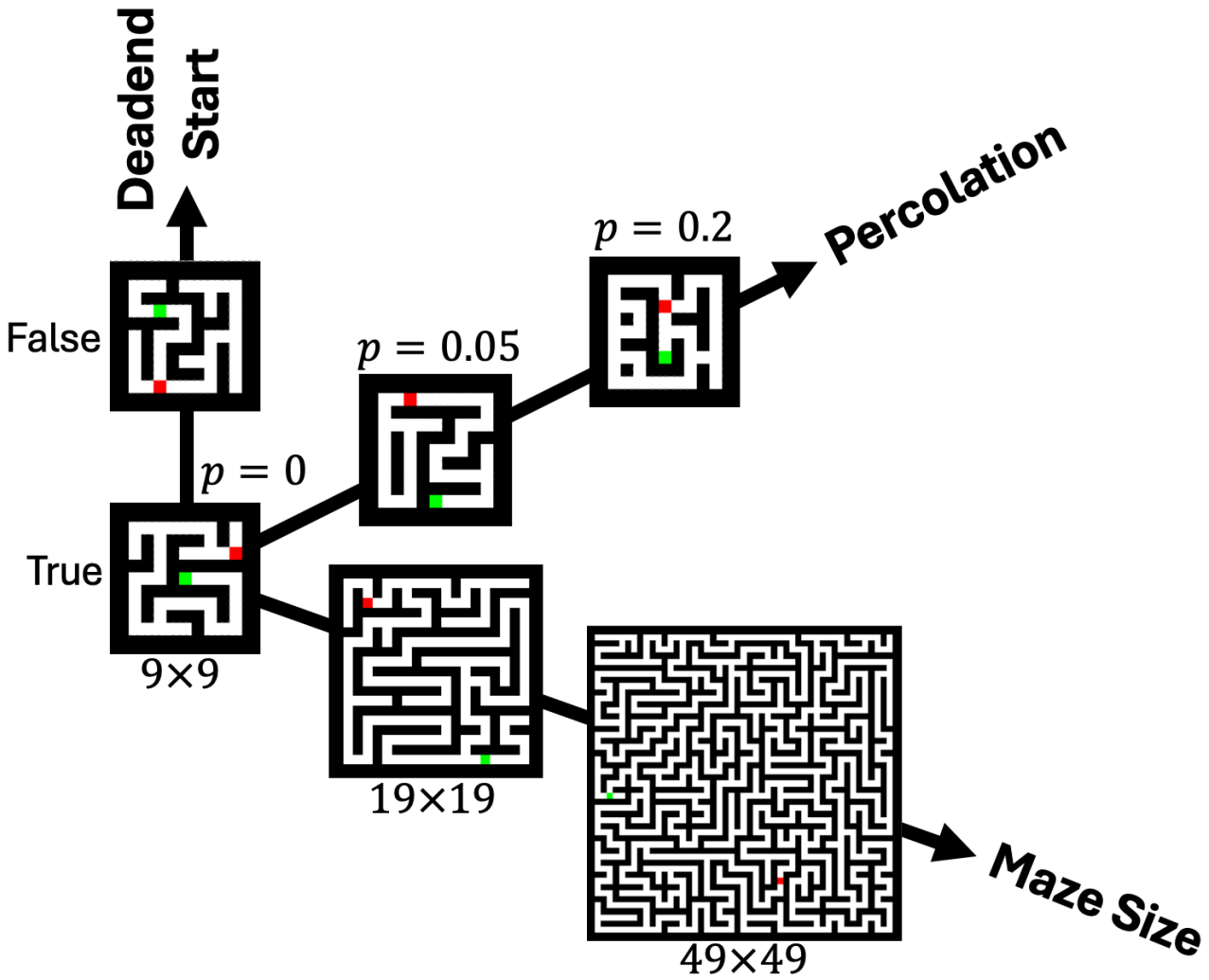
Problem



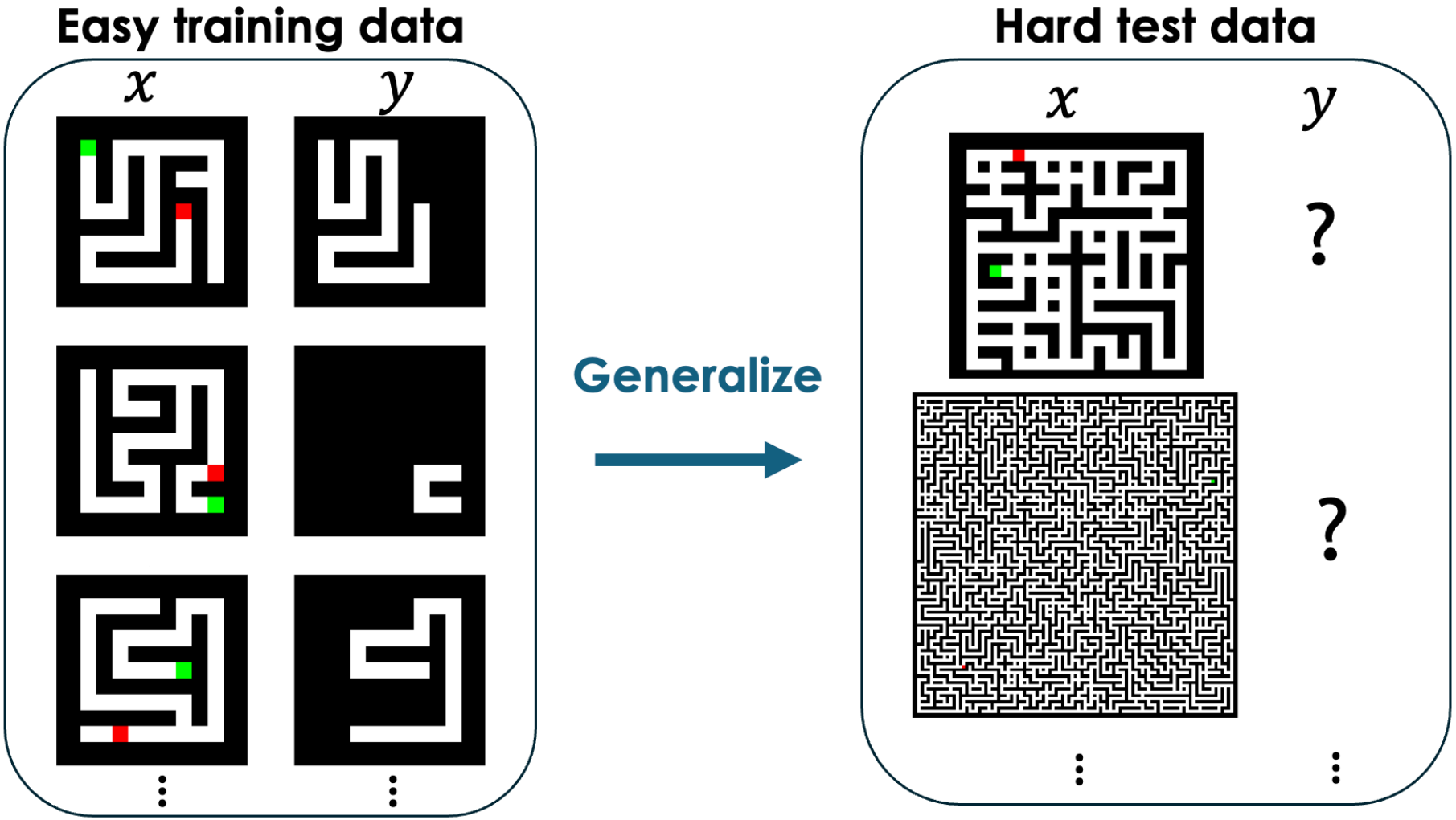
Solution



Out-of-distribution shifts



Extrapolating with mazes



Do RNN/INNs learn a known algorithm?

Do RNN/INNs learn a known algorithm?

- RNN approximately learns the “deadend-filling” algorithm

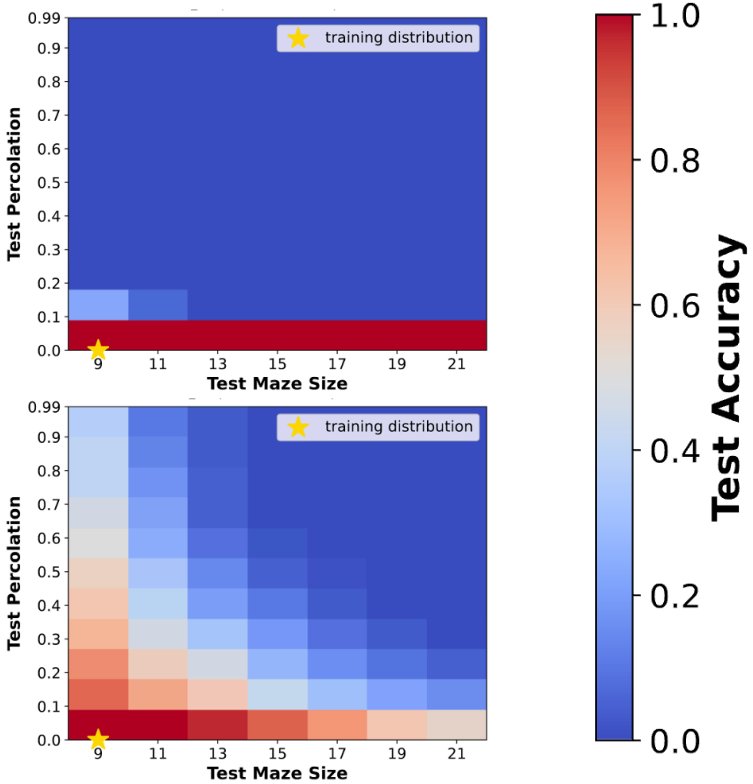
Do RNN/INNs learn a known algorithm?

- RNN approximately learns the “deadend-filling” algorithm
- Models fail on mazes with loops (goal misgeneralization)

Train Percolation: 0.000

RNN:

INN



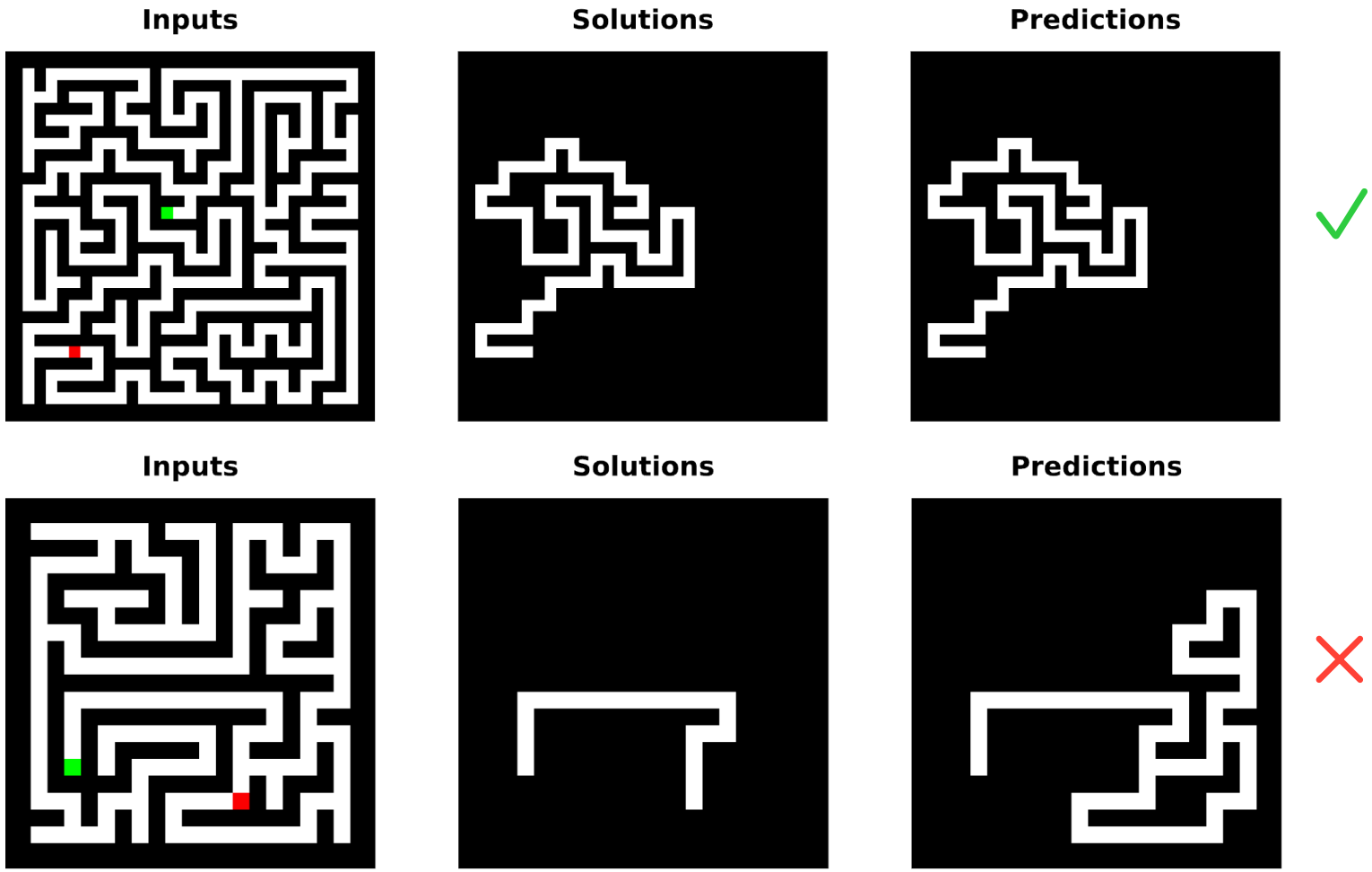
→: larger mazes

↑: more loops

★: training distribution

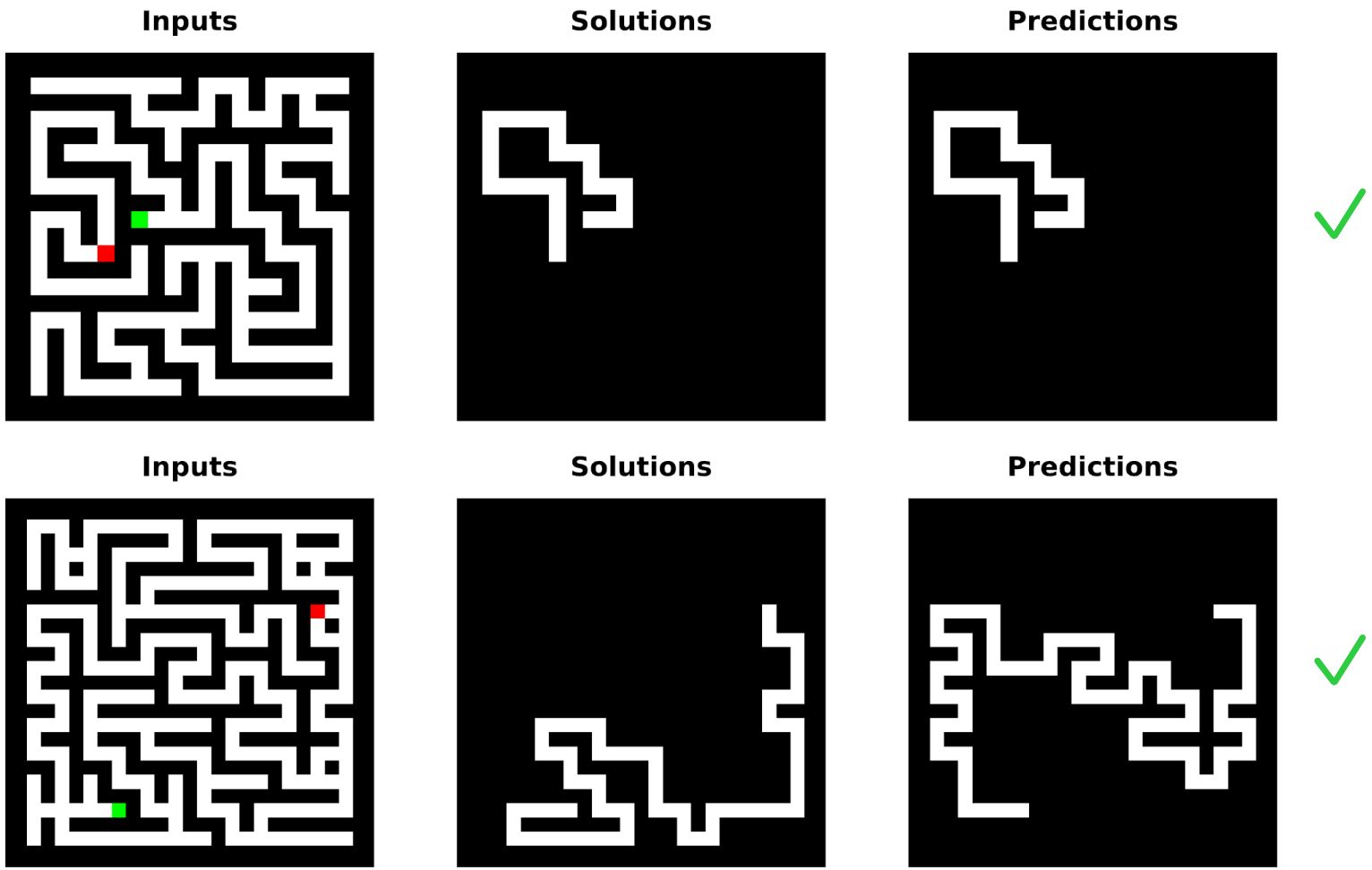
RNN predictions

The RNN solves much larger mazes but fails to solve mazes with loops.



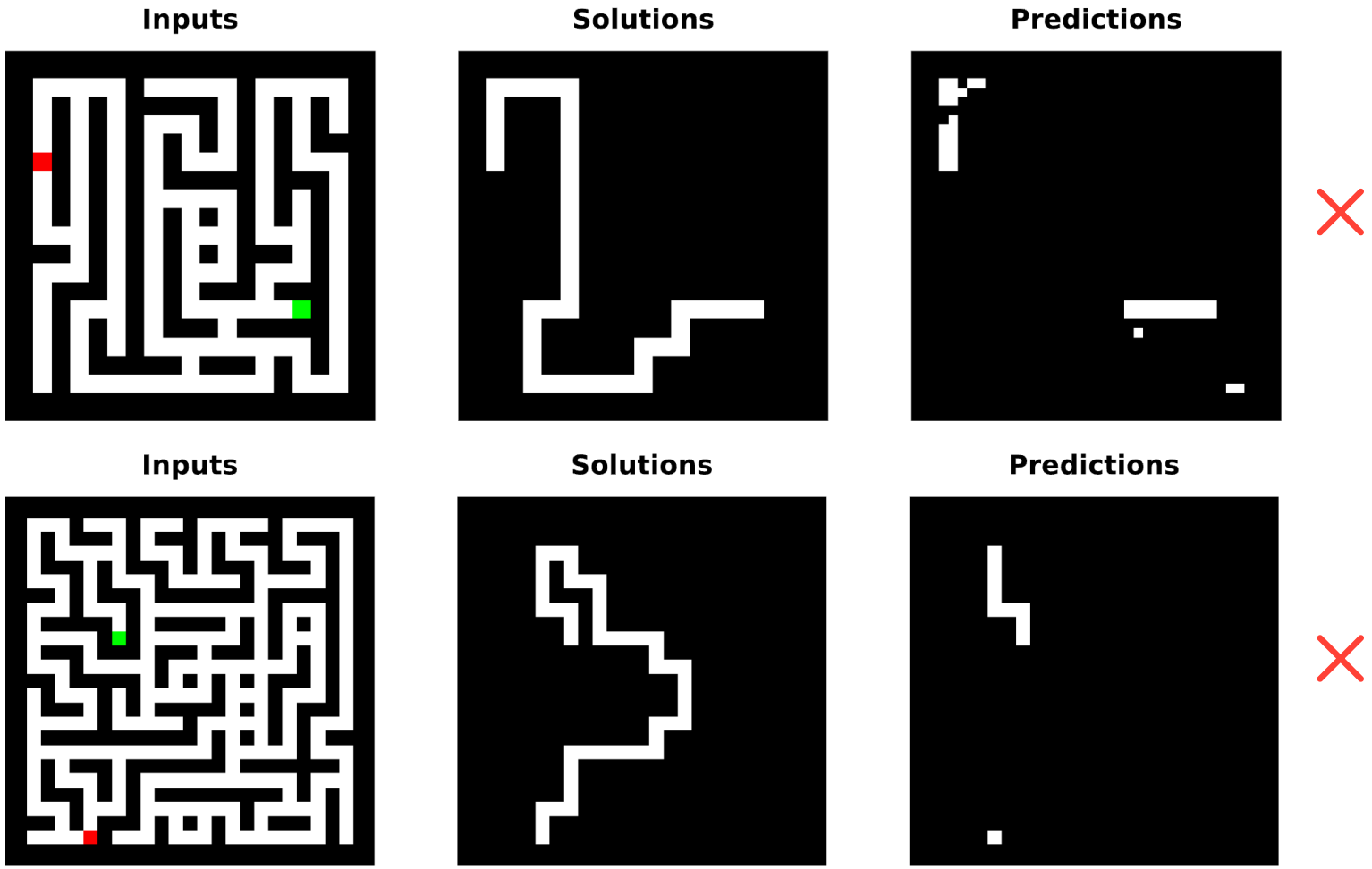
INN predictions

The INN can solve larger mazes with loops, and sometimes finds alternate paths.

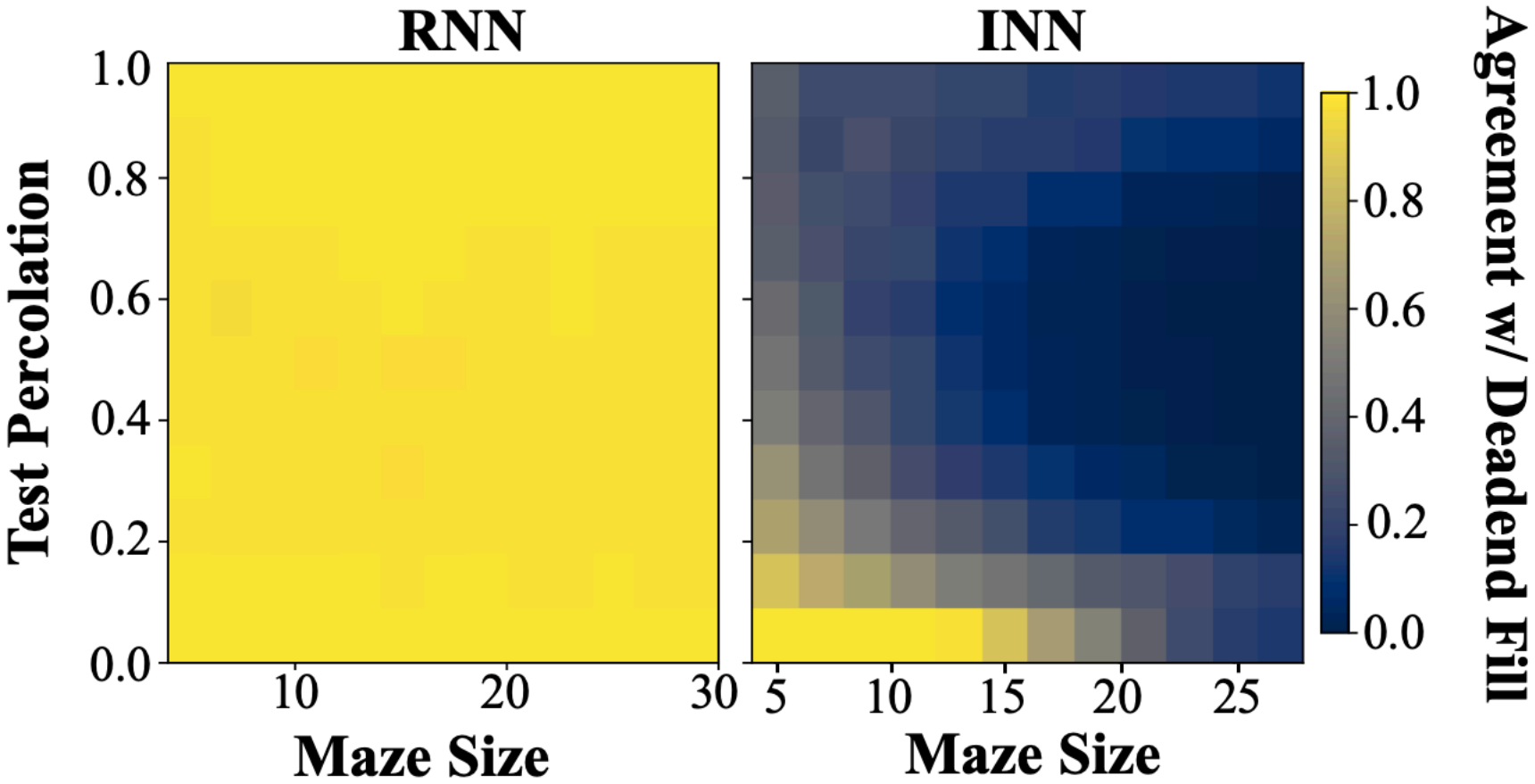


INN predictions

But like the RNN, the INN fails on many mazes.

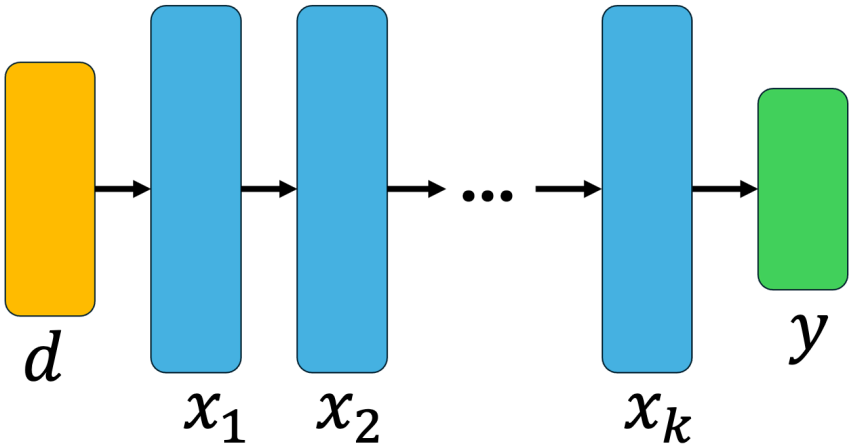


Agreement with deadend filling algorithm

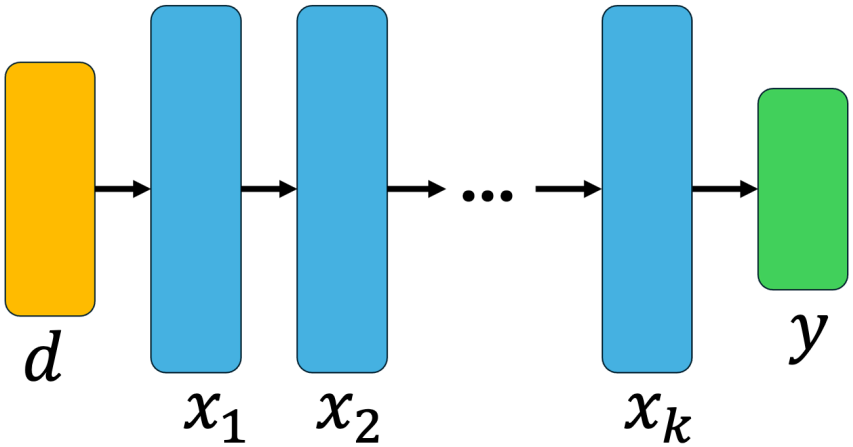


What do iterates converge to, and does it matter?

- **Prior hypothesis**: iterates $\{x_1, \ldots, x_k\} \subseteq \mathbb{R}^n$ converge to a fixed point

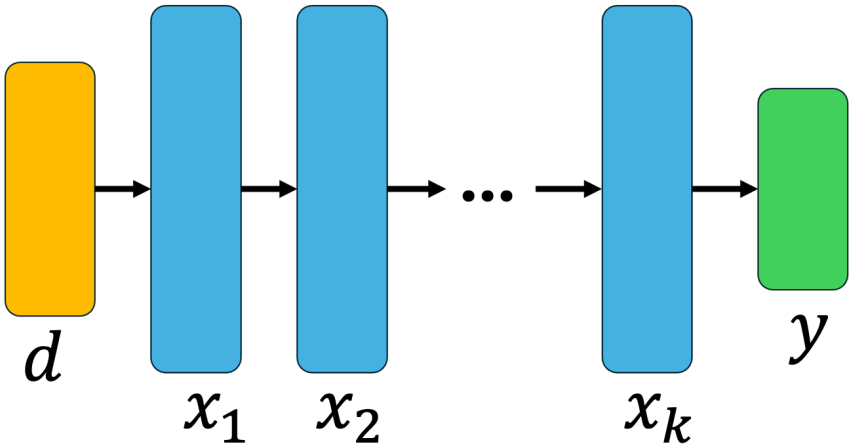


What do iterates converge to, and does it matter?



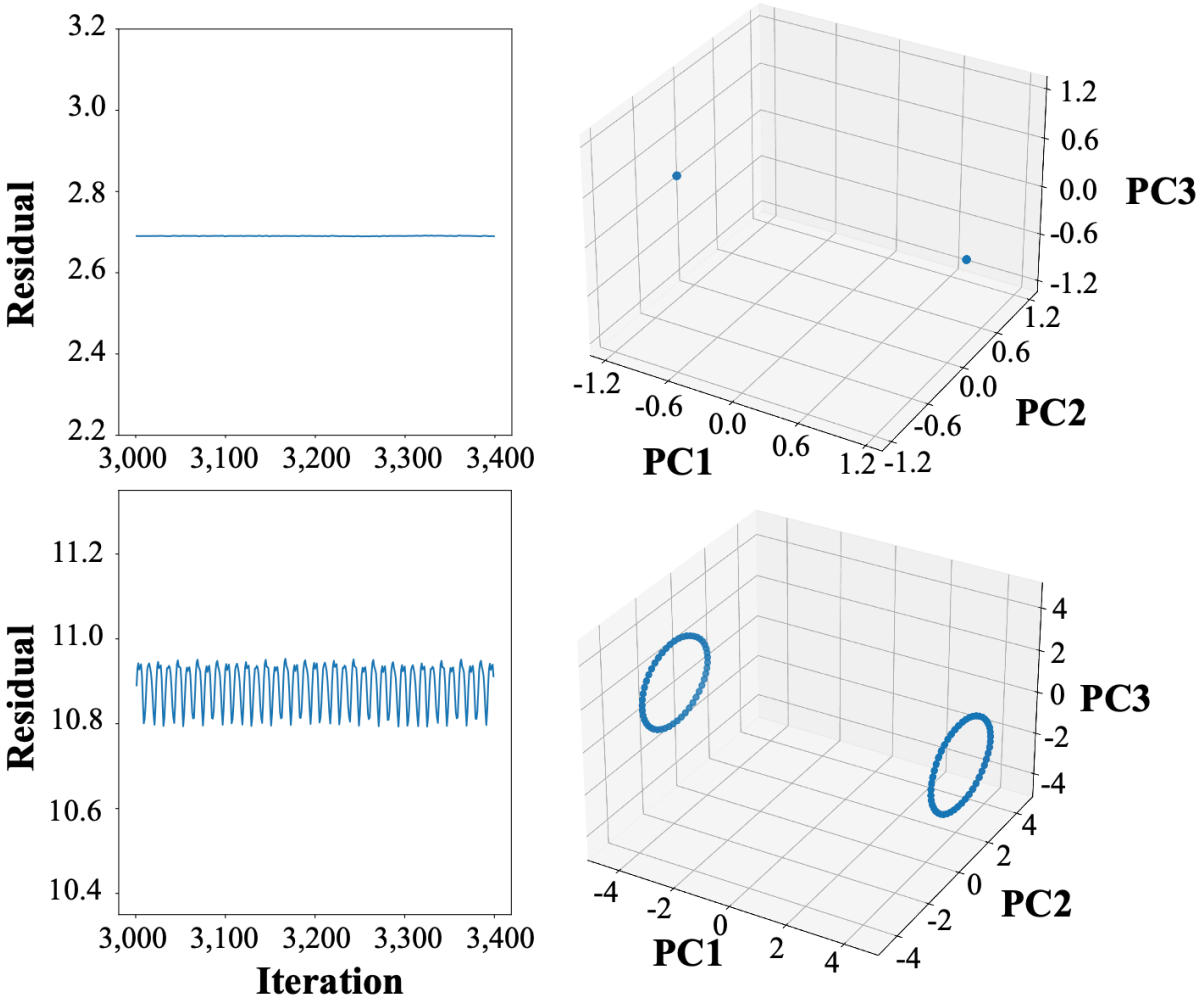
- **Prior hypothesis**: iterates $\{x_1, ..., x_k\} \subseteq \mathbb{R}^n$ converge to a fixed point
- We use **Topological Data Analysis (TDA)** to characterize limiting behavior

What do iterates converge to, and does it matter?



- **Prior hypothesis**: iterates $\{x_1, ..., x_k\} \subseteq \mathbb{R}^n$ converge to a fixed point
- We use **Topological Data Analysis (TDA)** to characterize limiting behavior
- Found four limiting behaviors:
  - Fixed point: $[B_0, B_1] = [1, 0]$
  - Two-point cycle: $[B_0, B_1] = [2, 0]$
  - One-loop cycle: $[B_0, B_1] = [1, 1]$
  - Two-loop cycle: $[B_0, B_1] = [2, 2]$

Iterate dynamics



TDA results

- **INN**: always converges to fixed point

TDA results

- **INN**: always converges to fixed point
- **RNN**: also converges to two-point, one-loop, and two-loop cycles

TDA results

- **INN**: always converges to fixed point
- **RNN**: also converges to two-point, one-loop, and two-loop cycles
- RNN extrapolates to larger mazes despite converging to cycles

TDA results

- **INN**: always converges to fixed point
- **RNN**: also converges to two-point, one-loop, and two-loop cycles
- RNN extrapolates to larger mazes despite converging to cycles
- No obvious correlation between limiting behavior and accuracy

TDA results

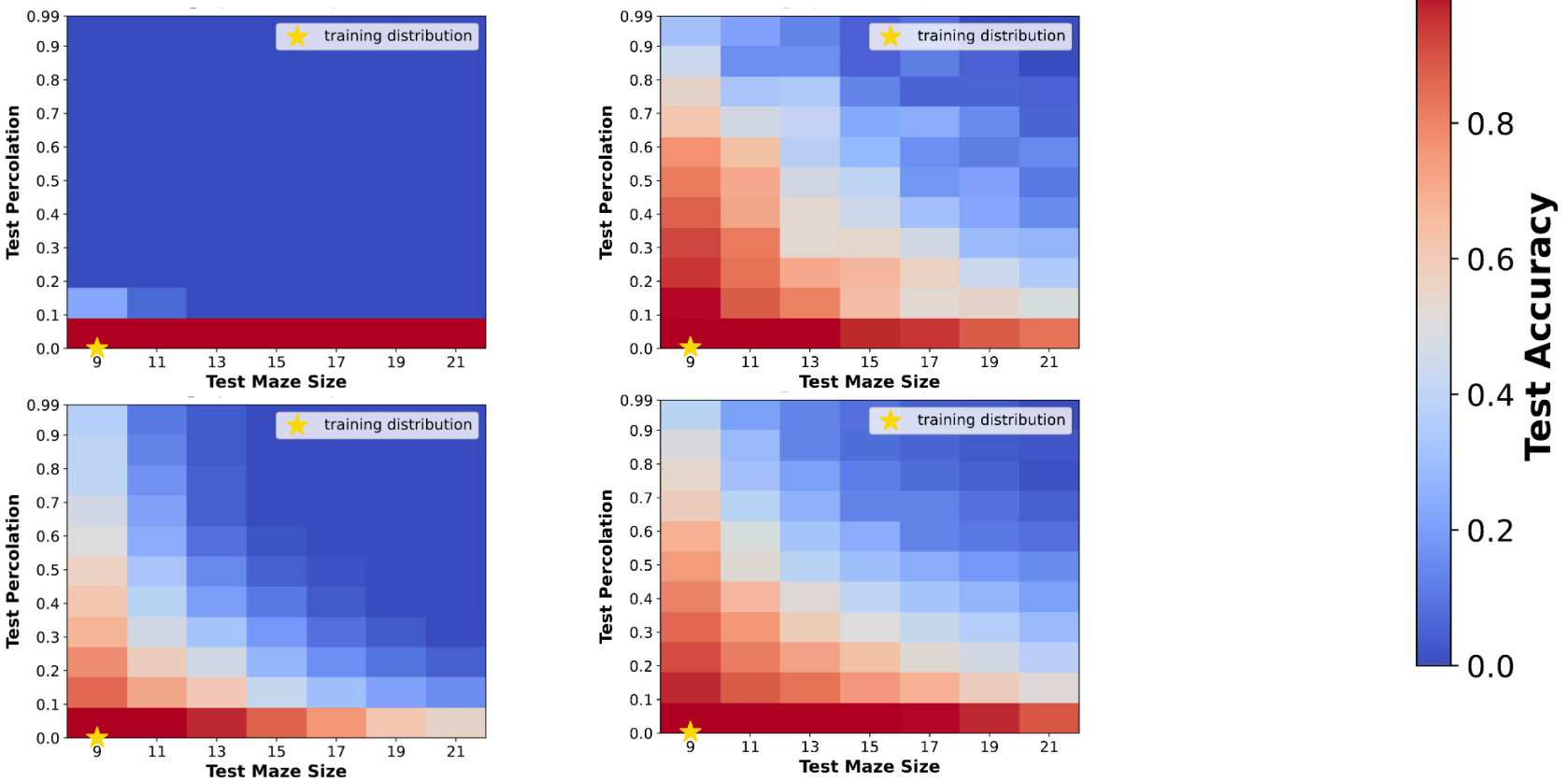
- **INN**: always converges to fixed point
- **RNN**: also converges to two-point, one-loop, and two-loop cycles
- RNN extrapolates to larger mazes despite converging to cycles
- No obvious correlation between limiting behavior and accuracy
- Fixed-point convergence is not necessary for extrapolation.

How does training diversity affect extrapolation?

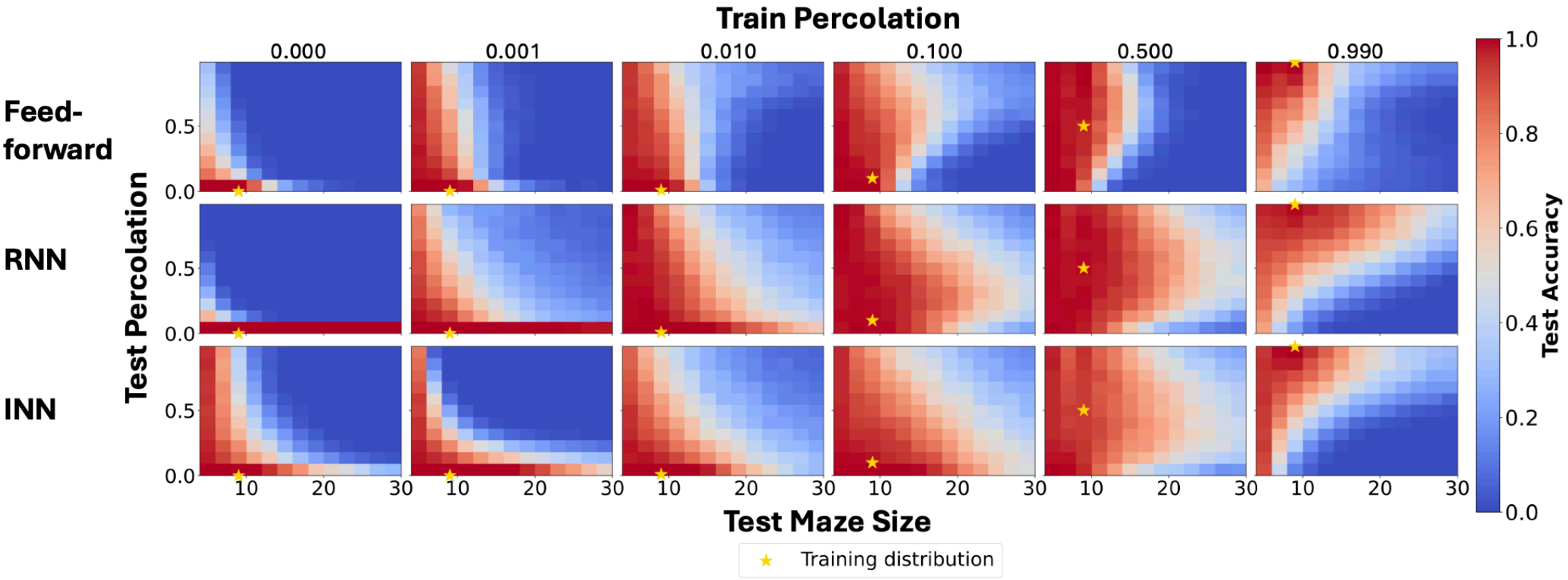
Train Percolation: **0.000** **0.003**

RNN:

INN



Testing extrapolation



Thank you!

(1) **Do RNN/INNs learn a known algorithm?**
- Partially. An RNN approximately learns deadend-filling.

(2) **What do iterates converge to, and does it matter?**
- Fixed points or cycles. No, it doesn't seem to matter.

(3) **How does training diversity affect extrapolation?**
- Slight diversification dramatically improves performance on mazes with loops. But diversified models don't appear algorithmic, and extrapolation to larger mazes diminishes.

paper

maze dataset package

This work was supported by the National Science Foundation under Grant DMS-2309810.

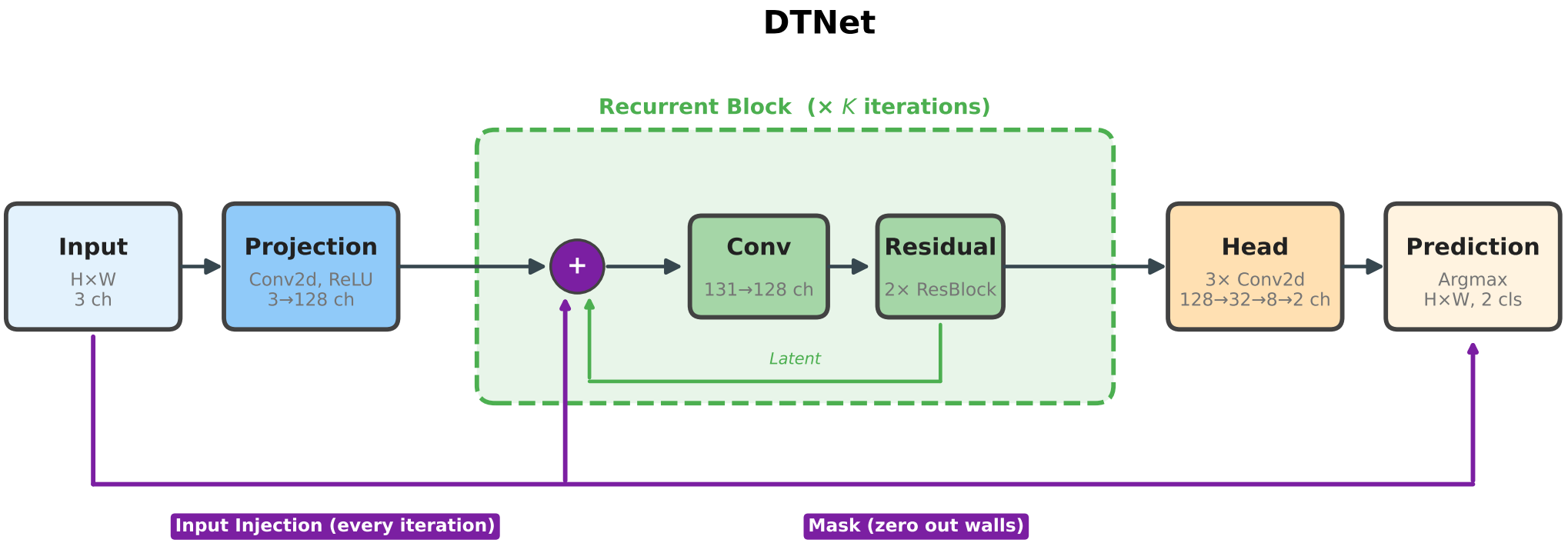
Bibliography

Anil, C., Pokle, A., Liang, K., Treutlein, J., Wu, Y., Bai, S., Kolter, J. Z., & Grosse, R. B. (2022). Path Independent Equilibrium Models Can Better Exploit Test-Time Computation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems: Vol. 35. Advances in Neural Information Processing Systems.*

Bansal, A., Schwarzschild, A., Borgnia, E., Emam, Z., Huang, F., Goldblum, M., & Goldstein, T. (2022). End-to-end Algorithm Synthesis with Recurrent Networks: Extrapolation without Overthinking. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems: Vol. 35. Advances in Neural Information Processing Systems.*

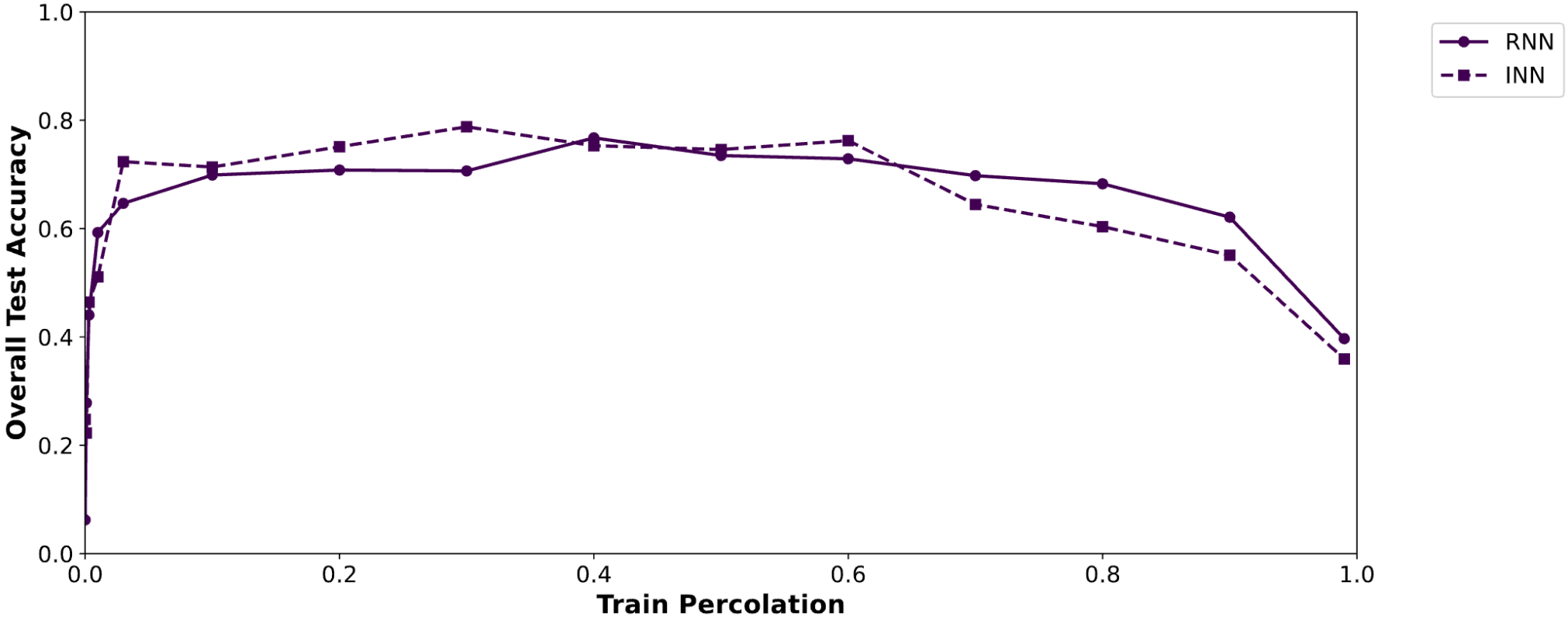
Appendix

Architecture diagram



Test accuracy (appendix)

Overall extrapolation



Betti number frequencies

Model	$[B_0, B_1]$	9	19	29	39	49	59
DT-Net	$[1,0]^*$	15	0	0	0	0	0
	$[1,1]$	6	6	0	0	0	0
	$[2,0]$	79	88	95	100	98	98
	$[2,1]$	0	1	0	0	0	0
	$[2,2]$	0	5	5	0	2	2
PI-Net	$[1,0]^*$	100	100	100	100	100	100
IT-Net	$[1,0]^*$	100	100	100	100	100	100

Frequencies (%) across maze sizes $n \times n$. DT-Net usually converges to a two-point cycle ($[2,0]$), while PI-Net and IT-Net always converge to a fixed point ($[1,0]$). $*$Converged to within 0.01.