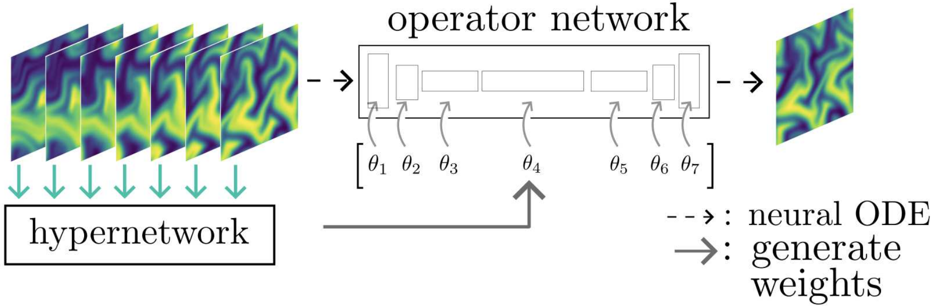
Learning Evolution Operators Across PDE Systems: Meta-Learning and Test-Time Generalization

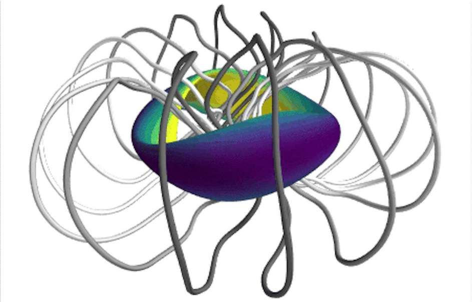
Jiequn Han
Center of Computational Mathematics
Flatiron Institute, Simons Foundation



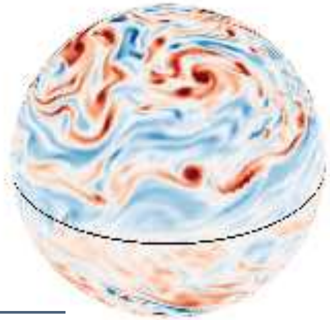
IPAM
April 2026

Machine Learning Models for Physics

Expensive simulations

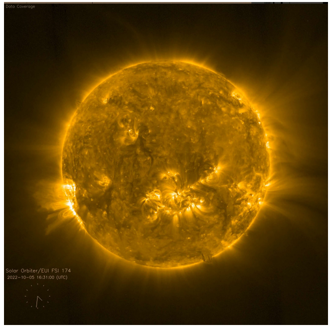


Weather forecast

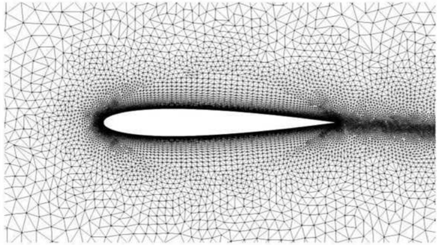


Various
ML Models

Unknown physics



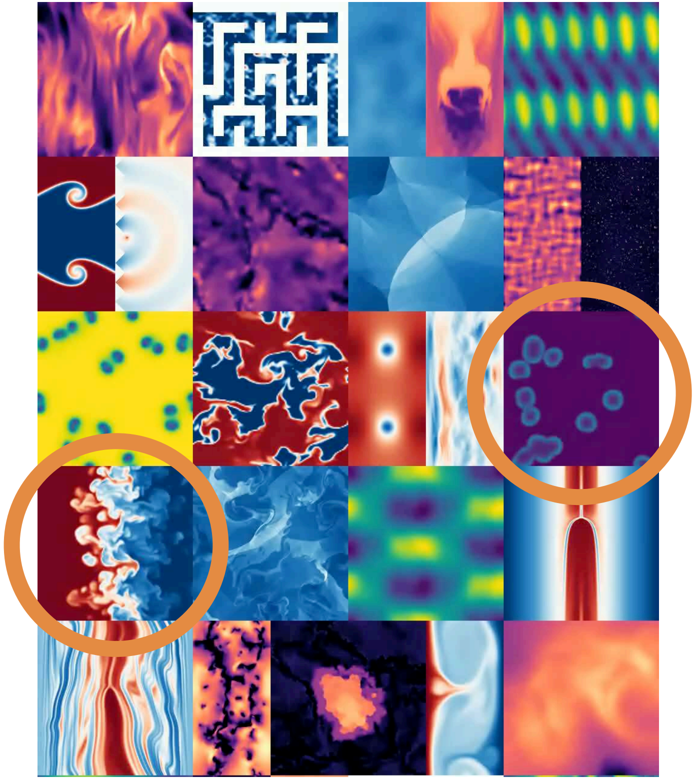
Non-uniform grids



fast surrogate for inference, design,
and control

Why training from scratch
every time?!

Data Variability

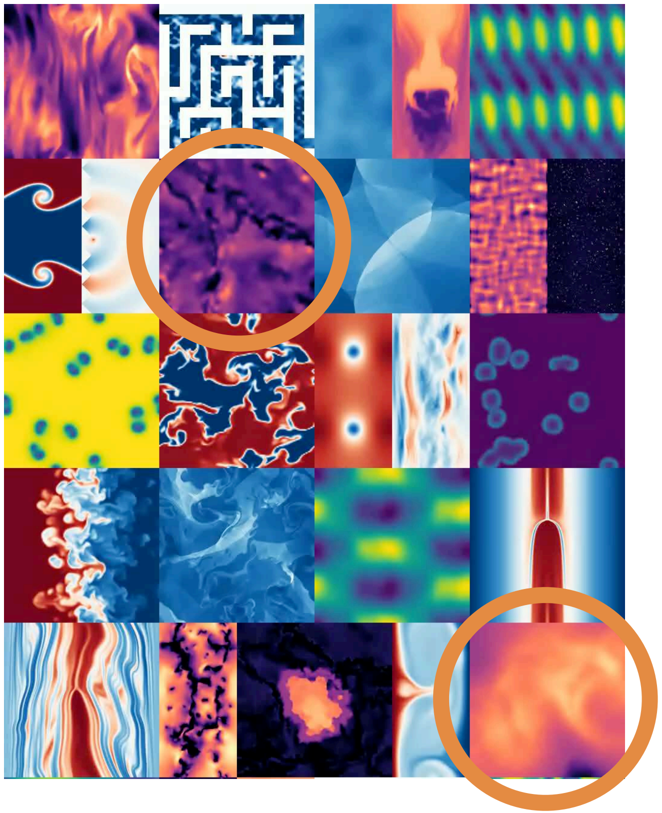


PDE class lvl 1

PDE

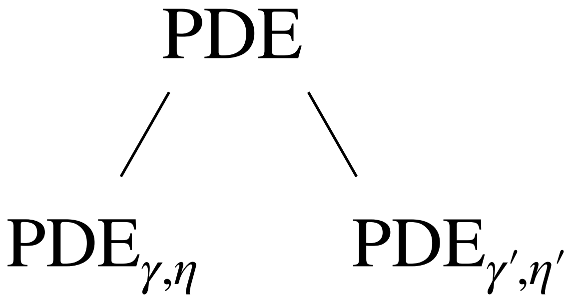
The Well, R. Ohana, et al. lead by Polymathic, 2024*

Data Variability



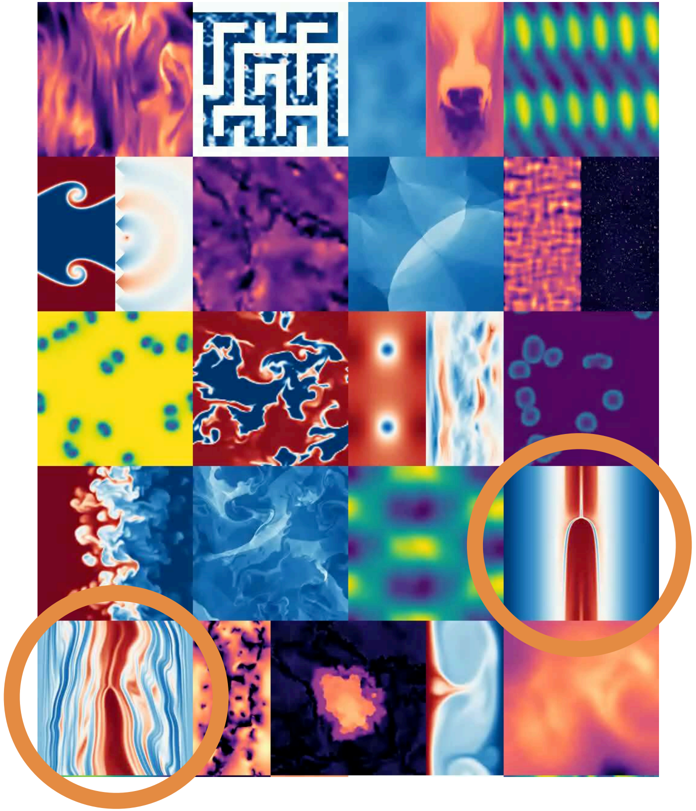
PDE class | lvl 1

parameter geometry | lvl 2



The Well, R. Ohana, et al. lead by Polymathic, 2024*

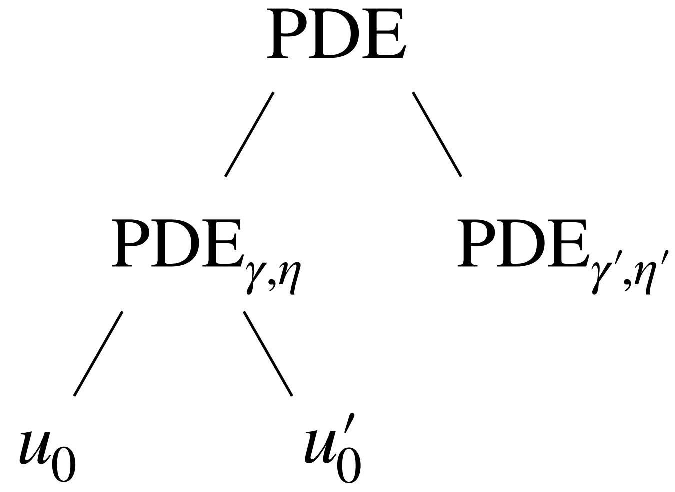
Data Variability



PDE class — lvl 1

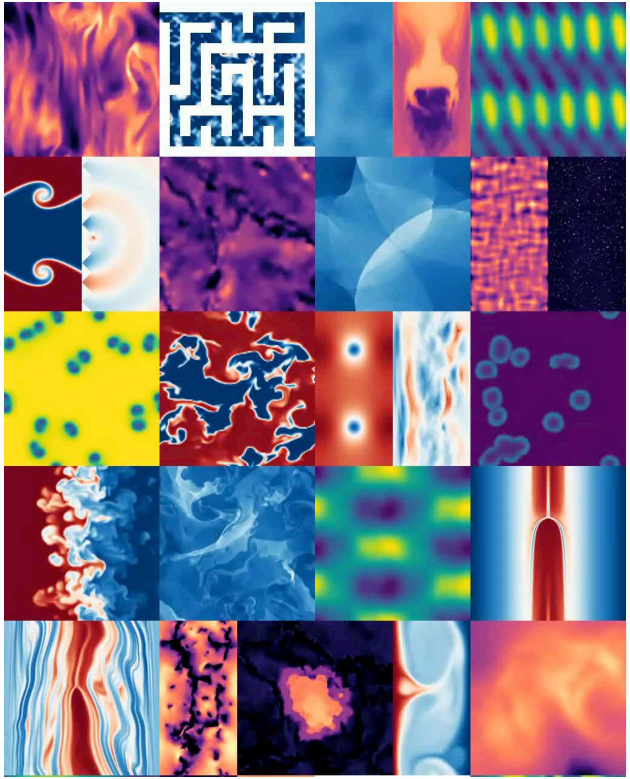
parameter geometry — lvl 2

init cond — lvl 3



The Well, R. Ohana, et al. lead by Polymathic, 2024*

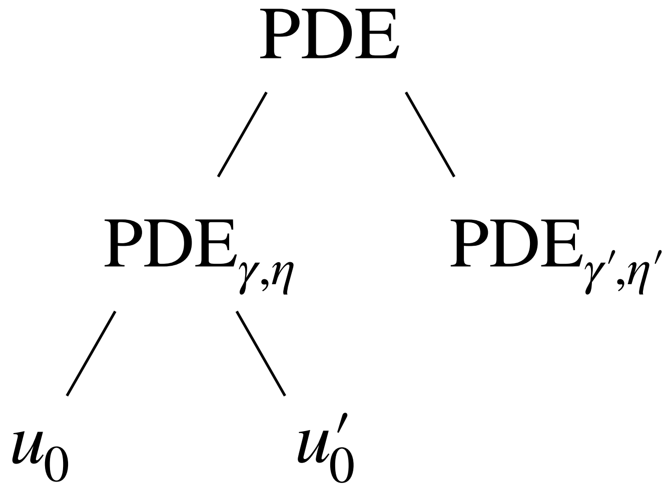
Data Variability



PDE class — lvl 1

parameter geometry — lvl 2

init cond — lvl 3

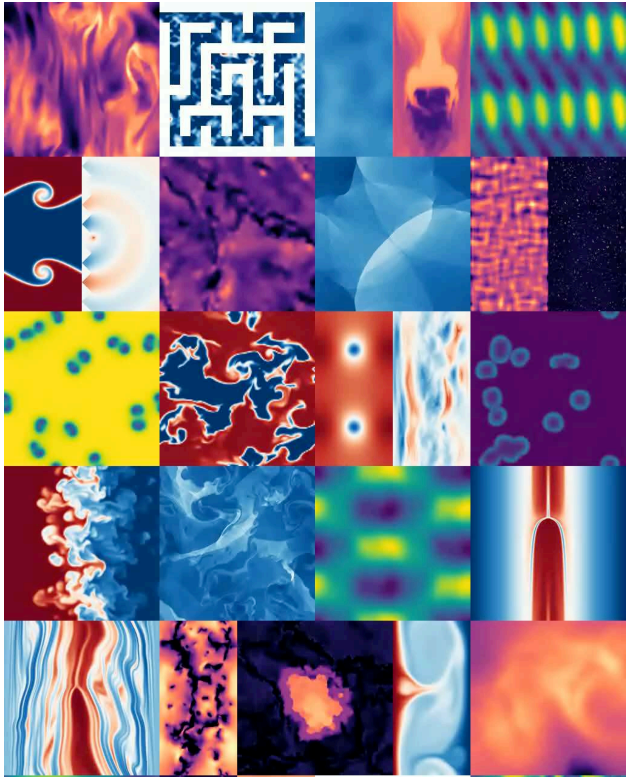


neural operator

Insufficient for complex physical phenomenon

The Well, R. Ohana, et al. lead by Polymathic, 2024*

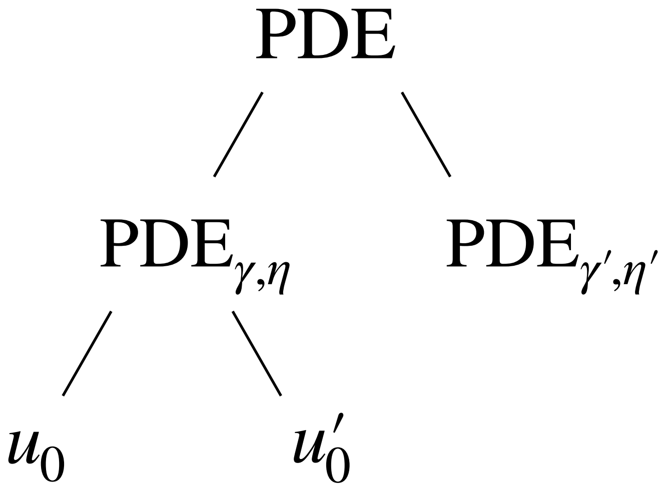
Data Variability



PDE class — lvl 1

parameter geometry — lvl 2

init cond — lvl 3



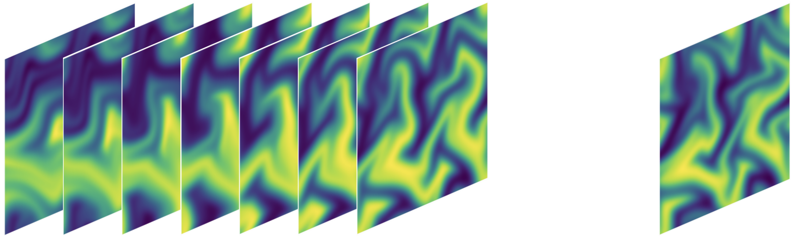
multi-physics-agnostic prediction

We need various (multi-fidelity) physics for approximating the complex physics

The Well, R. Ohana, et al. lead by Polymathic, 2024*

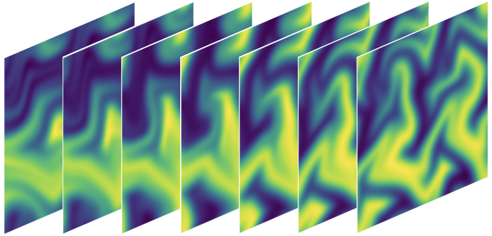
Multi-Physics-Agnostic Prediction

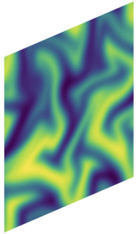
Task: From a context of T states $(u_{t-T+1}, \ldots, u_t)$ predict the next state u_{t+1}



Multi-Physics-Agnostic Prediction

Task: From a context of T states $(u_{t-T+1}, \ldots, u_t)$ predict the next state u_{t+1}





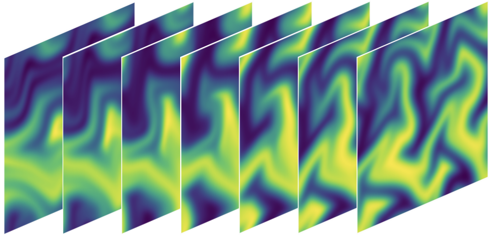
task introduced in [Multiple Physics Pretraining, McCabe et al., Polymathic, 2024]

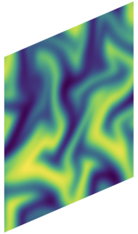
Transformer Neural Network

just like next token prediction

Difference between Physics and Language

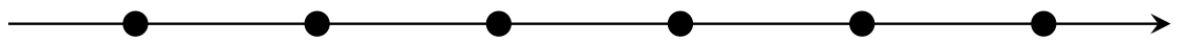
Task: From a context of T states $(u_{t-T+1}, \ldots, u_t)$ predict the next state u_{t+1}





inherently **continuous** in time

inherently **discrete** in time



How to Solve the Continuous-Time Physics?

$$\partial_t u_t(x) = f\left(u_t(x), \nabla u_t(x), \nabla^2 u_t(x)\right)$$

Numerical solver

time discretization

$$u_{t+\Delta t} \approx u_t + f\left(u_t, \nabla u_t, \nabla^2 u_t\right) \Delta t$$

space discretization

$$u_{t+\Delta t} \approx u_t + f_\theta\left(u_t\right) \Delta t \quad \left(e.g.\ u'(x) \approx \frac{u(x+\Delta x) - u(x-\Delta x)}{2\Delta x}\right)$$

Neural solver ⟶ learn θ from data

[Neural ODE, Chen et al., 2018]

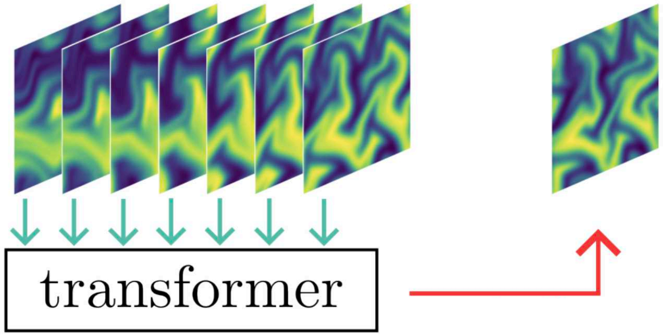
[Bar-Sinai, Hoyer et al. 2019]

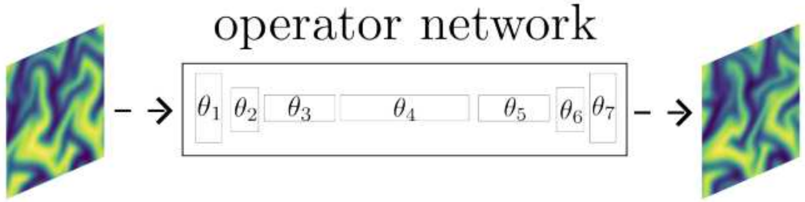
[Brandstetter, Worrall, Welling, 2022]

Task: From a context of T states $(u_{t-T+1}, \ldots, u_t)$ predict the next state u_{t+1}

Two subtasks: (1) infer the operator (2) apply the operator to evolve the state

How to Solve the Continuous-Time Physics?





pure transformer mixes the two

relatively easy for neural operator

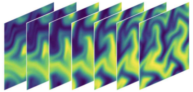
Task: From a context of T states $(u_{t-T+1}, \ldots, u_t)$ predict the next state u_{t+1}

Two subtasks: (1) infer the operator (2) apply the operator to evolve the state

⟶ How to infer the operator f_θ from each context $(u_{t-T+1}, \ldots, u_t)$?

Meta Learning the Operator from Context

1. Gradient adaptation. Learn an operator network f_θ every new context



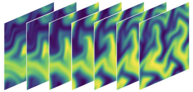
gradient descent on ξ

$$u_{t+\Delta t} \approx u_t + f_\theta\left(u_t\right)\Delta t \qquad \theta = \theta_c + W\xi$$

very costly

[GEPS, Koupaï et al. 2024]

2. DISCO (ICML 2025): output the operator itself



single forward

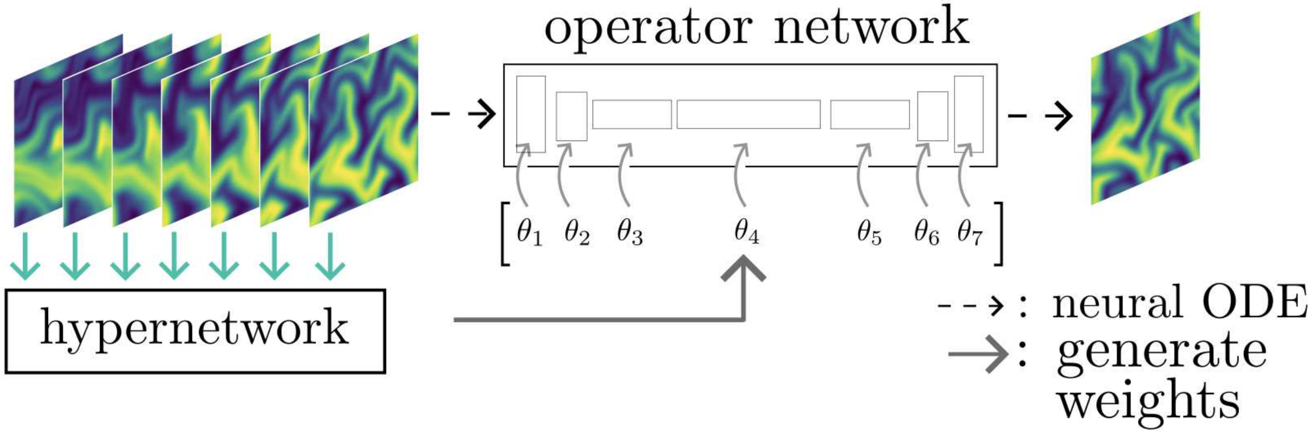
$$\theta$$

Rudy Morel

Edouard Oyallon

DISCO: learning to DISCover an evolution Operator for multi-physics-agnostic prediction

Learning to DISCover an evolution Operator from data



$$\min \frac{1}{|\mathcal{D}|} \sum_{k \in \mathcal{D}} \mathrm{Loss}(u_{t+1}^k, \hat{u}_{t+1}^k)$$

$$\hat{u}_{t+1}^k = f_\theta(u_t^k), \quad \theta = \mathrm{Transformer}(u_{t-T+1}^k, \dots, u_t^k)$$

- decouple operator inference from state evolution
- enforce an “information bottleneck” in the operator: intrinsic $\dim \theta = 384$

DataSets

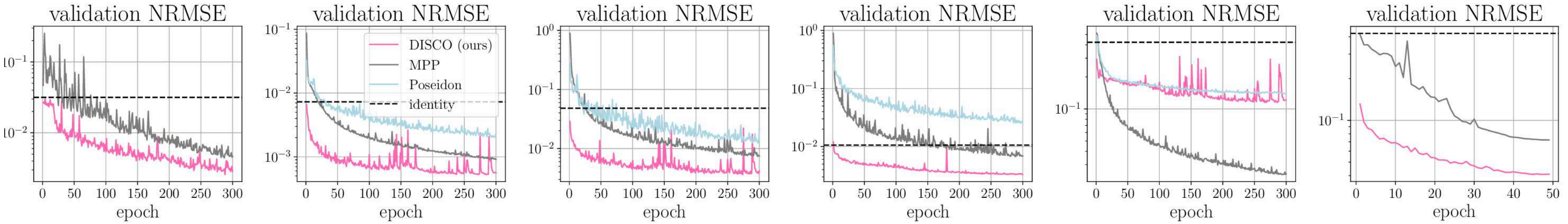
PDEBench + The Well:
1D, 2D, 3D, different resolution/quantities/boundary conditions

lvl1 lvl2 lvl3

Table 1. The datasets from PDEBench (Takamoto et al., 2022) and The Well (Ohana et al., 2024) used in this paper.

Dataset Name	Physical Dimension	# of Fields	Resolution (time)	Resolution (space)	Boundary Conditions
Burgers	1D	1	200	1024	periodic
Shallow water equation	2D	1	100	128×128	open
Diffusion-reaction	2D	2	100	128×128	Neumann
Incomp. Navier-Stokes (INS)	2D	3	1000	512×512	Dirichlet
Comp. Navier-Stokes (CNS)	2D	4	21	512×512	periodic
active matter	2D	11	81	256×256	periodic
Euler multi-quadrants	2D	5	100	512×512	periodic / open
Gray-Scott reaction-diffusion	2D	2	1001	128×128	periodic
Rayleigh-Bénard	2D	4	200	512×128	periodic × Dirichlet
shear flow	2D	4	200	256×512	periodic
turbulence gravity cooling	3D	6	50	$64 \times 64 \times 64$	open
MHD	3D	7	100	$64 \times 64 \times 64$	periodic
Rayleigh-Taylor instability	3D	4	120	$64 \times 64 \times 64$	periodic × periodic × slip
supernova explosion	3D	6	59	$64 \times 64 \times 64$	open

State-of-the-art Prediction on PDEBench



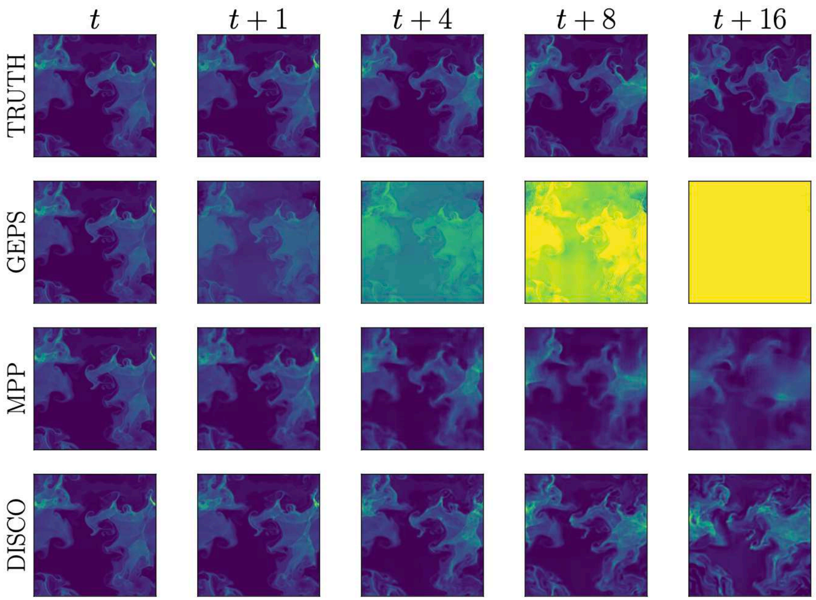
(a) Burgers (b) Shallow water (c) Diff.-reaction (d) Incomp. NS (e) Comp. NS (f) Comp. NS (alone)

Table 5. **Number of epochs to reach SOTA performance on PDEBench datasets.**

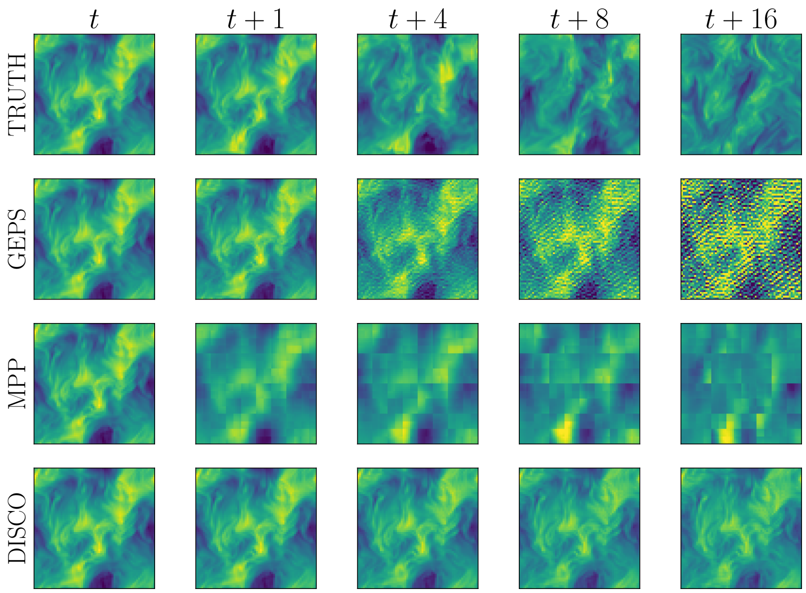
Model	# parameters	Burgers	SWE	DiffRe2D	INS	CNS	CNS (alone)
MPP (retrained)	160m	500	500	500	500	**500**	50
DISCO (ours)	119m	**277**	**70**	**55**	**35**	-	**7**

→ requires far fewer epochs on most datasets

Rollout on the Well dataset



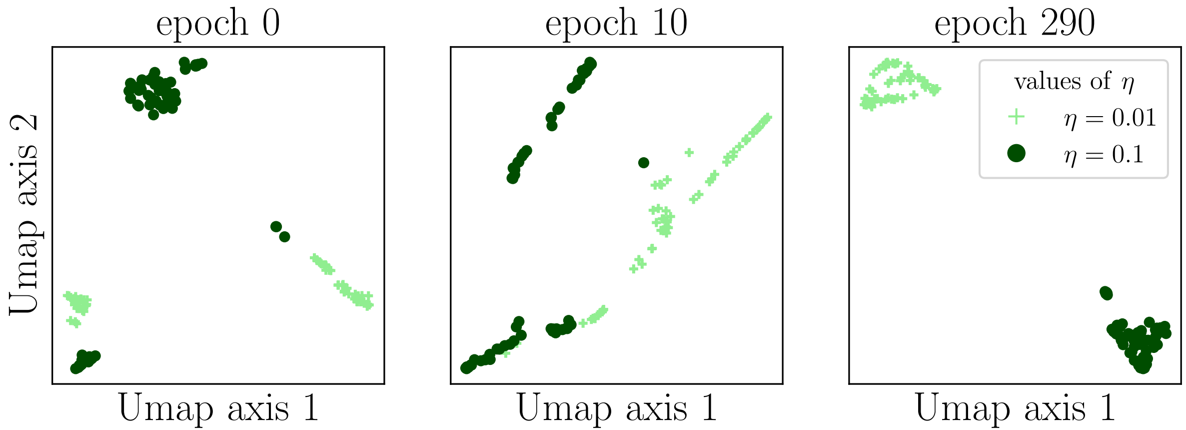
Euler



2d slices of 3d MHD

A Shared Latent Space for Physics

Space of the evolution operators



Significantly better for physical parameter prediction than full-pixel models

	MSE (↓)		
	active matter	shear flow	Rayleigh-Bénard convection
DISCO	**0.057**	**0.13**	**0.01**
MPP (full finetuning)	0.230	0.59	0.08

Qu, Morel, et al. 2026

The Out-of-Distribution Challenge

DISCO and other foundation models can achieve good accuracy in-distribution

But performance degrades sharply out-of-distribution:

- Unseen PDE coefficients
- Unseen forcing terms
- Novel combinations of physics

Current remedy: large-scale pretraining + fine-tuning on OOD data

→ Requires examples from the new dynamics
→ Falls short of true zero-shot generalization

Test-Time Compute: The LLM Paradigm

paradigm shift
for challenging tasks

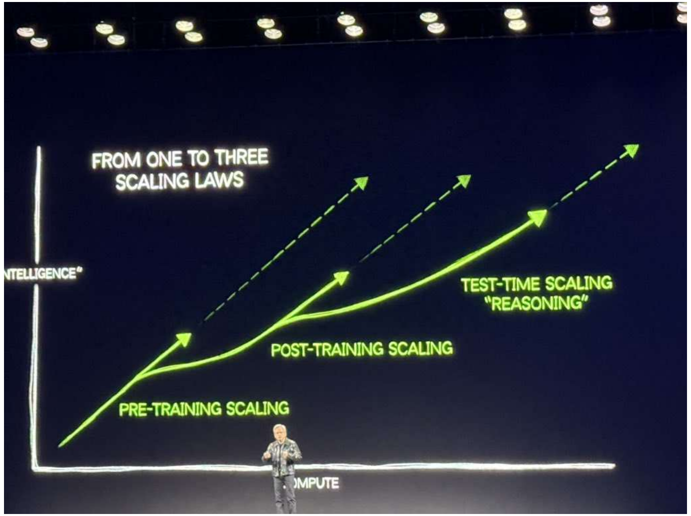
Train-Time Scaling

GPT-2
1.5B

GPT-3
175B

GPT-4
~1.8T

↓ More parameters

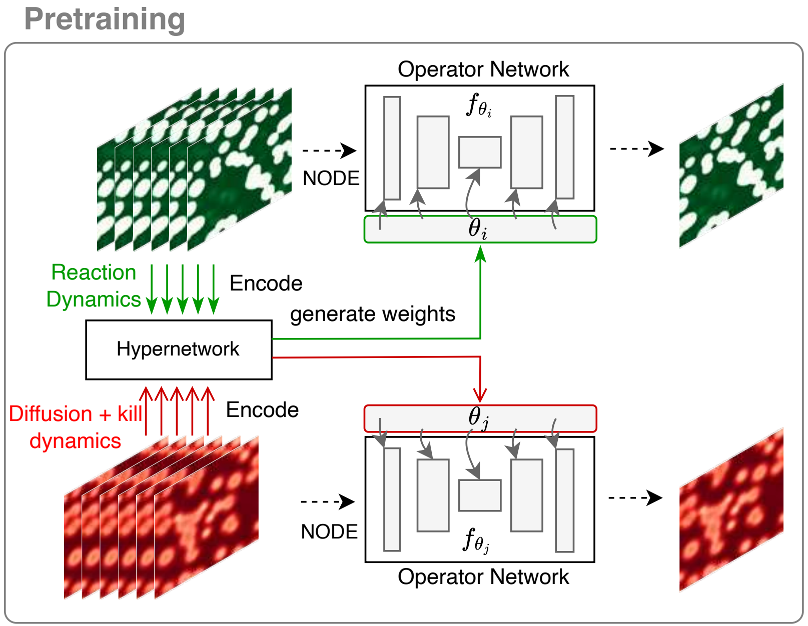


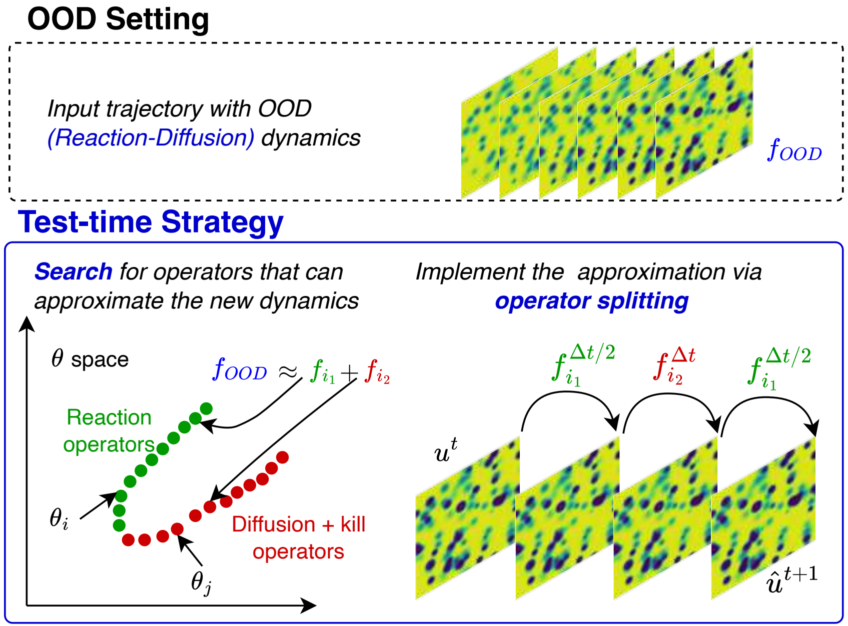
Test-Time Scaling

- Chain-of-Thought
  Step-by-step reasoning
- Beam Search
  Explore multiple paths
- Self-Refinement
  Iterative improvement

↓ More compute at inference

Part II: Test-Time Scaling and Generalization for Physics





Extract a dictionary of operators $\{f_1, \cdots, f_N\}$ after pretraining

Search for the optimal operator composition via operator splitting on the test trajectory

Operator Composition Search

Find a small subset $\{f_{i_1}, \cdots, f_{i_m}\}$ from the dictionary $\{f_1, \cdots, f_N\}$ such that $\hat{f} = f_{i_1} + \cdots + f_{i_m}$ best approximate the test dynamics

$$\min_{\hat{f}} \sum_t \mathrm{Loss}(u_t^{\mathrm{test}}, \hat{u}_t^{\mathrm{test}})$$

Search strategies:

1. Uniform search: randomly sample K operator subsets, pick the best

2. Beam search: progressively expand operator sets

$\mathscr{B}_0$ = top-B singletons from $\{f_1, \cdots, f_N\}$

$\mathscr{B}_{m+1}$ = top-B from $\{S \cup \{f_i\} : S \in \mathscr{B}_m\}$

Complexity: $O(BN)$ evaluations per iteration

Test-time Generalization for Physics through Neural Operator Splitting


Louis Serrano


Edouard Oyallon


Shirley Ho

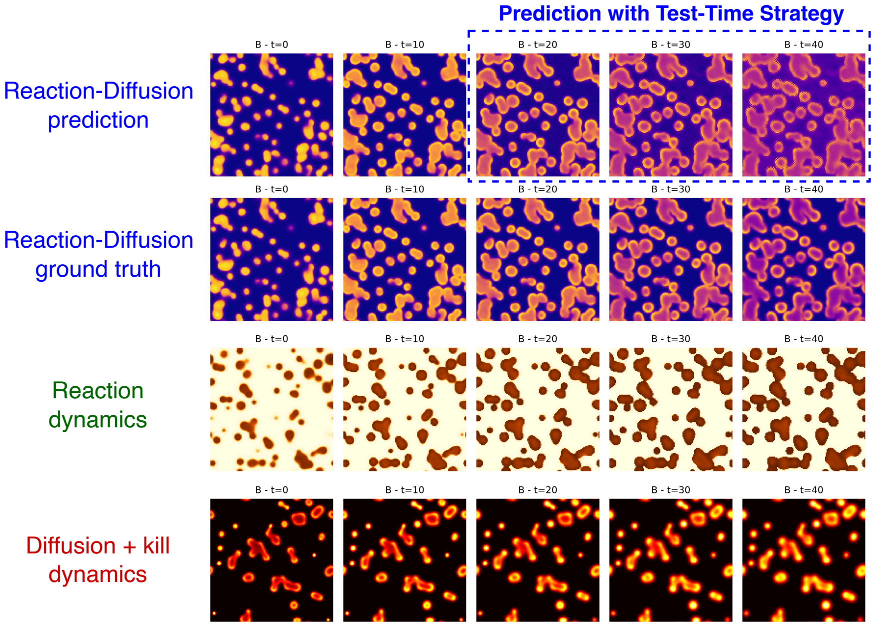

Rudy Morel

Out-of-Distribution for Reaction-Diffusion

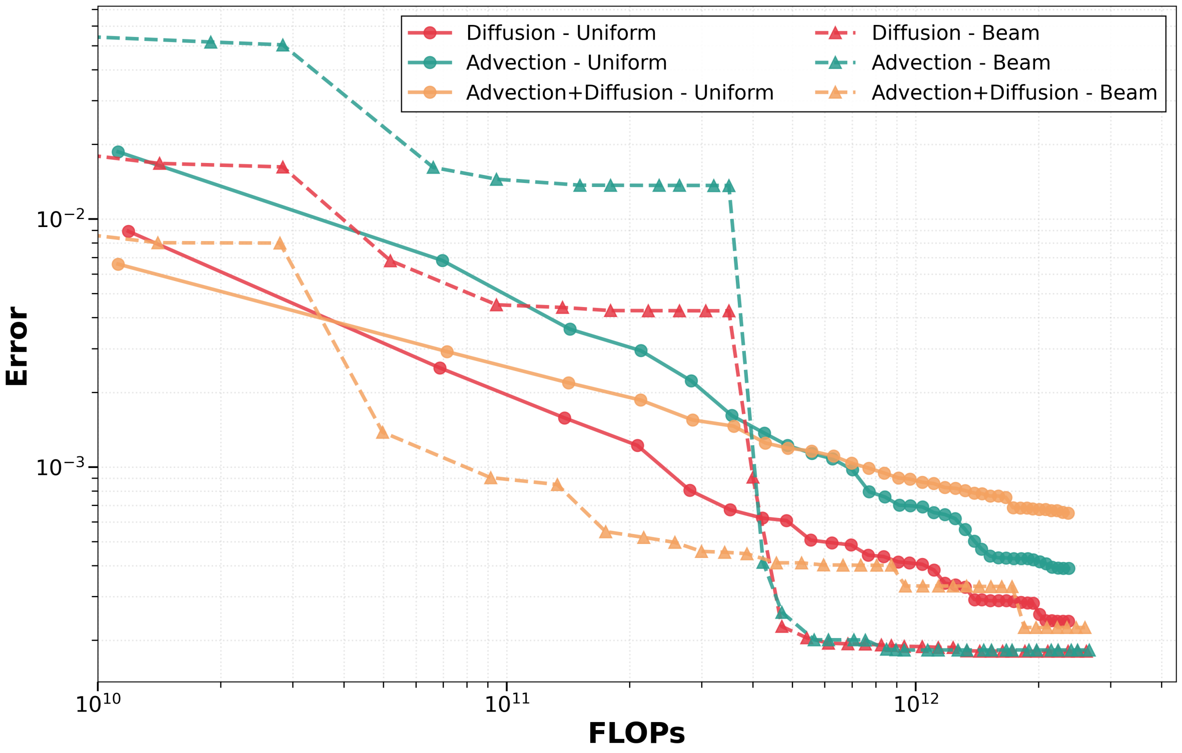
$$\frac{\partial A}{\partial t} = D_A \nabla^2 A - \delta A B^2 + F(1 - A) \ ,$$
$$\frac{\partial B}{\partial t} = D_B \nabla^2 B + \delta A B^2 - (F + k) B \ .$$

Two types of training data

1. pure reaction ($D_A = D_B = 0, \delta = 1$ $k = 0$, varying F)
2. diffusion-kill (fixed nonzero D_A, D_B $\delta = F = 0$, varying k)



Scaling Performance



Takeaway

FLATIRON INSTITUTE

- Multi-physics-agnostic prediction with multi-level variability is essential for generalization
- DISCO exploits the the structure of evolution operator while harnessing the power of transformers
- The decoupled design of DISCO offers test-time generalization via operator composition

paper 1

paper 2

One Big Open Question for Training Data

Classical ML: Natural images, natural language

Scientific ML: Natural amino acids, natural weather/climate data

For PDE problems: what is natural data for training?

1. Clearer motivation and smarter design for science and engineering applications
2. Generative modeling can help infer data distributions from indirect measurements

Thanks for your attention