Reasoning Powered Learning Using Scientific Approaches

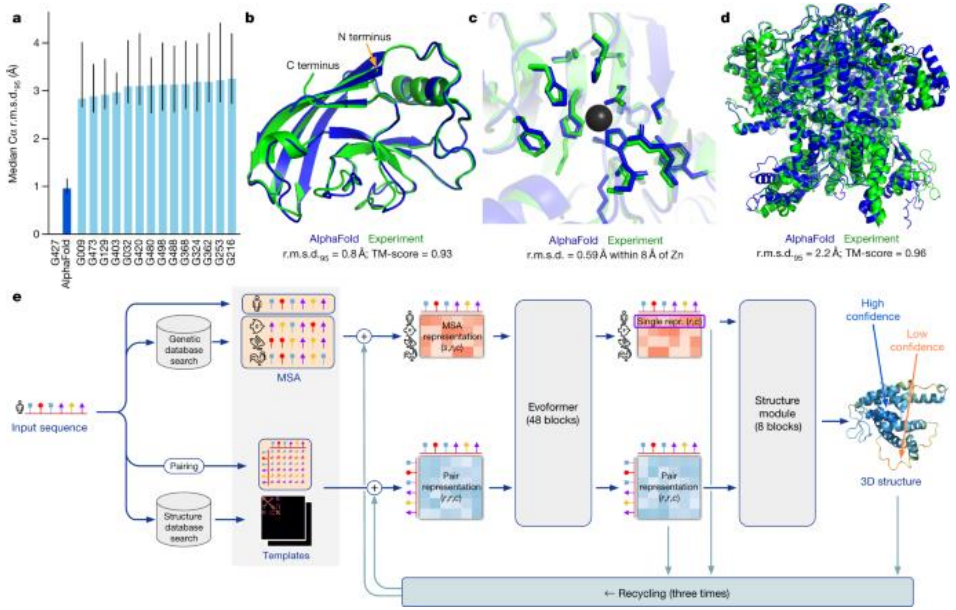
Yexiang Xue (*yexiang@purdue.edu*)
Department of Computer Science
Purdue University, USA

In collaboration with Nan Jiang, Md Nasim, Chonghao Sima, Xinghang Zhang, Anter El-Azab.

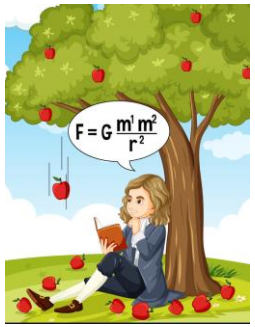
Deep Learning Lacks Deep Reasoning

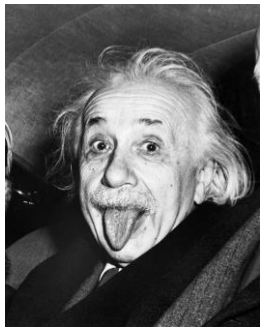
- Exciting progress in deep learning

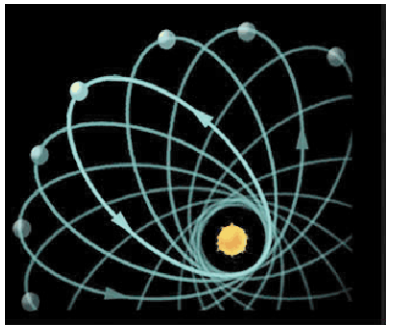
[AlphaFold]



- Human learning (discovery) is **better**!







  - Learning from an incredibly small set of “surprising” samples
  - Interpretable, elegant models & equations
  - Active exploration with a purpose
  - Machine learning based on Stochastic Gradient Descend (SGD) hardly captures its essence

$$\mathrm{F} = \mathrm{G}\frac{m_1 m_2}{r^2} \longrightarrow \text{2.1?}$$

Reasoning Powered Learning using scientific approaches

Newton and Einstein's examples show the role of ***active reasoning*** in scientific discoveries (***learning***).

- What new hypothesis can explain the data?
- What new experiments we can design to validate the hypothesis?
- What conclusions we can draw from the observations?

This talk: explore how "***control variable experiments***" as a classical scientific approach *expedites* scientific machine learning.

Symbolic regression

X_1	X_2	X_3	Y
2.5	1.0	9.5	12
3.0	-1.0	4.0	1
1.6	3.5	5.2	10.8
1.8	1.0	3.2	5
7.1	8.6	3.8	64.9
1.7	1.0	2.3	4
2.5	2.6	3.1	9.6
8.9	1.1	2.0	11.8
4.2	-1.0	2.2	-2
5.8	1.0	7.2	13
1.6	5.7	1.2	10.3
9.7	-1.0	1.7	-8

- Learning a symbolic expression from data
  - A good benchmark mimicking scientific discovery process.
- Incredibly difficult because of the large search space of all possible expressions.

- Can you guess which equation $y = f(x_1, x_2, x_3)$ generates the data shown in the left table?

Symbolic regression

X_1	X_2	X_3	Y
2.5	1.0	9.5	12
1.8	1.0	3.2	5
1.7	1.0	2.3	4
5.8	1.0	7.2	13

- Learning a symbolic expression from data
  - A good benchmark mimicking scientific discovery process.
- Incredibly difficult because of the large search space of all possible expressions.

- Can you guess which equation $y = f(x_1, x_2, x_3)$ generates the data shown in the left table?
- How about if I only ask you to look into these rows?

$$y = x_1 + x_3?$$

Symbolic regression

X_1	X_2	X_3	Y
3.0	-1.0	4.0	1
4.2	-1.0	2.2	-2
9.7	-1.0	1.7	-8

- Learning a symbolic expression from data
  - A good benchmark mimicking scientific discovery process.
- Incredibly difficult because of the large search space of all possible expressions.

- Can you guess which equation $y = f(x_1, x_2, x_3)$ generates the data shown in the left table?
- How about if I only ask you to look into these rows?

$$y = x_1 + x_3?$$

- How about these rows?

$$y = -x_1 + x_3?$$

Symbolic regression

Red and blue data are two control variable experiment trials (X_2 controlled)!
Control variable experiments ***simplify*** symbolic regression!

X_1	X_2	X_3	Y
2.5	1.0	9.5	12
3.0	-1.0	4.0	1
1.8	1.0	3.2	5
1.7	1.0	2.3	4
4.2	-1.0	2.2	-2
5.8	1.0	7.2	13
9.7	-1.0	1.7	-8

- Learning a symbolic expression from data
  - A good benchmark mimicking scientific discovery process.
- Incredibly difficult because of the large search space of all possible expressions.

- Can you guess which equation $y = f(x_1, x_2, x_3)$ generates the data shown in the left table?
- How about if I only ask you to look into these rows?

$$y = x_1 + x_3?$$

- How about these rows?

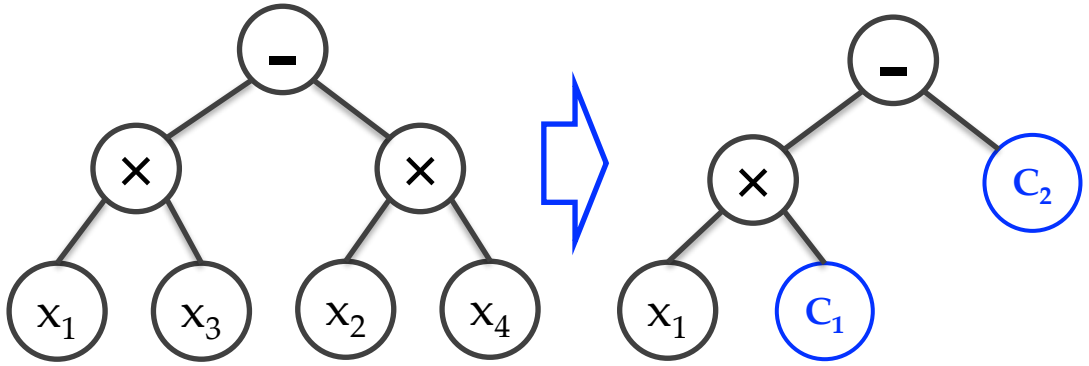
$$y = -x_1 + x_3?$$

- Maybe the equation is:

$$y = x_2 x_1 + x_3?$$

INDEED!

Control Variable Experiments



(a) Ground-truth expression

(b) Reduced form after controlling x_2, x_3, x_4

x_1	x_2	x_3	x_4	y
0.3	0.5	0.1	0.7	-0.32
0.6	0.5	0.1	0.7	-0.29
0.2	0.5	0.1	0.7	-0.33
0.9	0.5	0.1	0.7	-0.26

controlled

(c) Trial T_1

x_1	x_2	x_3	x_4	y
0.6	0.3	0.8	0.2	0.42
0.1	0.3	0.8	0.2	0.02
0.2	0.3	0.8	0.2	0.10
0.9	0.3	0.8	0.2	0.66

controlled

(d) Trial T_2

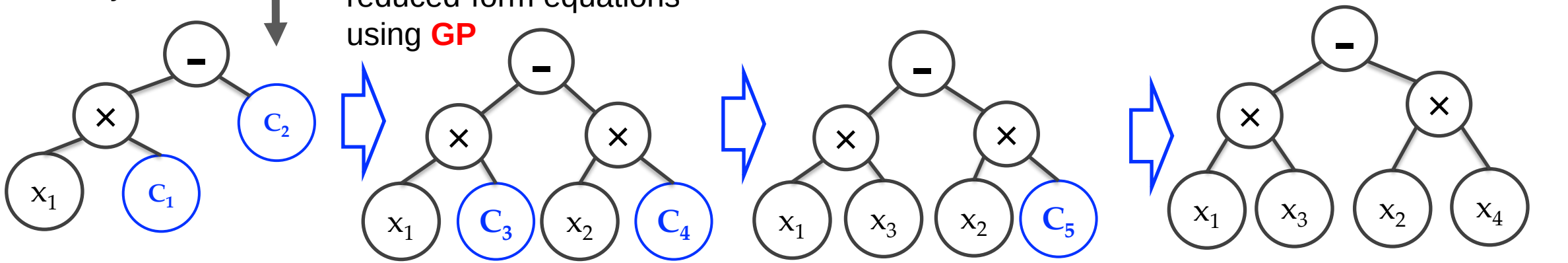
- **Control variable experimentation** – a classic procedure widely implemented and proven useful in science.
- **Controlled variables**: take the same value in a trial, but vary in values across trials
- **Free variables**: values change within a trial
- **Ground-truth equation**: the hidden equation that generates the data
- **Reduced form equation**: Under a controlled experiment, the data looks “as if” generated by the reduced equation, in which controlled variables are replaced with constants.

Control Variable Genetic Programming (CVGP)

Best Expression

Found by **GP**

Extended from previous reduced-form equations using **GP**



Multiple Trials

x_1	x_2	x_3	x_4	y
0.3	0.5	0.1	0.7	-.32
0.6	0.5	0.1	0.7	-.29
0.2	0.5	0.1	0.7	-.33
0.9	0.5	0.1	0.7	-.26

(a) Control x_2,x_3,x_4

x_1	x_2	x_3	x_4	y
0.6	0.1	0.8	0.4	0.44
0.4	0.9	0.8	0.4	0.04
0.3	0.2	0.8	0.4	0.16
0.7	0.4	0.8	0.4	0.40

(b) Control x_3,x_4

x_1	x_2	x_3	x_4	y
0.7	0.8	0.1	0.2	-.09
0.5	0.4	0.6	0.2	0.22
0.2	0.1	0.9	0.2	0.16
0.3	0.5	0.1	0.2	-.07

(c) Control x_4

x_1	x_2	x_3	x_4	y
0.2	0.4	0.2	0.7	-.24
0.9	0.3	0.5	0.5	0.30
0.5	0.4	0.8	0.1	0.36
0.1	0.8	0.7	0.6	-.41

(d) No control

Experiment Results

Ops	Dataset configs	CVGP (ours)		GP		DSR		PQT		VPG		GPMeld	
		50%	75%	50%	75%	50%	75%	50%	75%	50%	75%	50%	75%
inv	(2,1,1)	0.198	0.490	**0.024**	**0.053**	0.032	3.048	0.029	0.953	0.041	0.678	0.387	22.806
	(4,4,6)	**0.036**	**0.088**	0.038	0.108	1.163	3.714	1.016	1.122	1.087	1.275	1.058	1.374
	(5,5,5)	0.076	0.126	**0.075**	**0.102**	1.028	2.270	1.983	4.637	1.075	2.811	1.479	2.855
	(5,5,8)	**0.061**	**0.118**	0.121	0.186	1.004	1.013	1.005	1.006	1.002	1.009	1.108	2.399
	(6,6,8)	**0.098**	**0.144**	0.104	0.167	1.006	1.027	1.006	1.020	1.009	1.066	1.035	2.671
	(6,6,10)	**0.055**	**0.097**	0.074	0.132	1.003	1.009	1.005	1.008	1.004	1.015	1.021	1.126
sin, cos	(3,2,2)	**0.098**	**0.165**	0.108	0.425	0.350	0.713	0.351	1.831	0.439	0.581	0.102	0.597
	(4,4,6)	**0.078**	**0.121**	0.120	0.305	7.056	16.321	5.093	19.429	2.458	13.762	2.225	3.754
	(5,5,5)	**0.067**	**0.230**	0.091	0.313	32.45	234.31	36.797	229.529	14.435	46.191	28.440	421.63
	(5,5,8)	**0.113**	**0.207**	0.119	0.388	195.22	573.33	449.83	565.69	206.06	629.41	363.79	666.57
	(6,6,8)	**0.170**	**0.481**	0.186	0.727	1.752	3.824	4.887	15.248	2.396	7.051	1.478	6.271
	(6,6,10)	**0.161**	**0.251**	0.312	0.342	11.678	26.941	5.667	24.042	7.398	25.156	11.513	28.439
sin, cos, inv	(3,2,2)	0.049	**0.113**	**0.023**	0.166	0.663	2.773	1.002	1.992	0.969	1.310	0.413	2.510
	(4,4,6)	**0.141**	**0.220**	0.238	0.662	1.031	1.051	1.297	1.463	1.051	1.774	1.093	1.769
	(5,5,5)	**0.157**	0.438	0.195	**0.337**	1.098	3.617	1.018	5.296	1.012	1.27	1.036	3.617
	(5,5,8)	**0.122**	**0.153**	0.166	0.186	1.009	1.103	1.017	1.429	1.007	1.132	1.07	2.904
	(6,6,8)	**0.209**	**0.590**	**0.209**	0.646	1.003	1.153	1.047	1.134	1.059	1.302	1.029	3.365
	(6,6,10)	0.139	0.232	**0.073**	**0.159**	1.654	3.408	1.027	1.069	1.009	1.654	1.445	2.106

Median (50%) and 75%-quantile NMSE values of the symbolic expressions found by all the algorithms on several noisy benchmark datasets. Our CVGP finds symbolic expressions with the smallest NMSEs.

Discover New Physics from Data

Physics Model

Cahn-Hilliard:

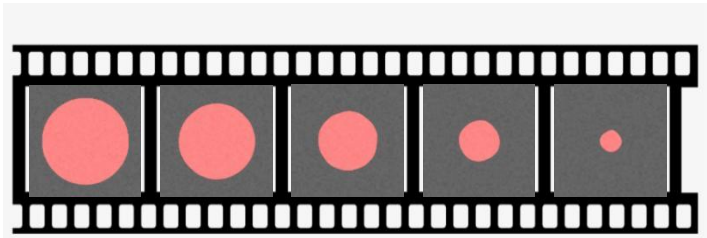
$$\frac{\partial u}{\partial t} = \nabla\left(M\nabla\frac{1}{N}\frac{\delta F}{\delta u}\right)$$

Allen-Cahn:

$$\frac{\partial v}{\partial t} = -L\frac{\delta F}{\delta v}$$

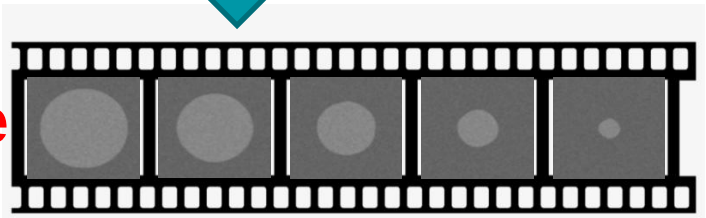
Simulation

Physics Knowledge Discovery



Simulation

Match?



Experiment

Discover New Physics from Data

Physics Model

Cahn-Hilliard:

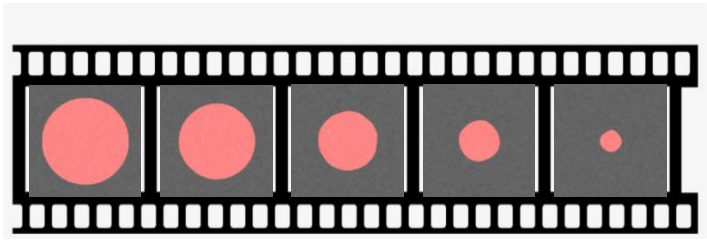
$$\frac{\partial u}{\partial t} = \nabla\left(M\nabla\frac{1}{N}\frac{\delta F}{\delta u}\right)$$

Allen-Cahn:

$$\frac{\partial v}{\partial t} = -L\frac{\delta F}{\delta v}$$

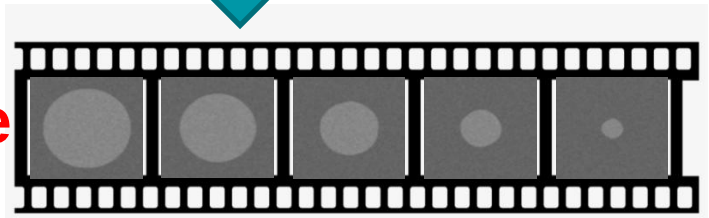
Simulation

Physics Knowledge Discovery



Simulation

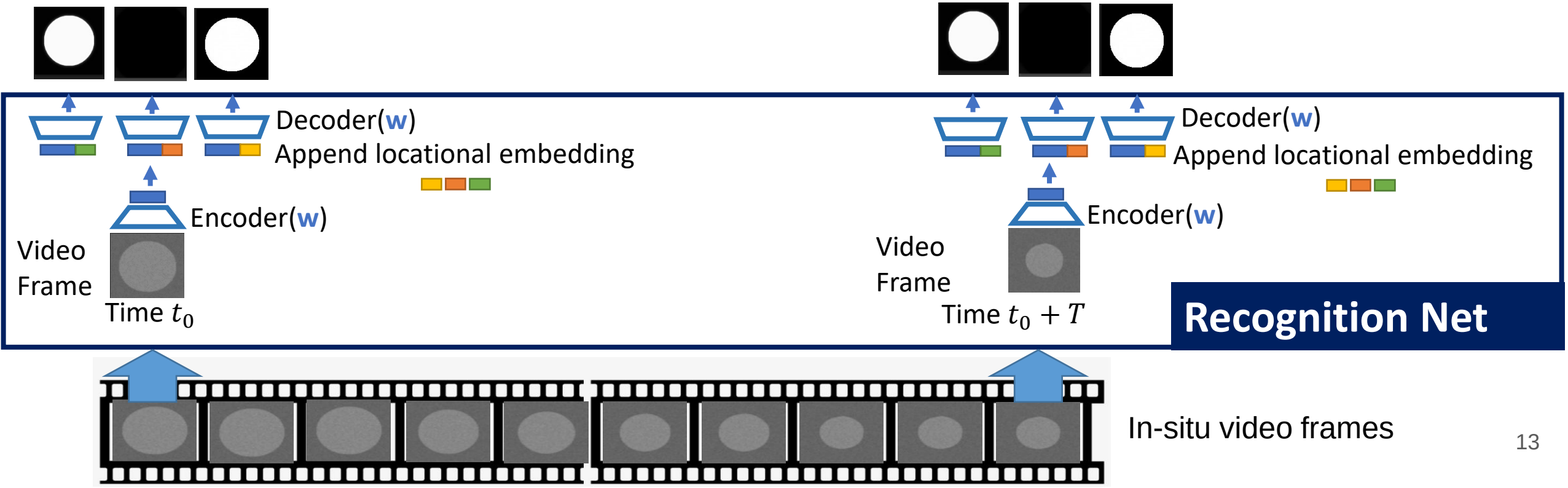
Match?



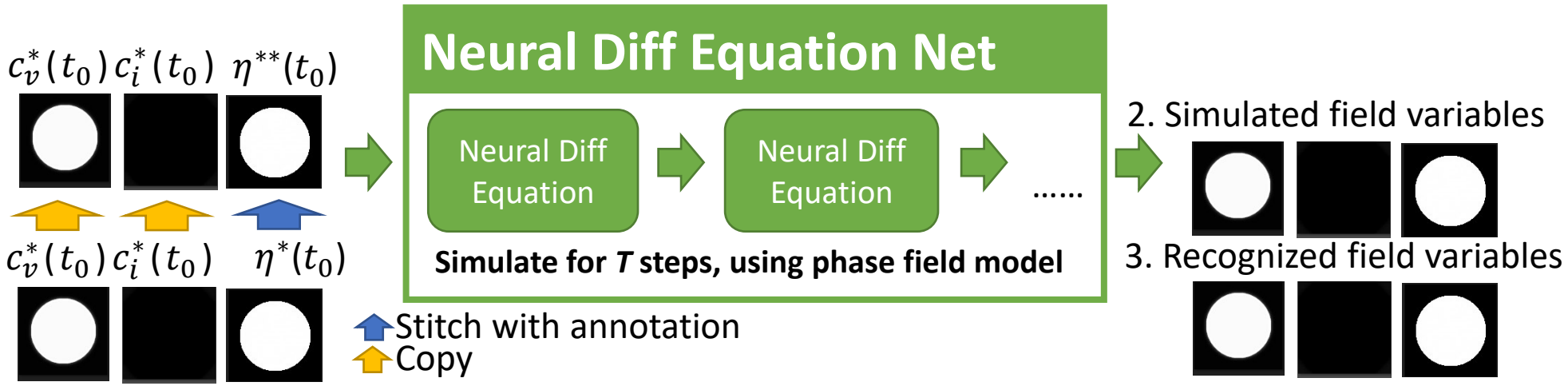
Experiment

Can machine learning automatically *discover new science*?

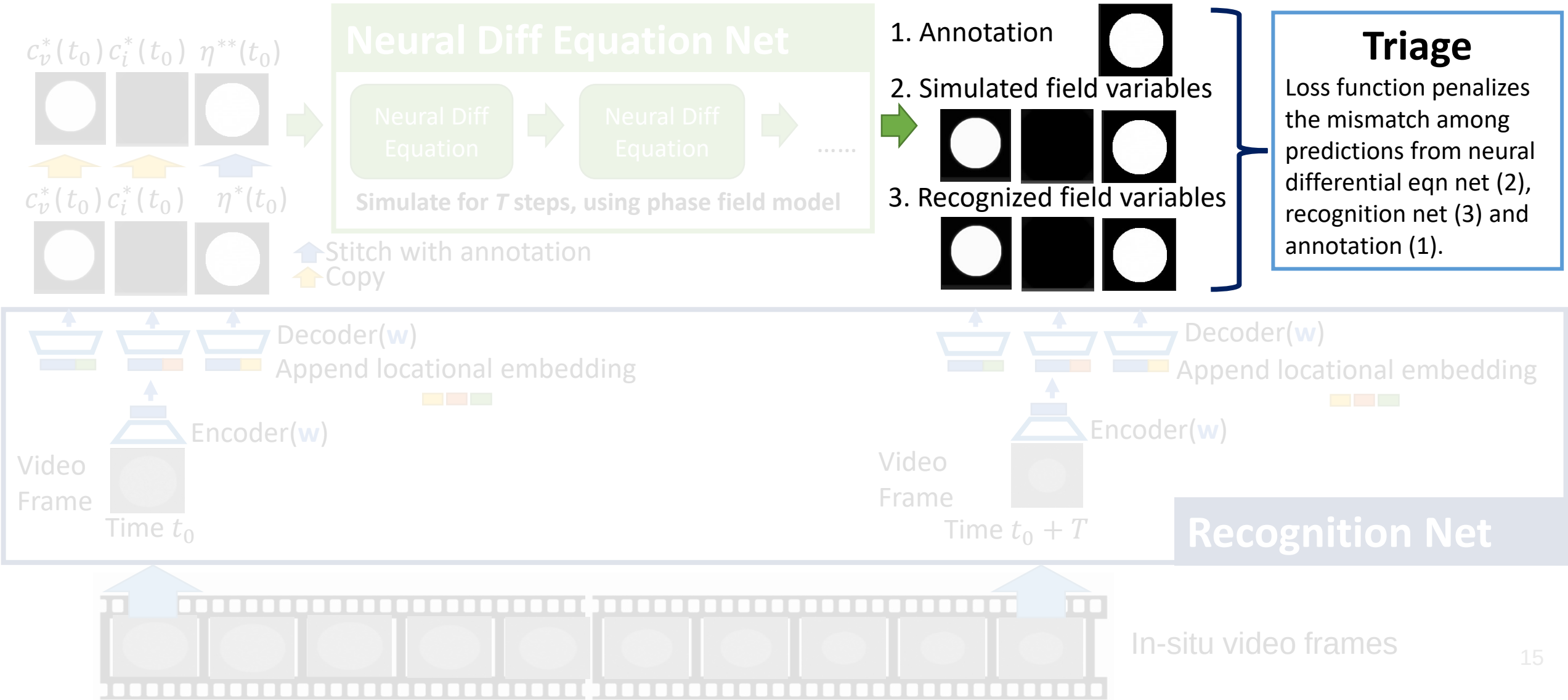
NeuraDiff: High Level Idea



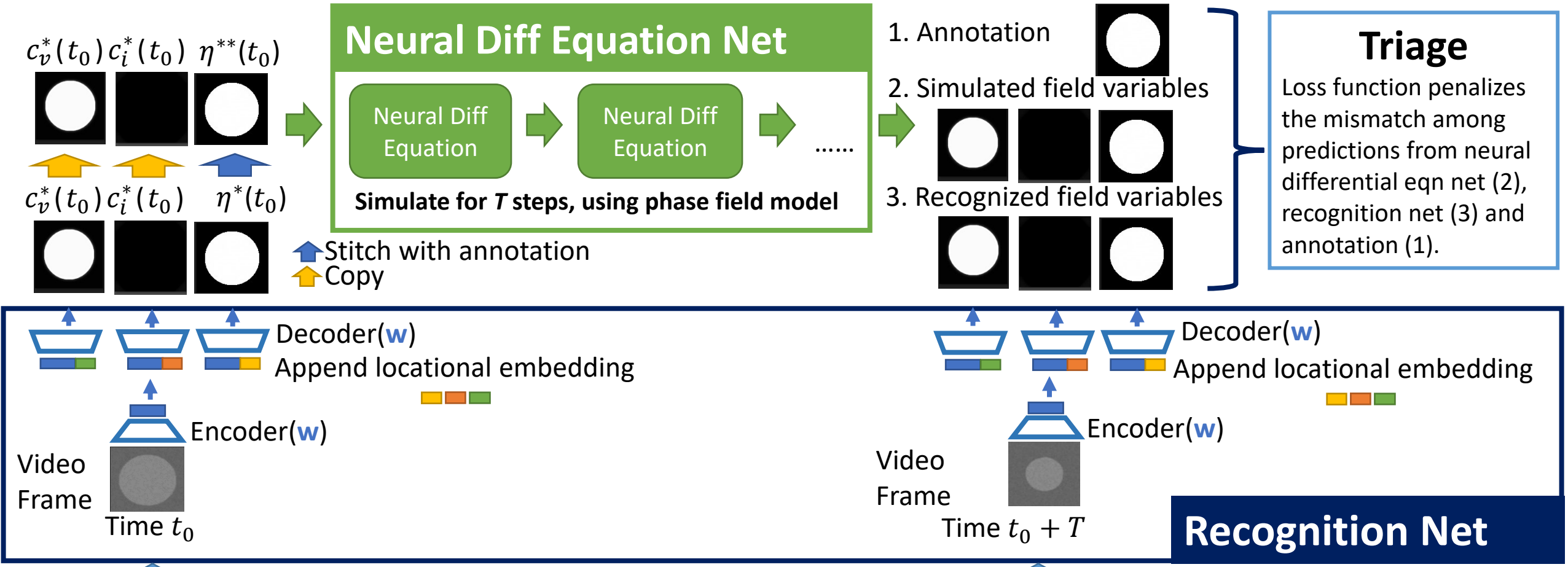
NeuraDiff: High Level Idea

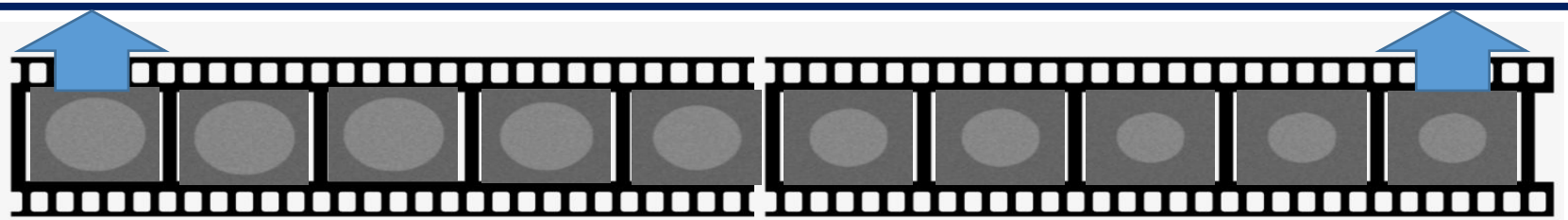


NeuraDiff: High Level Idea



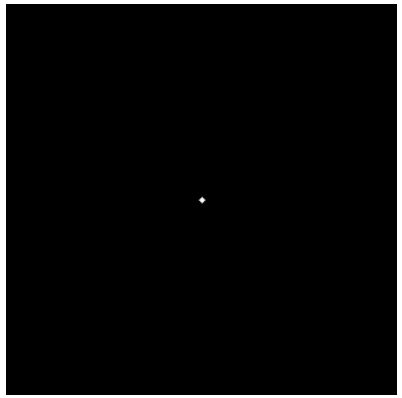
NeuraDiff: High Level Idea





In-situu video frames

Learning models for dendritic solidification



Ground-truth ϕ

Phase-field model:

$$F(\phi, m) = \int \left(\frac{1}{2}\epsilon^2 |\nabla\phi|^2 + f(\phi, m)\right) dv,$$

$$f(\phi, m) = \frac{1}{4}\phi^4 - \left(\frac{1}{2} - \frac{1}{3}m\right)\phi^3 + \left(\frac{1}{4} - \frac{1}{2}m\right)\phi^2,$$

$$\epsilon = \bar{\epsilon}\sigma(\theta),$$

$$\sigma(\theta) = 1 + \delta\cos(j(\theta - \theta_0)),$$

$$\theta = \tan^{-1}\left(\frac{\partial\phi/\partial y}{\partial\phi/\partial x}\right),$$

$$m(T) = (\alpha/\pi)\tan^{-1}[\gamma(T_{eq} - T)],$$

Dendritic growth follows Allen-Cahn equation:

$$\tau\frac{\partial\phi}{\partial t} = -\frac{\delta F}{\delta\phi}$$

Temperature follows conservation law:

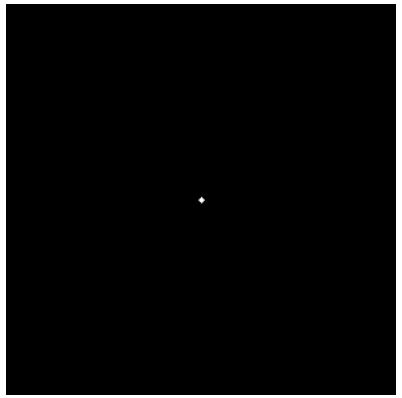
$$\frac{\partial T}{\partial t} = \nabla^2 T + \kappa\frac{\partial\phi}{\partial t}$$

Controlled learning experiment

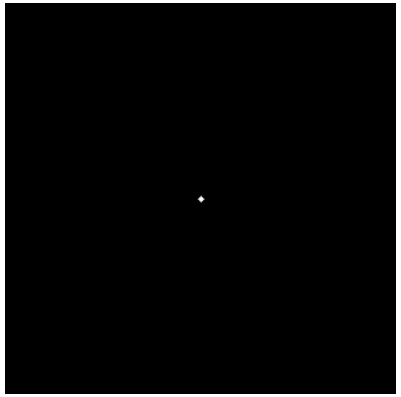
- Intentionally first learn on data in which $\nabla\phi = 0$;
- In this case, blue parameters do not affect dynamics;
- Focus on learning red parameters.

- Allow $\nabla\phi$ to vary in the second stage, hence start to learn blue parameters.

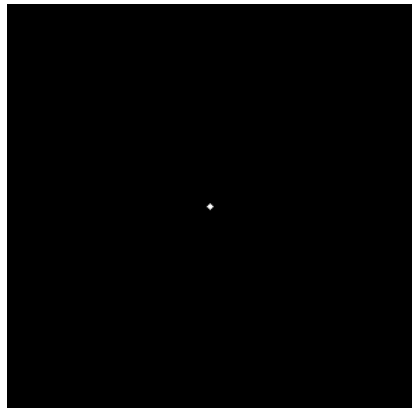
Comparison



Ground-truth ϕ



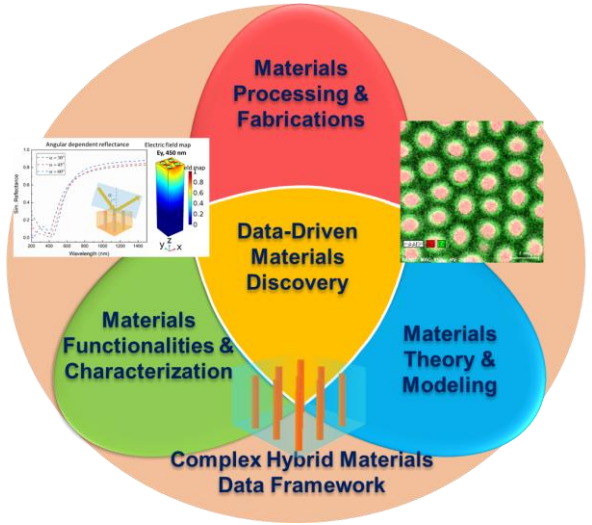
Learning all parameters at once

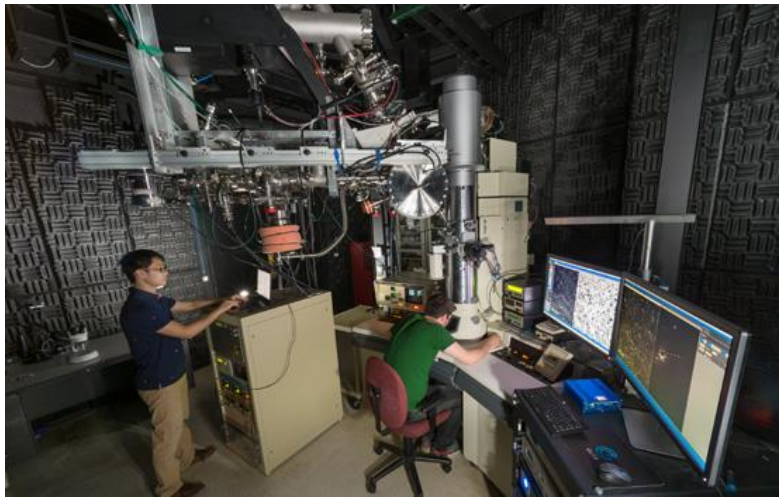


Controlled learning experiments

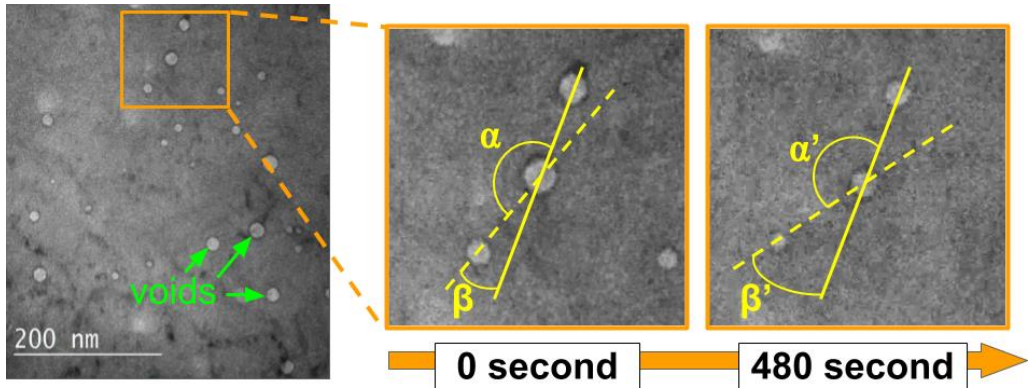
AI Driven Materials Discovery in Extreme Conditions

- Search for strong materials under heavy irradiation and extreme high temperature
- Understand defect formation, migration in extreme conditions
- Better materials for future nuclear reactors
- In-situ experimentation



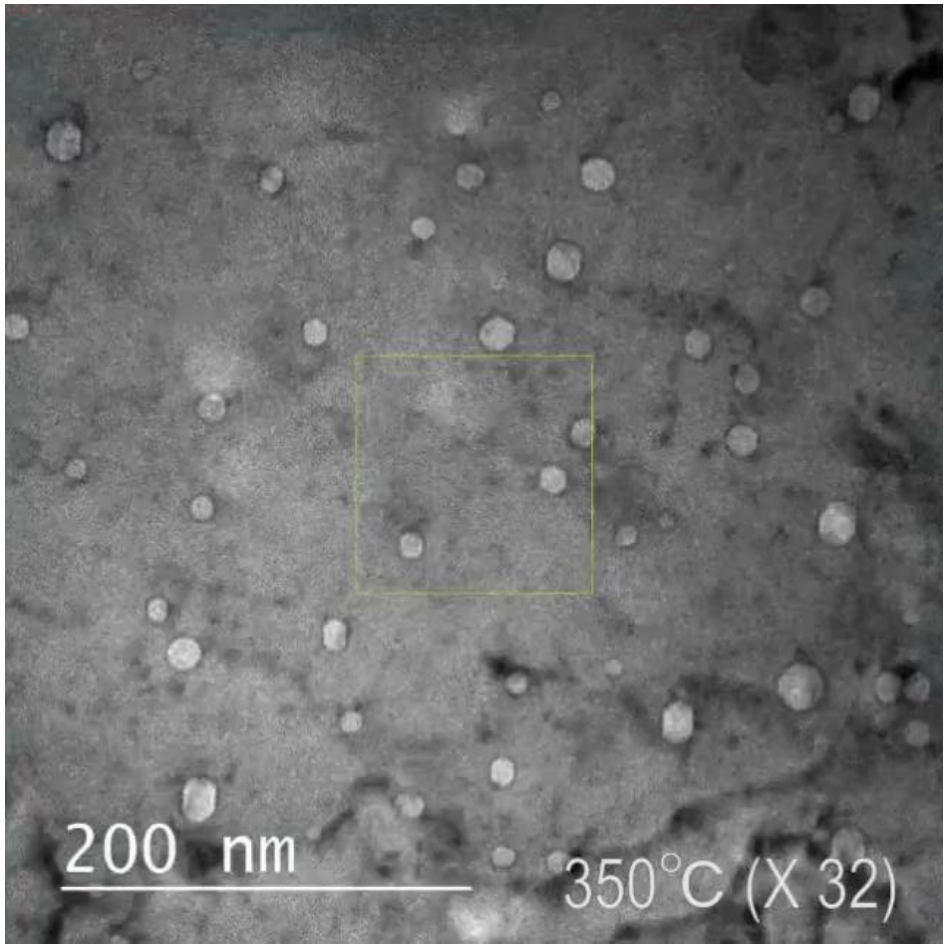


In-situ experiment setup (Argonne National Lab)



Nanovoid defects captured by in-situ experiment

In situ Video

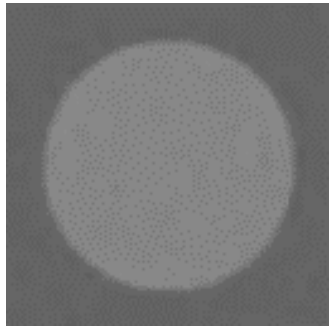


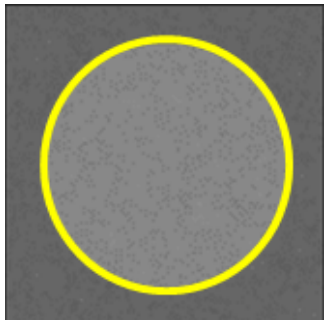
Novel AI techniques are needed!

- Terabytes of data
- Beyond manual effort
- 3.75 months work (40 hours per week) analyzing a 10-minute video if spend 5 minutes per frame

Track Nanovoids + Learn Phase Field Model

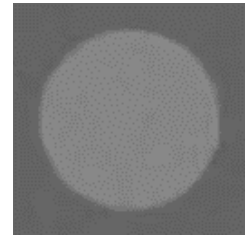
TEM Video





Partial Annotation
(every 10th frame)

NN cannot predict future well



Nanovoid Tracking

Identify phase field model parameters

$$F = N\int\left[h(\eta)f^{s}(c_v,c_i) + j(\eta)f^{v}(c_v,c_i) + \frac{\kappa_v}{2}|\nabla c_v|^2 + \frac{\kappa_i}{2}|\nabla c_i|^2 + \frac{\kappa_\eta}{2}|\nabla\eta|^2\right]dV,$$

$$\frac{\partial c_v}{\partial t} = \nabla\cdot\left(M_v\nabla\frac{1}{N}\frac{\delta F}{\delta c_v}\right) + \xi + P_v - R_{iv},$$

$$\frac{\partial c_i}{\partial t} = \nabla\cdot\left(M_i\nabla\frac{1}{N}\frac{\delta F}{\delta c_i}\right) + \xi + P_i - R_{iv},$$

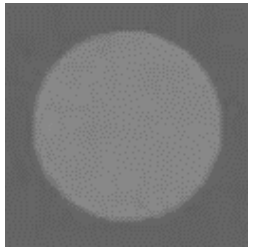
$$\frac{\partial \eta}{\partial t} = -\mathrm{L}\frac{\delta F}{\delta \eta} + \xi + P_{vi}.$$

Computational Materials Science
Volume 50, Issue 3, January 2011, Pages 949-959

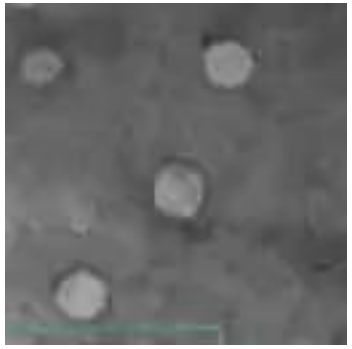
Phase-field simulation of irradiated metals: Part I: Void kinetics

Paul C. Millett, A.m. Anter El-Azab, Srujan Rokkam, Michael Tonks, Dieter Wolf

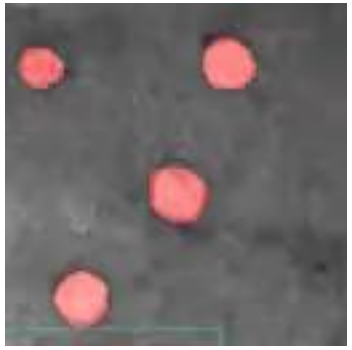
Simulate void evolution according to learned model



NeuraDiff for Real-world In-situ Video Data



In situ video



NeuraDiff Tracking Output

Compared to baseline methods, NeuraDiff shows **similar tracking** accuracy, and **superior learning** of physics model

Pixelwise Tracking Accuracy

	NeuraDiff	UNet Baseline
Synthetic Data	98.5%	99.9%
Real-world In-situ Data	96.2%	96.4%

Conclusions

- Control Variable Genetic Programming (CVGP) for symbolic regression
  - Learning from control variable experiments
  - Incrementally build complex equations from simple ones using genetic programming
- Neuradiff: learning partial differential equations from experiment data
  - Integrate recognition neural net with neural PDE net
- Controlled experiments improve learning dendritic solidification
- Learning nano-structure evolutions from experiment data
  - Applications in the search of strong materials for high temperature and irradiation applications
- Look into future: passive learning vs. active probing
  - Science progress resulted from insightful experiment design, courageous hypothesis forming (reasoning) + high-capacity modeling (learning)