AI + Science: motivation, advances and open questions

Tailin Wu Stanford University 3/26/2023 @ 集智斑图







Outline

- Why AI + Science?
- Al for science: important questions, recent advances and relations with machine learning
 - Al for scientific simulation
 - Al for scientific design
 - Al for scientific discovery
- Science for AI: important questions and recent advances
- Discussion: where to find the next AlphaFold and chatGPT?
 - What is the next big problems for AI + Science?
 - Potential bottlenecks and mid-term roadmap

Outline

• Why AI + Science?

- Al for science: important questions, recent advances and relations with machine learning
 - Al for scientific simulation
 - Al for scientific design
 - Al for scientific discovery
- Science for AI: important questions and recent advances
- Discussion: where to find the next AlphaFold and chatGPT?
 - What is the next big problems for AI + Science?
 - Potential bottlenecks and mid-term roadmap

AI + Science = (AI for Science) + (Science for AI)

What disciplines?











AI + Science = (AI for Science) + (Science for AI)



Why AI + Science?

Why AI for Science:

(*) Science solves critical problems
(1) AI significantly boosts the speed and accuracy
(2) AI helps explore vast design/control space
(3) AI helps uncover scientific knowledge

Why Science for AI:

. . .

. . .

(1) Novel challenges(2) Novel insights(3) Novel tools

Why AI for Science?

Why Science?

- Solves **critical problems** for humans and society:
 - Biomedicine
 - Materials
 - Energy
 - Mechanical engineering
 - Aerospace
 - Manufacturing
 - Revert global warming
 - ...





















Why AI for Science?

Why Science?

- Long-term survival of human as a species:
 - If the world GDP grows 2% per year, in 200 years, it needs to be 1.02¹⁰⁰=52x current GDP; in 1000 years, it needs to be 4x10⁸ times.
 - If slowed down, involution (内卷) will be more severe.
 - The long-term solution: make cake bigger (做 大蛋糕). Where is the cake?
 - The cake is at Solar system:
 - The Sun outputs 10⁹ more energy than that reaches Earth, can support an economy 10⁹ times the current GDP.
 - Requires advances in all scientific disciplines: materials, aerospace, biomedicine, ME, energy, ...





















Why AI for Science:

. . .

(*) Science solves critical problems
(1) AI significantly boosts the speed and accuracy
(2) AI helps explore vast design/control space
(3) AI helps uncover scientific knowledge

(1) AI significantly boosts the speed and accuracy

GraphCast [1] by DeepMind outperforms traditional methods in **accuracy** for 10-day forecast. Takes only 60s to predict on a TPU, while traditional method takes 1h running on 11,664 CPU cores.



[1] Lam, Remi, et al. "GraphCast: Learning skillful medium-range global weather forecasting." *arXiv preprint arXiv:2212.12794* (2022).

(1) AI significantly boosts the speed and accuracy

AlphaFold 2 [1] successfully predicts 98.5% of human proteins with atomic accuracy



[1] Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." *Nature* 596.7873 (2021): 583-589.

(2) AI helps explore vast design/control space

DeepMind used deep reinforcement learning (RL) for controlling [1] fusion plasma, and control novel plasma shapes.



[1] Degrave, Jonas, et al. "Magnetic control of tokamak plasmas through deep reinforcement learning." *Nature* 602.7897 (2022): 414-419.

(3) AI helps uncover scientific knowledge

AI Feynman 2.0 [1]: rediscover top-100 physics equations in Feynman lectures

Al Poincaré [2]: discover conservation laws from data

[3] learns physical concepts, and re-discovers that solar system is heliocentric

Some equations discovered by AI Feynman 2.0:

[1] Udrescu, Silviu-Marian, et al. "AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity." *NeurIPS 2020*

[2] Liu, Ziming, and Max Tegmark. "Machine learning conservation laws from trajectories." *Physical Review Letters* 126.18 (2021): 180604.

[3] Iten, Raban, et al. "Discovering physical concepts with neural networks." Physical review letters 124.1 (2020): 010508.

	Equation	Symmetries
1	$\delta = -5.41 + 4.9 \frac{\alpha - \beta + \gamma/\chi}{3\chi}$	TC
2	$\chi = 0.23 + 14.2 rac{lpha + eta}{3 \gamma}^{lpha}$	TS
3	$eta = 213.80940889 \left(1 - e^{-0.54723748542lpha} ight)$	
4	$\delta = 6.87 + 11 \sqrt{lpha eta \gamma}$	Р
5	$V = \left[R_1^{-1} + R_2^{-1} + R_3^{-1} + R_4^{-1} \right]^{-1} I_0 \cos \omega t \text{(Parallel resistors)}$	PGSM
6	$I_0 = \frac{V_0}{\sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}} (\text{RLC circuit })$	MG
7	$I = \frac{V_0 \cos \omega t}{\sqrt{R^2 + (\omega L - \frac{1}{\omega C})^2}} (\text{RLC circuit})$	MG
8	$V_2 = \left(\frac{R_2}{R_1 + R_2} - \frac{R_x}{R_x + R_3}\right) V_1 \text{(Wheatstone bridge)}$	SGMA
9	$v = c rac{(v_1 + v_2 + v_3)/c + v_1 v_2 v_3/c^3}{1 + (v_1 v_2 + v_1 v_3 + v_2 v_3)/c^2}$ (Velocity addition)	AG
0	$v = c \frac{(v_1 + v_2 + v_3 + v_4)/c + (v_2 v_3 v_4 + v_1 v_3 v_4 + v_1 v_2 v_4 + v_1 v_2 v_3)/c^3}{1 + (v_1 v_2 + v_1 v_3 + v_1 v_4 + v_2 v_3 + v_2 v_4 + v_3 v_4)/c^2 + v_1 v_2 v_3 v_4/c^4} $ (Velocity addition)	GA
.1	$z = (x^4 + y^4)^{1/4}$ (L ₄ -norm)	AC
2	$w = xyz - z\sqrt{1 - x^2}\sqrt{1 - y^2} - y\sqrt{1 - x^2}\sqrt{1 - z^2} - x\sqrt{1 - y^2}\sqrt{1 - z^2}$	GA
3	$z=rac{xy+\sqrt{1-x^2-y^2+x^2y^2}}{y\sqrt{1-x^2}-x\sqrt{1-y^2}}$	А
4	$z = y\sqrt{1-x^2} + x\sqrt{1-y^2}$	Α
5	$z=xy-\sqrt{1-x^2}\sqrt{1-y^2}$	Α
6	$r = rac{a}{\cot{(\alpha/2)} + \cot{(\beta/2)}}$ (Incircle)	GMAC

Why AI + Science?

AI for Science: how AI boosts science

(*) Science solves critical problems
(1) AI significantly boosts the speed and accuracy
(2) AI helps explore vast design/control space
(3) AI helps uncover scientific knowledge

Science for AI: how science boosts AI

(1) Novel challenges(2) Novel insights(3) Novel tools

. . .

. . .

(1) Novel challenges

A. Large-scale

- Large Hadron Collider (LHC): Each run (12h) generates 2000,000TB of data
 - Compare that chatGPT uses 570GB of data to train



- (1) Novel challenges
 - A. Large-scale
 - Simulation of realistic systems may need millions/billions of particles/cells per time step



astrophysics

Controlled nuclear fusion

hypersonic flow

@Fiuza

(1) Novel challenges

B. Generalization and robustness

- How to **generalize** to test dataset with distribution shift?
- How to be more **robust** to noise?
- How to deal with **small** number of examples



(2) Novel insigh	nts	Proposed by Surya Ganguli <i>et al.</i> , who is Stanford applie physics professor	
Concepts	Source	Application	
Symmetry	physics	equivariant GNN [1]	
Diffusion	physics	Diffusion models [2]	
Energy	physics	Energy-based models, Boltzmann machines	
Hamiltonian	physics	Hamiltonian Monte Carlo [3], Hamiltonian Neural Networks [4]	
Grid cells	neuroscience	Grid cells for navigation [5]	
Working memory	neuroscience	Working Memory Graphs [6]	
		•••	

Proposed by Max Welling et al. who has a physics background

[1] Satorras, Victor Garcia, Emiel Hoogeboom, and Max Welling. "E (n) equivariant graph neural networks." ICML 2021 [2] Sohl-Dickstein, Jascha, et al. "Deep unsupervised learning using nonequilibrium thermodynamics." ICML 2015. [3] Duane, S. "Kennedy, AD, Pendleton, BJ, and Roweth,

D.(1987), "Hybrid Monte Carlo,"." Physics Letters.

[4] Greydanus, Samuel, Misko Dzamba, and Jason Yosinski. "Hamiltonian neural networks." Advances in neural information processing systems 32 (2019). [5] Banino, Andrea, et al. "Vector-based navigation using grid-like representations in artificial agents." Nature 557.7705 (2018): 429-433. [6] Loynd, Ricky, et al. "Working memory graphs." ICML 2020

(3) Novel tools



Nahuku board





Quantum computing for AI

Neuromorphic computing [1] 1000x more energy efficient

Photonic crystal

up to 1000x faster in computing speed with light

24

[1] Rao, Arjun, et al. "A long short-term memory for AI applications in spike-based neuromorphic hardware." *Nature Machine Intelligence* 4.5 (2022): 467-479.

Why AI + Science?

AI for Science: how AI boosts science

(*) Science solves critical problems
(1) AI significantly boosts the speed and accuracy
(2) AI helps explore vast design/control space
(3) AI helps uncover scientific knowledge

Science for AI: how science boosts AI

(1) Novel challenges(2) Novel insights(3) Novel tools

. . .

. . .

Why AI + Science?

Because it is interesting!

Outline

- Why AI + Science?
- Al for science: important questions, recent advances and relations with machine learning (ML)
 - Al for scientific simulation
 - Al for scientific design
 - Al for scientific discovery
- Science for AI: important questions and recent advances
- Discussion: where to find the next AlphaFold and chatGPT?
 - What is the next big problems for AI + Science?
 - Potential bottlenecks and mid-term roadmap

AI for science: questions, advances and relations with ML



AI for science: questions, advances and relations with ML

- 1. Al for scientific simulation
- 2. Al for scientific design
- 3. Al for scientific discovery











Fundamental across science













AI for science: questions, advances and relations with ML

- 1. Al for scientific simulation
- 2. Al for scientific design
- 3. Al for scientific discovery











Fundamental across science

AI for scientific simulation

Scientific simulation: simulate the dynamics or steady state of the system, given initial state, boundary condition and parameters of the system

Al for scientific simulation: develop machine learning (ML) methods for scientific simulation, improving its **speed** and/or **accuracy**

- ML for simulating dynamics
- ML for simulating steady state

Dynamical system: forward simulation



u^t: original **state** of the system. Can be an infinite-dimensional function u(t, x) as solution to a PDE, or a graph (e.g., mesh, particles, molecules)

 f^* : Evolution. Can be a PDE evolved by classical solver, or evolution in the real world

a: static parameters of the system that does not change with time (e.g., parameters of PDE, spatially varying diffusion coefficient)

 $\partial \mathbb{X}$: boundary condition of the system

Spectrum of methods for simulating dynamics



Classical solvers and limitations

Classical solvers:

Based on Partial Differential Equations (PDEs)

 $\frac{\partial \mathbf{u}}{\partial t} = F(x, \mathbf{u}, \frac{\partial \mathbf{u}}{\partial \mathbf{x}}, \frac{\partial^2 \mathbf{u}}{\partial \mathbf{x}^2}, \dots) \quad \begin{array}{l} \mathbf{u} : \text{ state} \\ \mathbf{x} : \text{ spatial coordinate} \\ t : \text{ time} \end{array}$

Discretize the PDE, then use finite difference, finite element, finite volume, *etc*. to evolve the system.

Pros and challenges:

- **Pros:** (1) Based on first principles and interpretable, (2) accurate, (3) have error guarantee.
- Challenges: Slow and computational expensive, due to
 - (1) Small time interval to ensure numerical stability, or use implicit method.
 - (2) For multi-resolution systems, typically need to resolve to the lowest resolution



Deep learning-based surrogate models

Recently, deep learning based surrogate modeling has emerged as attractive alternative to replace or complement classical solvers. They:

- Offer **speedup** via:
 - Larger spatial resolution
 - Larger time intervals
 - Use explicit forward
 - Better representations

Neural surrogate models

Goal: learn f_{θ} that maps u^t to u^{t+1} :



u^t: original **state** of the system. Can be an infinite-dimensional function u(t, x) as solution to a PDE, or a graph (e.g., mesh, particles, molecules)

 f_{θ} : neural surrogate models (CNN, GNN, neural operators)

a: static parameters of the system that does not change with time (e.g. parameters of PDE, spatially varying diffusion coefficient)

 $\partial \mathbb{X}$: boundary condition of the system
Neural surrogate models

Goal: learn f_{θ} that maps U^t to U^{t+1} and the encoder q_{φ} :



Ut: representation (表示) of the system.

 f_{θ} : neural surrogate models (CNN, GNN, neural operators)

a: static parameters of the system that does not change with time (e.g. parameters of PDE, spatially varying diffusion coefficient)

 $\partial \mathbb{X}$: boundary condition of the system

Neural surrogate models

Challenges:

- Large-scale: have only been applied to relatively small scale (~10k state size) systems, much less than millions or billions of state size in real systems
- Multi-resolution: how to simulate multi-resolution/multi-scale system both accurately and fast
- Long-term prediction accuracy: how can it reduce long-term error accumulation

Spectrum of methods for simulating dynamics





u^t: **state** of the system. Represented as a **graph** (e.g., mesh, particles, molecules) f_{θ} : Graph Neural Network (GNN)

Such graph-structured data is universal across disciplines:



Fluid dynamics, computer graphics



Mesh-based simulation for PDEs



Dislocation in materials



Proteins and small molecules



Graph Network Simulator (GNS) [1] introduced a GNN-based simulator that learns to simulate particle-based systems



GNN Model predicts particle positions and velocity

Predicted simulation after rendering

[1] Sanchez-Gonzalez, Alvaro, et al. "Learning to simulate complex physics with graph networks." *International conference on machine learning*. PMLR, 2020.



Training: trained with 1-step prediction, minimizing MSE. To improve long-term prediction, add Gaussian noise on the input

$$L = \mathbb{E}[(f_{\theta}(\boldsymbol{u}_t + \boldsymbol{\sigma} \cdot \boldsymbol{\epsilon}) - \boldsymbol{u}_{t+1})^2]$$

 $\epsilon \sim N(0, I), \sigma$: amplitude of noise on each feature

Inference: autoregressively rollout for hundreds of steps

Result:





Task: Subsurface fluid simulation (critical in energy, carbon capture, etc.)

Main contribution: Introduced HGNS [1] for fluid simulation, which use

- multi-step prediction during training to improve long-term prediction accuracy
- Sector-based training and inference

Results: Up to 18x faster than classical solver. Apply to 10million cells per step. Deployed in industry



[1] Wu, Tailin, *et al.* "Learning large-scale subsurface simulations with a hybrid graph network simulator." SIGKDD 2022.



Subsurface (consisting of cells, wells, fractures, etc.) 47



Case study: GraphCast

Task: Weather forecasting (mid-range, 10-day)

Main contribution: Introduced GraphCast [1]:

- Multi-scale GNN
- Annealed multi-step learning objective

Results: outperforms state-of-the-art weather forecasting method (HRES) in 10-day prediction acc.





Prediction by GraphCast

[1] Lam, Remi, et al. "GraphCast: Learning skillful medium-range global weather forecasting." *arXiv preprint arXiv:2212.12794* (2022).



Case study: MeshGraphNets

Task: Mesh-based simulation

Main contribution: Introduced MeshGraphNets [1]:

- World-space edges
- Supervised remeshing

Results: accurate prediction on many different systems.







Example predictions

[1] Pfaff, Tobias, et al. "Learning mesh-based simulation with graph networks." *ICLR* 2021





[1] Satorras, Victor Garcia, Emiel Hoogeboom, and Max Welling. "E (n) equivariant graph neural networks." ICML, 2021.

[2] Zhang, Linfeng, et al. "Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics." *Physical review letters* 120.14 (2018): 143001.



Challenges:

- Large-scale
- Multi-resolution
- Long-term accuracy

u^t: **state** of the system. Represented as a **graph** (e.g., mesh, particles, molecules) f_{θ} : Graph Neural Network (GNN)

Such graph-structured data is universal across disciplines:



Fluid dynamics, computer graphics



Mesh-based simulation for PDEs



Dislocation in materials



Proteins and small molecules

Case study: neural operator-based simulation



Challenges:

- Large-scale
- **Multi-resolution**
- Long-term accuracy

u^t: state of the system. Represented as a infinite-dimensional function u(x) or u(x, t).

 f_{θ} : Neural operator

Benefits:

- Mesh-free ٠
- Allow super-resolution ٠

Main categories:

- Fourier Neural Operator (FNO) [1]
- DeepONet [2]
- **Physics Informed Neural Network** (PINN) [3]



[1] Li, Zongyi, et al. "Fourier neural operator for parametric partial differential equations." ICLR 2021 [2] Lu, Lu, et al. "Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators." Nature machine intelligence 3.3 (2021): 218-229. [3] Raissi et al., Journal of Computational physics 378 (2019): 686-707

Case study: neural operator-based simulation



Challenges:

- Large-scale
- Multi-resolution
- Long-term accuracy

u^t: **state** of the system.

 f_{θ} : learned neural surrogate models

Other architectures are possible:

- Transformers
- CNNs, UNets
- Latent evolution
- Neural ODEs
- Solver-in-the-loop
- ...(your new invention)

ML techniques:

- Supervised learning
- Representation learning
- Reinforcement learning
- Diffusion models
- Uncertainty quantification and active learning
- Generalization bounds and certifiable prediction

AI for scientific simulation

Scientific simulation: simulate the dynamics or steady state of the system, given initial state, boundary condition and parameters of the system

Al for scientific simulation: develop machine learning (ML) methods for scientific simulation, improving its **speed** and/or **accuracy**

- ML for simulating dynamics
- ML for simulating steady state

AI for scientific simulation

Scientific simulation: simulate the dynamics or steady state of the system, given initial state, boundary condition and parameters of the system

Al for scientific simulation: develop machine learning (ML) methods for scientific simulation, improving its **speed** and/or **accuracy**

- ML for simulating dynamics
- ML for simulating steady state

Case study: neural operator-based simulation



Example 1: second order elliptic PDE (e.g., Poisson's equation):

$$-\nabla \cdot (a(x)\nabla u(x)) = f(x), \quad x \in D$$
$$u(x) = 0, \qquad x \in \partial D$$

Important in materials, plasma physics, elasticity, hydrology.

Techniques:

- Neural operators (e.g., FNO [1], GKN [2]), PINNs [3]
- GNNs

[1] Li, Zongyi, et al. "Fourier neural operator for parametric partial differential equations." *ICLR 2021* [2] Li, Zongyi, et al. "Neural operator: Graph kernel network for partial differential equations." arXiv preprint arXiv:2003.03485 (2020).

[3] Raissi et al., Journal of Computational physics 378 (2019): 686-707

Case study: neural operator-based simulation



Example 2: protein folding, polymer simulation

Techniques:

- AlphaFold 2 [1]
- GNNs
- ...

[1] Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." Nature 596.7873 (2021): 583-589.

AI for scientific simulation: broader impact

- A. Digital twins (数字孪生)
- B. Meta-verse (元宇宙)



Digital twins: can significantly boost the efficiency, safety, and fast iteration across **industry**



Meta-verse: creates a new digital universe

AI for science: questions, advances and relations with ML

- 1. Al for scientific simulation
- 2. Al for scientific design
- 3. Al for scientific discovery













AI for science: questions, advances and relations with ML

- 1. Al for scientific simulation
- 2. Al for scientific design
- 3. Al for scientific discovery

























u^t: original **state** of the system. Can be an infinite-dimensional function u(t,x) as solution to a PDE, or a graph (e.g., mesh, particles, molecules)

 f_{θ} : neural surrogate models (CNN, GNN, neural operators)

a: static parameters of the system that does not change with time (e.g., parameters of PDE, spatially varying diffusion coefficient)

 $\partial \mathbb{X}$: boundary condition of the system

Al for scientific design:

• boundary condition ∂X : plane shape, rocket shape, underwater robot shape Objective







 $\mathbf{a},\partial\mathbb{X}$

u¹

 $\mathbf{a},\partial\mathbb{X}$

 $\mathbf{a},\partial\mathbb{X}$

fθ

u²

 $\mathbf{a},\partial\mathbb{X}$

 L_d

• system itself u^0 , a: design proteins, small molecules, materials; or state estimation



• external control: control pulses for controlled nuclear fusion, revert global warming





Method categories:

- 1. (Learned) simulator as an inner loop
- Physical design with GNS [1]





- 2. Iterative convergence
- Diffusion models (e.g., GeoDiff [2]), energy-based models



- 3. Direct mapping
- Transformer for boundary value inverse problems [3]



Allen, Kelsey R., et al. "Physical design using differentiable learned simulators." NeurIPS 2022
 Xu, Minkai, et al. "Geodiff: A geometric diffusion model for molecular conformation generation." ICLR 2022
 Guo, Ruchi, Shuhao Cao, and Long Chen. "Transformer meets boundary value inverse problems." NeurIPS 2022

Case study: physical design with differentiable learned simulators

Task: design boundary for particle-based fluid simulation

Main contribution:

Backpropagation through the entire GNN-based simulation

Results: able to design various boundaries, outperforms sample-based traditional methods





Method categories:

4. Reinforcement learning (RL) for designing the system:



GCPN [1]

GFlowNet [2]

[1] You, Jiaxuan, et al. "Graph convolutional policy network for goal-directed molecular graph generation." Advances in neural information processing systems 31 (2018).
[2] Jain, Moksh, et al. "Biological sequence design with gflownets." ICML 2022

Method categories:4. RL for controlling the system:



Controlling nuclear fusion plasma [1]

[1] Degrave, Jonas, et al. "Magnetic control of tokamak plasmas through deep reinforcement learning." *Nature* 602.7897 (2022): 414-419.

Method categories:
4. RL for designing experiment:



Design quantum computing experiment

[1] Nautrup, Hendrik Poulsen, et al. "Optimizing quantum error correction codes with reinforcement learning." Quantum 3 (2019): 215.

Related machine learning techniques:

- Reinforcement learning
- Diffusion models
- Graph Neural Networks
- Active learning
- Representation learning
- ...

AI for scientific design: broader impact

- A. Digital twins (数字孪生)
- B. Aerospace, energy, materials, etc.



Digital twins: can significantly boost the efficiency, safety, and fast iteration across **industry**



Aerospace, energy, materials, etc.
AI for science: questions, advances and relations with ML

- 1. Al for scientific simulation
- 2. Al for scientific design
- 3. Al for scientific discovery























Galileo and the Lamp

As a student, Galileo famously observed a lamp swinging in Pisa Cathedral and timed its swing against his pulse.

He concluded: the lamp's period was constant and independent of its amplitude.

Galileo then suggested a pendulum could control a clock.

Key components of the scientific discovery process:

- 1. Focus on the relevant concept
- 2. Simplicity
- 3. Generalization to novel concepts
- 4. Transfer this knowledge to other human.

Based on this scientific discovery process, we **developed complex theories**, e.g., theory of light, gravity, quantum theory, central dogma.



Lamp in Pisa Cathedral



young Galileo

AI for scientific discovery



AI for scientific discovery: 1. Concept identification & representation

- Discover and represent important **aspect** of the system, e.g.,
 - Cell type discovery and universal cell embeddings
 - Rare-event selection in particle physics
- Discover the structure and causal relations in a system
 - Causal learning
 - Structure learning
- Discover important invariances of the system [1]

Al for scientific discovery: 2. theory learning



Al for scientific discovery: 2. theory learning

• Discover interpretable, universal equations on the relevant concepts

Techniques:

- Recursion and divide and conquer [1]
- AI Physicist [2]
- Reinforcement learning [3]

[1] Udrescu, Silviu-Marian, et al. "AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity." *NeurIPS 2020*[2] Wu, Tailin, and Max Tegmark. "Toward an artificial intelligence physicist for unsupervised learning." Physical Review E 100.3 (2019): 033311.
[3] Mundhenk, Terrell, et al. "Symbolic regression via deep reinforcement learning enhanced genetic programming seeding." *NeurIPS* 2021

Al for scientific discovery: 2. theory learning case study

AI Feynman 2.0 [1]: rediscover top-100 **physics equations** in Feynman lectures Some equations discovered by AI Feynman 2.0:

	Equation	Symmetries
1	$\delta = -5.41 + 4.9 \frac{\alpha - \beta + \gamma/\chi}{3\chi}$	TC
2	$\chi = 0.23 + 14.2 \frac{\alpha + \beta}{3\alpha}$	TS
3	$\beta = 213.80940889'(1 - e^{-0.54723748542\alpha})$	
4	$\delta = 6.87 + 11 \sqrt{lpha eta \gamma}$	Р
5	$V = \left[R_1^{-1} + R_2^{-1} + R_3^{-1} + R_4^{-1} \right]^{-1} I_0 \cos \omega t \text{(Parallel resistors)}$	PGSM
6	$I_0 = \frac{V_0}{\sqrt{R^2 + (\omega L - \frac{1}{\omega C})^2}} \text{(RLC circuit)}$	MG
7	$I = \frac{\sqrt{V_0 \cos \omega t}}{\sqrt{R^2 + (\omega L - \frac{1}{\omega C})^2}} \text{(RLC circuit)}$	MG
8	$V_2 = \left(\frac{R_2}{R_1 + R_2} - \frac{R_2}{R_x + R_3}\right) V_1$ (Wheatstone bridge)	SGMA
9	$v = c \frac{(v_1 + v_2 + v_3)/c + v_1 v_2 v_3/c^3}{1 + (v_1 v_2 + v_1 v_3 + v_2 v_3)/c^2}$ (Velocity addition)	AG
10	$v = c \frac{(v_1 + v_2 + v_3 + v_4)/c + (v_2 v_3 v_4 + v_1 v_3 v_4 + v_1 v_2 v_4 + v_1 v_2 v_3)/c^3}{1 + (v_1 v_2 + v_1 v_3 + v_1 v_4 + v_2 v_3 + v_2 v_4 + v_3 v_4)/c^2 + v_1 v_2 v_3 v_4/c^4}$ (Velocity addition)	GA
11	$z = (x^4 + y^4)^{1/4}$ (L ₄ -norm)	AC
12	$w = xyz - z\sqrt{1 - x^2}\sqrt{1 - y^2} - y\sqrt{1 - x^2}\sqrt{1 - z^2} - x\sqrt{1 - y^2}\sqrt{1 - z^2}$	GA
13	$z=rac{xy+\sqrt{1-x^2-y^2+x^2y^2}}{y\sqrt{1-x^2}-x\sqrt{1-y^2}}$	А
14	$z = y\sqrt{1-x^2} + x\sqrt{1-y^2}$	Α
15	$z = xy - \sqrt{1 - x^2}\sqrt{1 - y^2}$	A
16	$r = \frac{a}{\cot(\alpha/2) + \cot(\beta/2)}$ (Incircle)	GMAC

[1] Udrescu, Silviu-Marian, et al. "AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity." *NeurIPS 2020*

Training data: $\{(x, y)\}$ **Goal:** discover symbolic function $f: x \rightarrow y$



It detects symmetry from first training a neural net

Al for scientific discovery: 3. concept generalization



Al for scientific discovery: 3. concept generalization

• Generalize to more complex concept in inference

Techniques:

- Energy-based models [1][2]
- Few-shot learning [3]
- In-context learning [4] with large-language models

[1] Du, Yilun, Shuang Li, and Igor Mordatch. "Compositional visual generation with energy based models." *NeurIPS 2020*[2] Wu, Tailin, et al. "Zeroc: A neuro-symbolic model for zero-shot concept recognition and acquisition at

inference time." NeurIPS 2022

[3] Cao, Kaidi, Maria Brbic, and Jure Leskovec. "Concept learners for few-shot learning." ICLR 2021

[4] Brown, Tom, et al. "Language models are few-shot learners." NeurIPS 2020

Outline

- Why AI + Science?
- Al for science: important questions, recent advances and relations with machine learning
 - Al for scientific simulation
 - Al for scientific design
 - Al for scientific discovery
- Science for AI: important questions and recent advances
- Discussion: where to find the next AlphaFold and chatGPT?
 - What is the next big problems for AI + Science?
 - Potential bottlenecks and mid-term roadmap

Outline

- Why AI + Science?
- Al for science: important questions, recent advances and relations with machine learning
 - Al for scientific simulation
 - Al for scientific design
 - Al for scientific discovery

• Science for AI: important questions and recent advances

- Discussion: where to find the next AlphaFold and chatGPT?
 - What is the next big problems for AI + Science?
 - Potential bottlenecks and mid-term roadmap

Science for AI: how science brings new insights to AI

Famous concepts:

Concepts	Source	Application
Symmetry	physics	equivariant GNN [1]
Diffusion	physics	Diffusion models [2]
Energy	physics	Energy-based models, Boltzmann machines
Hamiltonian	physics	Hamiltonian Monte Carlo [3], Hamiltonian Neural Networks [4]
Grid cells	neuroscience	Grid cells for navigation [5]
Working memory	neuroscience	Working Memory Graphs [6]

 Satorras, Victor Garcia, Emiel Hoogeboom, and Max Welling. "E
 (n) equivariant graph neural networks." ICML 2021
 Sohl-Dickstein, Jascha, et al. "Deep unsupervised learning using nonequilibrium thermodynamics." *ICML 2015*.
 Duane, S. "Kennedy, AD, Pendleton, BJ, and Roweth, D.(1987), "Hybrid Monte Carlo,"." *Physics Letters*.

[4] Greydanus, Samuel, Misko Dzamba, and Jason Yosinski. "Hamiltonian neural networks." *Advances in neural information processing systems* 32 (2019).
[5] Banino, Andrea, et al. "Vector-based navigation using grid-like representations in artificial agents." *Nature* 557.7705 (2018): 429-433.
[6] Loynd, Ricky, et al. "Working memory graphs." *ICML* 2020

Science for AI: how science brings new insights to AI

Famous concepts:

Concepts	Source	Application
Fourier transform	math	FNO [1]
Path integral	physics	Path integral-based GNN [2]
Mean field theory	physics	Mean Field Multi-Agent Reinforcement Learning [3], How to Train 10,000-Layer Vanilla CNNs [4]
Phase transitions	physics	Phase transitions in the information bottleneck [5]
Optimal transport	math	W-GAN [6]
Arrow of time	physics	Learning the arrow of time for RL [7]

[1] Li, Zongyi, et al. "Fourier neural operator for parametric partial differential equations." ICLR 2021
[2] Ma, Zheng, et al. "Path integral based convolution and pooling for graph neural networks." *NeurIPS* 2020
[3] Yang, Yaodong, et al. "Mean field multi-agent reinforcement learning." *ICML* 2018
[4] Xiao, Lechao, et al. "Dynamical isometry and a mean field theory of cnns: How to train 10,000-layer vanilla convolutional neural networks." ICML 2018
[5] Wu, Tailin, and Ian Fischer. "Phase transitions for the information bottleneck in representation learning." *ICLR* 2020
[6] Arjovsky, 2017
[7] Rahaman, Nasim, et al. "Learning the arrow of time for problems in reinforcement learning." ICLR 2020

85

Science for AI: how science brings new insights to AI

Physics-inspired generative models:

- Diffusion models
- Energy-based models
- Electrodynamics-based models
- Quantum generative models



Poisson Flow Generative Models [1]

Physics-inspired learning theory:

- Phase transitions
- Particle interactions
- Field theory



[1] Xu, Yilun, et al. "Poisson flow generative models." *NeurIPS* 2022
[2] Liu, Ziming, et al. "Towards understanding grokking: An effective theory of representation learning." *Advances in Neural Information Processing Systems* 35 (2022): 34651-34663.

Understanding grokking [2]

Ad

I will join Westlake University (西湖大学) as an assistant professor in the engineering department, establishing AI + Science group

Research directions:

- AI for scientific simulation and design
 - For fluid, materials, biomedicine, ME
- Al for scientific discovery
 - For biomedicine
- Representation learning
 - With GNNs

Have close collaborations with Stanford CS, EE, ME and MIT CS

Please reach out to me (tailin@cs.stanford.edu) for

- Collaboration
- Postdoc, PhD, internship, visiting scholar positions



Outline

- Why AI + Science?
- Al for science: important questions, recent advances and relations with machine learning
 - Al for scientific simulation
 - Al for scientific design
 - Al for scientific discovery
- Science for AI: important questions and recent advances
- Discussion: where to find the next AlphaFold and chatGPT?
 - What are the next big problems for AI + Science?
 - Potential bottlenecks and mid-term roadmap

In 2019, when Demis Hassabis (CEO of DeepMind) gave a talk at MIT, he said that

• "The year 2019 is a watershed moment for DeepMind, where the full company turns its focus on AI for Science."

89

Now, we see that DeepMind has made exciting progress:

• AlphaFold 2

. . .

- RL for controlling fusion plasma
- AI for Mathematics

Criteria:

- 1. Universal and impactful: whose solution can enable the solutions for tens or hundreds of problems
- 2. Ambitious but still feasible within 2-3 years
- 3. Have enough data
- 4. Have a **clear** evaluation objective

Both AlphaFold and chatGPT satisfy the above criteria

Model

My idea:

A model & platform for general **inverse design** for engineering

Specification of objective

(in text, function, etc.)

Can be used across science and engineering

91







My idea:

A model & platform for general **inverse design** for engineering

Specification of objective

(in text, function, etc.)

Model





Does it satisfy the criteria?

- **1. Universal and impactful:** whose solution can enable the solutions for tens or hundreds of problems
- 2. Ambitious but still feasible within 2-3 years
- 3. Have enough data
- 4. Have a **clear** evaluation objective



Discussion: potential bottlenecks for AI + Science

- Easy to use **data** and **environment** for training and evaluation
 - ImageNet for image classification, OGB for GNNs, OpenAI Gym for RL
 - AI + Science community needs several these easy-to-use data/environment:
 - Great opportunity!
- Domain knowledge
- Computation
- Trustworthiness for model:
 - Uncertainty quantification
 - Generalization bounds
- Our imagination!

Solution:

- More open-discussions
- Cross-university/industry collaborations (build an ecosystem)
- ...

Outline

- Why AI + Science?
- Al for science: important questions, recent advances and relations with machine learning
 - Al for scientific simulation
 - Al for scientific design
 - Al for scientific discovery
- Science for AI: important questions and recent advances
- Discussion: where to find the next AlphaFold and chatGPT?
 - What is the next big problems for AI + Science?
 - Potential bottlenecks and mid-term roadmap