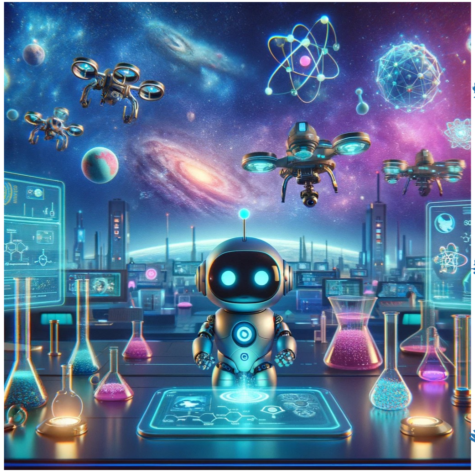
UCLA Samueli
Computer Science

Uncertainty-Aware Unsupervised and Robust Reinforcement Learning

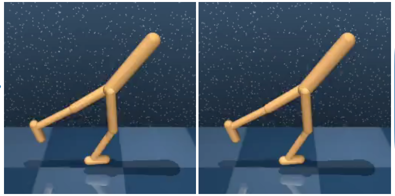
\

Unsupervised Data Collection & Exploration

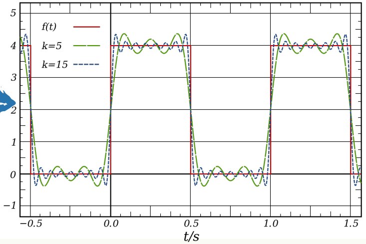
Science of Decision Making & Decision Making for Science



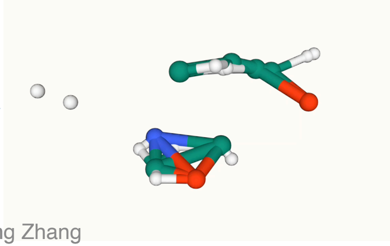
Endeavor: Decision making for scientific discoveries



Unsupervised data collection and exploration in reinforcement learning
[NeurIPS’21; ICML’23, ‘24]



Robust reinforcement learning under model error / misspecification
[ICML’23]

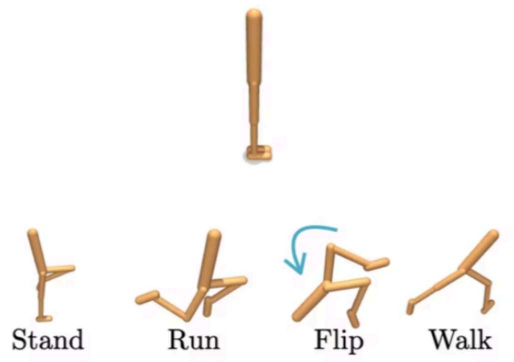


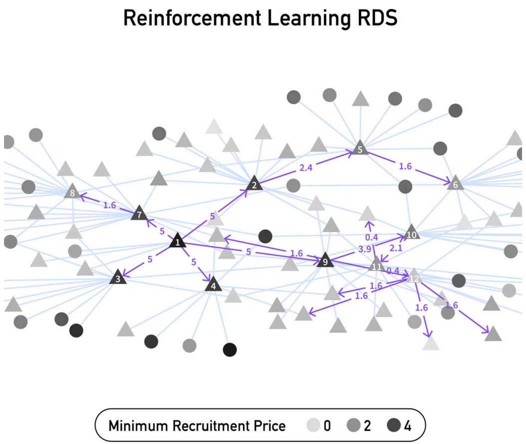
Frontier models / decision making for scientific tasks and drug design
[ACS Meas.Au’22, Nat. Comm.’24, etc.]

Unsupervised Data Collection & Exploration

REWARD-FREE EXPLORATION IN REINFORCEMENT LEARNING

Unsupervised RL — Explore without supervision





Multi-task robotics

- Explore and learn physics
- Execute the desired motion

Search engine (GPT4+Bing)

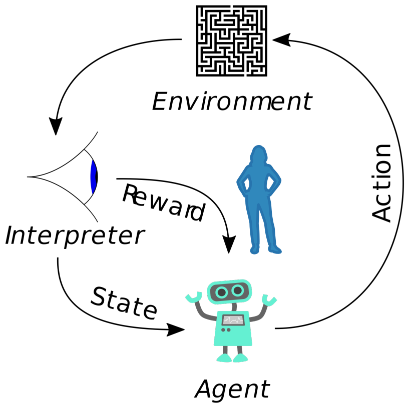
- Learn how to search result
- Search for specific result

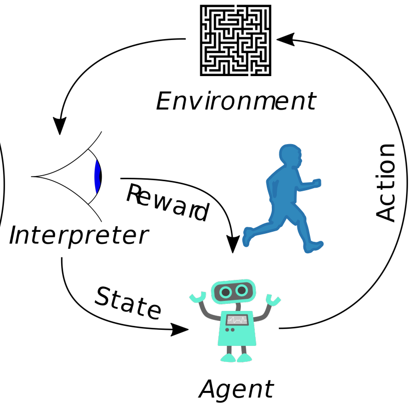
Field research for public health

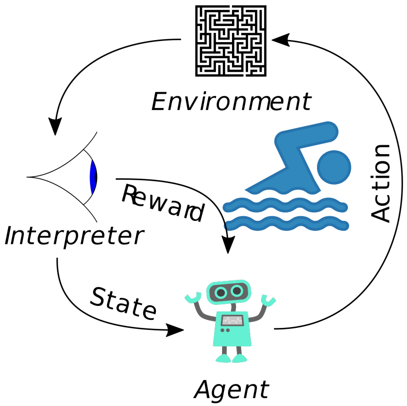
- Explore different groups
- Gain as much information as possible

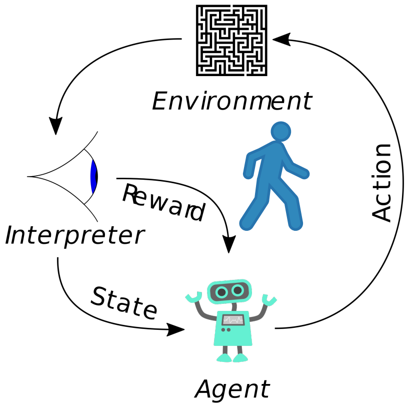
Motivation: Efficient exploration for various tasks

SUPERVISED REINFORCEMENT LEARNING — CONCERNS AND ISSUES





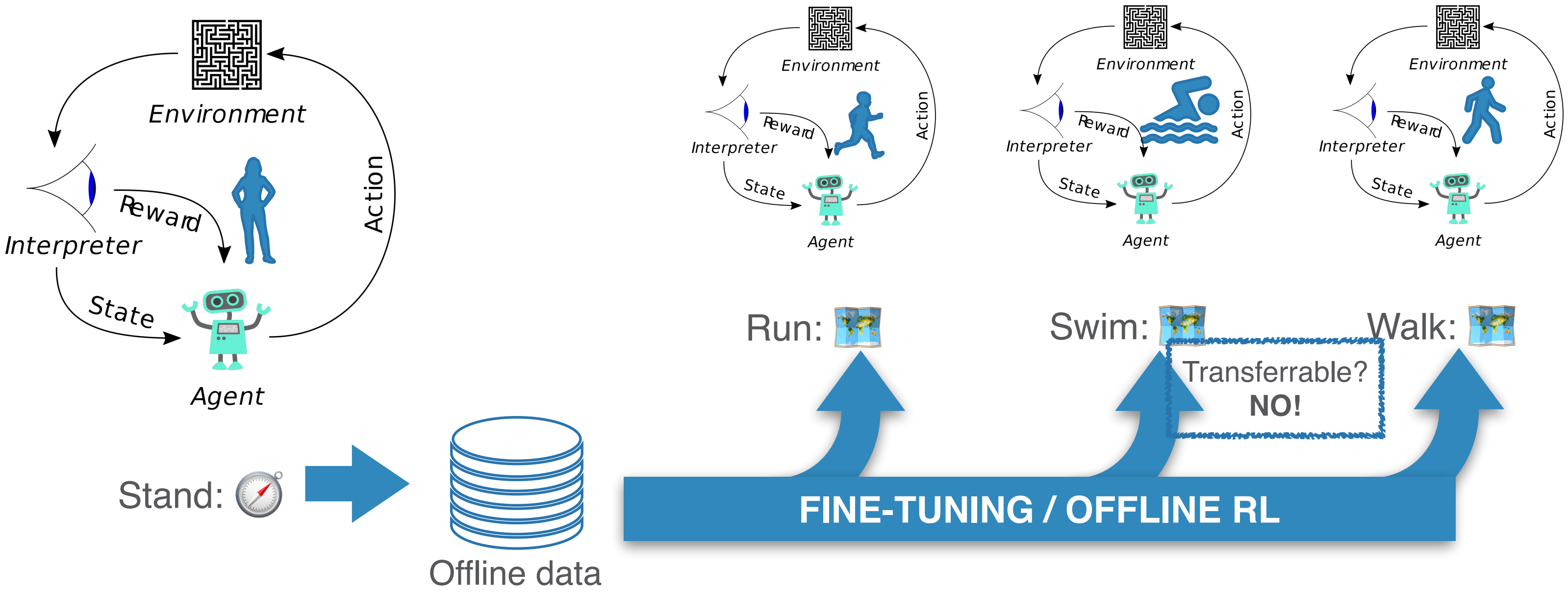




Stand: 🧭 Run: 🧭 Swim: 🧭 Walk: 🧭

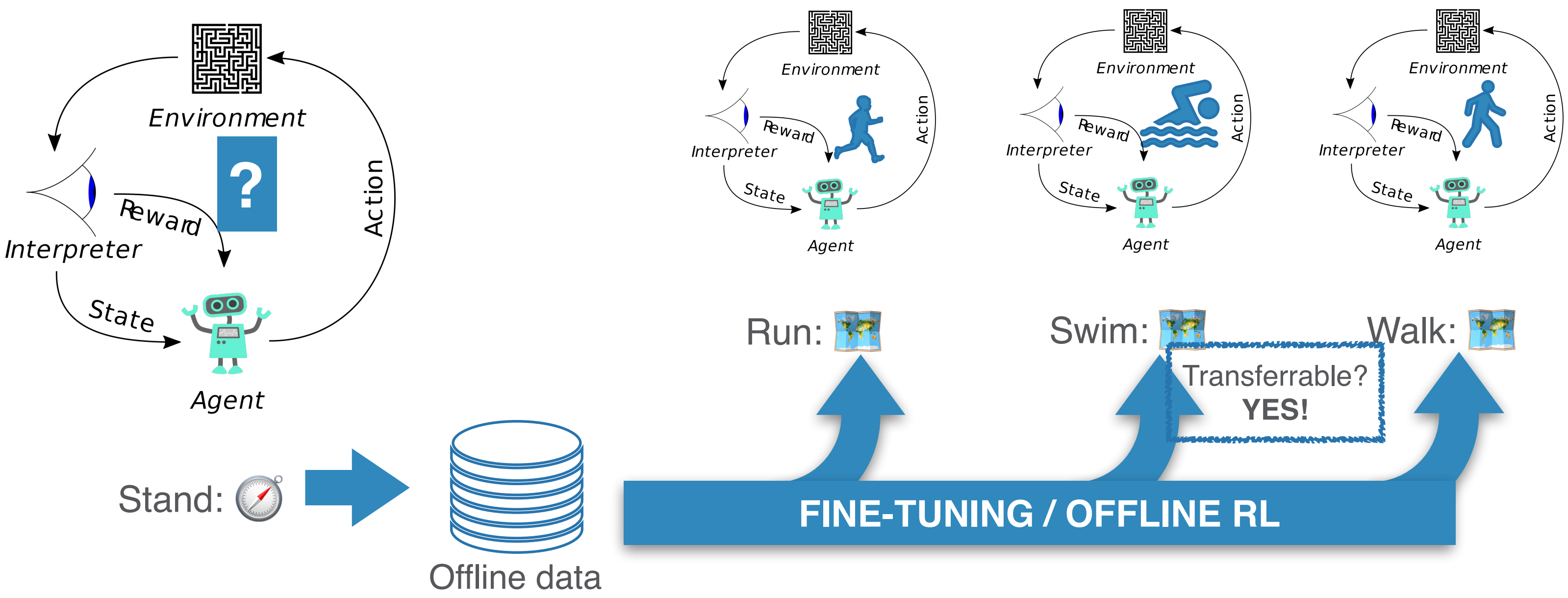
Efficient exploration for various tasks

OFFLINE RL WITH SUPERVISED DATA COLLECTION …



Unsupervised RL: Exploration for various tasks

DESIGNING UNSUPERVISED EFFICIENT EXPLORATION POLICY



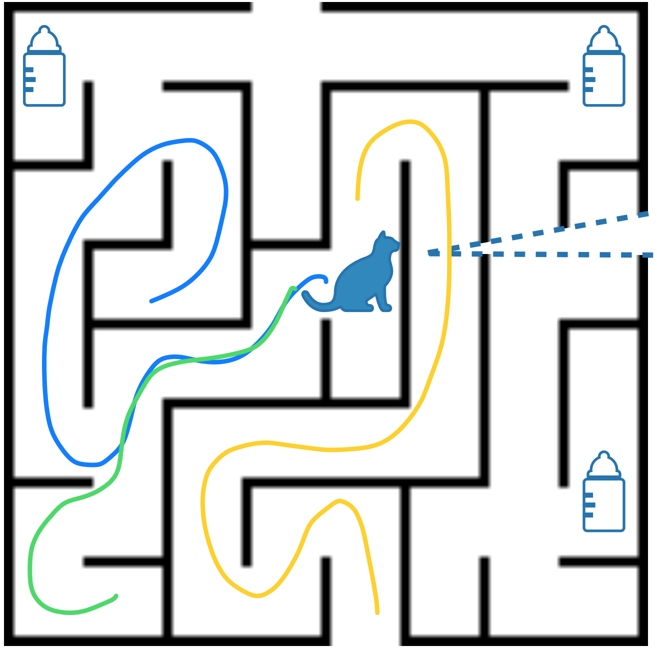
🤔 How to efficiently explore the environments without supervision?

🤓 Foundation of unsupervised RL for both practice and analysis!

[ZZG, NeurIPS'21]; [ZZG, ICML'23]; [ZZZG, ICML'24]

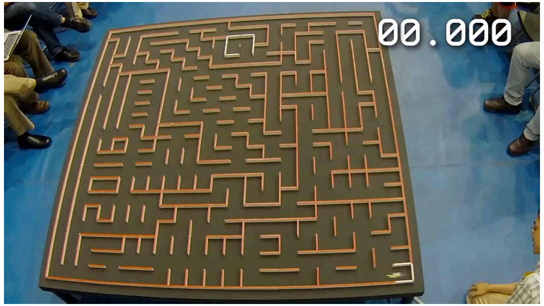
Designing efficient exploration policy

INTUITION — UNCERTAINTY AS REWARDS



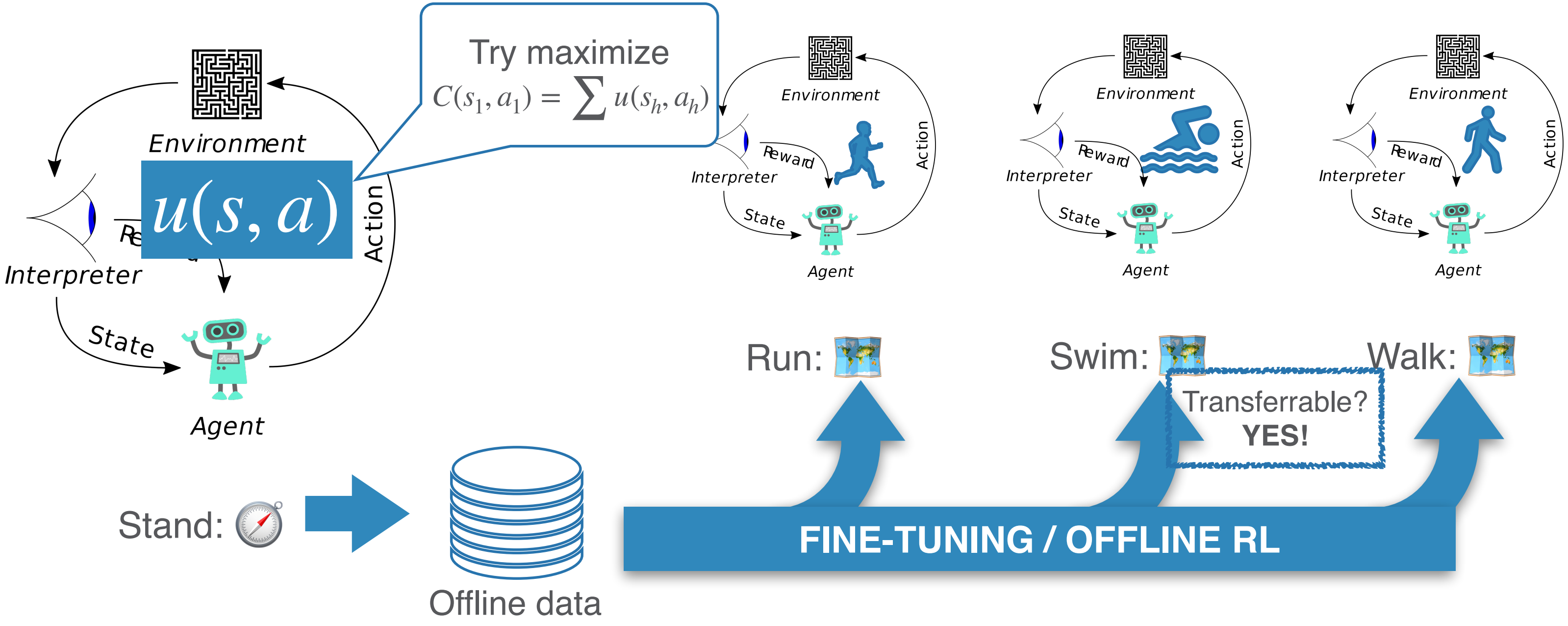
I should explore spaces that I'm **unfamiliar** with!

- Uncertainty / Curiosity
- Empirical Unsupervised RL:
  - Intrinsic reward [PA+'17]



Leveraging uncertainty for unsupervised RL

UNCERTAINTY AS PSEUDO REWARD FUNCTION [ZZG21]



Detour: How to determine uncertainty?

THEORETICAL FRAMEWORK

Function class $\mathscr{F}$ for approximating…

- State transition $P(s' \mid s, a)$
- Value function $Q(s, a) = \sum_h r(s_h, a_h)$

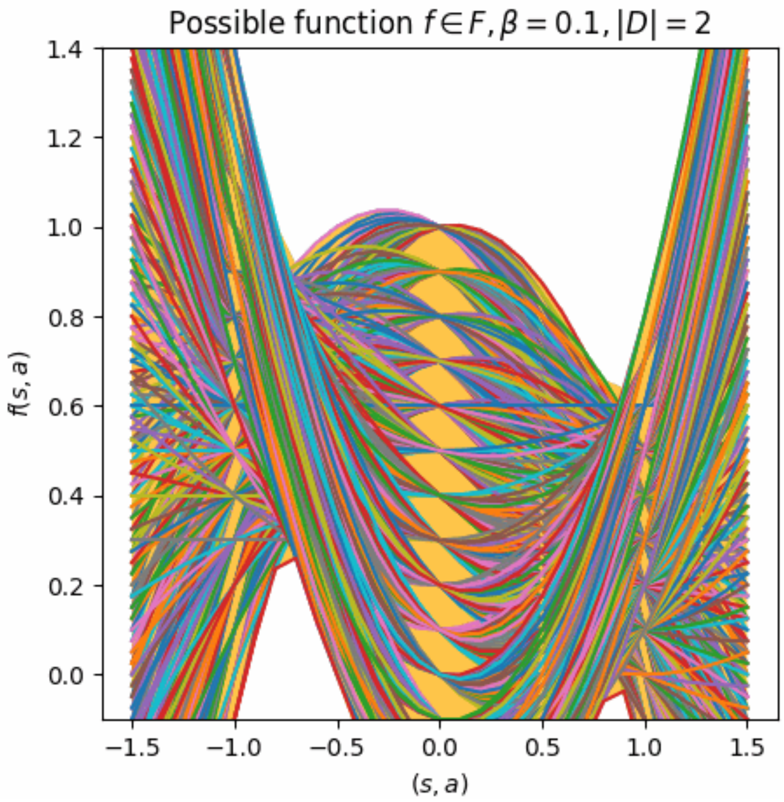
On historical dataset $\mathscr{D} = \{(s_i, a_i)\}$:

$$u(s, a) = \max_{f_1, f_2 \in \mathscr{F}} \left(f_1(s, a) - f_2(s, a)\right)^2$$

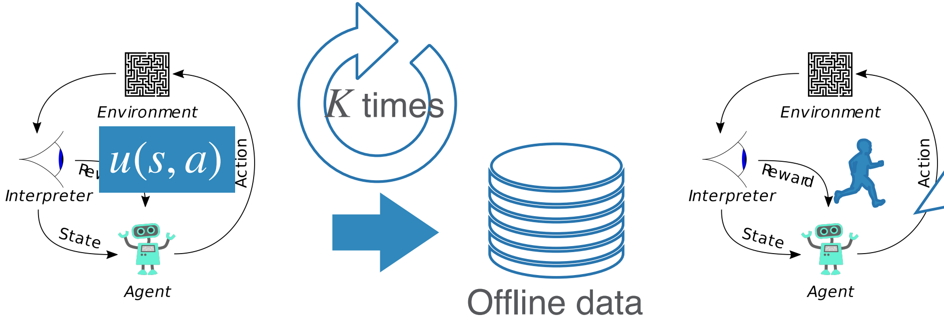
(radius of set)

$$\text{s.t.} \sum_{(s_i, a_i) \in \mathscr{D}} \left(f_{1,2}(s_i, a_i) - f^*(s_i, a_i)\right)^2 \le \beta$$

(well trained functions)



UCRL-RFE — Uncertainty as intrinsic reward



How many episodes K do we need for in exploration phase to get near-optimal policy in planning phase?

Unknown reward
Exploration phase

Any reward function
Planning phase

Theorem [ZZG21]. For UCRL-RFE algorithm, for any $0 < \epsilon < 1$, if $K = \tilde{O}(H^5 d^2 \epsilon^{-2})$ episodes are collected during exploration phase, then with high probability, for **any** reward function r, we can output a policy π such that $\mathbb{E}_s \left[V_1^*(s; r) - V_1^\pi(s; r) \right] \leq \epsilon$ in planning phase.

Theoretical results — Unsupervised RL

Theorem [ZZG21]. For UCRL-RFE algorithm, for any $0 < \epsilon < 1$, if $K = \tilde{O}(H^5 d^2 \epsilon^{-2})$ episodes are collected during exploration phase, then with high probability, for **any** reward function r, we can output a policy π such that $\mathbb{E}_s \left[V_1^*(s; r) - V_1^\pi(s; r) \right] \leq \epsilon$ in planning phase.

$V_1^\pi(s; r)$: Expected cumulative reward get from policy π

$$V_1^\pi(s; r) = \mathbb{E}\left[\sum_{h=1}^{H} r(s_h, a_h) \,\middle|\, \pi \right]$$

$V_1^*(s; r) = \max_\pi V_1^\pi(s; r)$: Maximum cumulative reward from optimal policy

H: length of decision process

$s_1, a_1, s_2, a_2, \cdots s_H, a_H$

e.g. At most $H = 100$ steps in Maze

d: dimension of features

$\phi(s, a, s') \in \mathbb{R}^d$

ϵ: precision of planning (most important)

No #state required!

AlphaGo:

$S \geq 10^{360}, d = 19 \times 19$

Discussion: Foundation of unsupervised RL

PSEUDO REWARD IS INTRINSIC REWARD [PA+17]

r_{int}: intrinsic reward — motivation, curiosity $\Leftrightarrow$ r_{ext}: extrinsic reward — target, goal

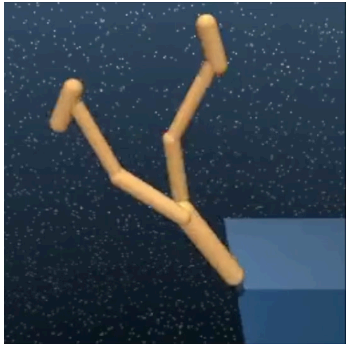
Exploration policy: $\pi = \arg\max_{\pi} V_1^{\pi}(s; r_{\text{int}})$

$$r_{\text{int}}(s,a) = \max_{f_1,f_2 \in \mathcal{F}} \left(f_1(s,a) - f_2(s,a)\right)^2$$
$$\text{s.t.} \sum_{(s_i,a_i)\in\mathcal{D}} \left(f_{1,2}(s_i,a_i) - f^*(s_i,a_i)\right)^2 \leq \beta$$

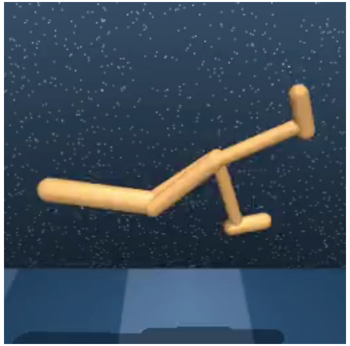
Various intrinsic rewards in unsupervised RL [Las+21]

Name	Intrinsic reward	Translation
ICM [PA+'17]	$\lVert f(s_{t+1} \mid s_t, a_t) - s_{t+1} \rVert_2^2$	$f_2(s,a) = f^*(s,a)$
Disagreement [PG+'19]	$\text{Var}[f_i(s_{t+1} \mid s_t, a_t)]_i$	Variance as radius
RND [BE+'18]	$\lVert f_1(s_t,a_t) - f_2(s_t,a_t) \rVert_2^2$	Only two function candidates

Experiments — Multi-task robotics



DeepMind Control Robotics:
Exploration: 1M frames, no reward
Only 10% of offline RL benchmarks!
(D4RL: 10M frames, expert agent)



Exploration (3, 2x speed)

Walk

Cumulative rewards (std) for various tasks

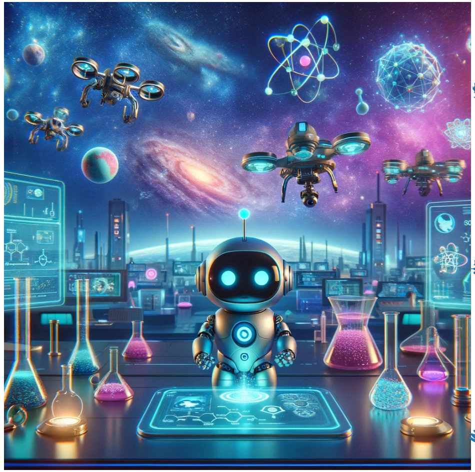
Task	ICM [PA+'17]	Disagreement [PG+'19]	RND [BE+'18]	Ours
Walk	411 (237)	851 (63)	709 (115)	826 (89)
Stand	466 (17)	726 (79)	750 (62)	925 (50)
Run	108 (41)	340 (37)	306 (34)	339 (64)

Run

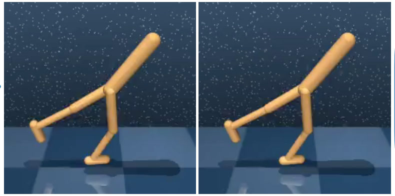
Stand

Uncertainty-aware curiosity helps exploration without supervision

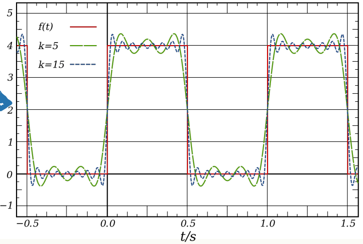
Misspecification-Robust Decision Making



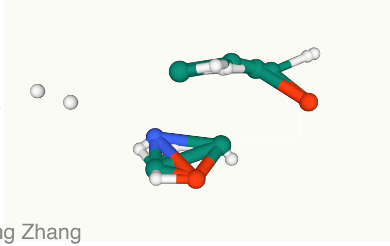
Endeavor: Decision making for scientific discoveries



Unsupervised data collection and exploration in reinforcement learning
[NeurIPS’21; ICML’23, ‘24]



Robust reinforcement learning under model error / misspecification
[ICML’23]

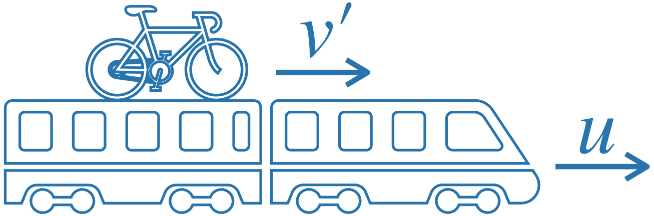


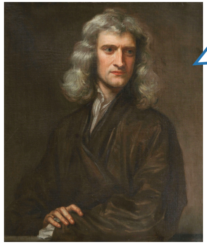
Frontier models / decision making for scientific tasks and drug design
[ACS Meas.Au’22, Nat. Comm.’24, etc.]

Misspecification-Robust Decision Making

REINFORCEMENT LEARNING WITH MODEL MISSPECIFICATION

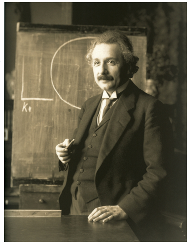
Model Misspecification Always Exists…



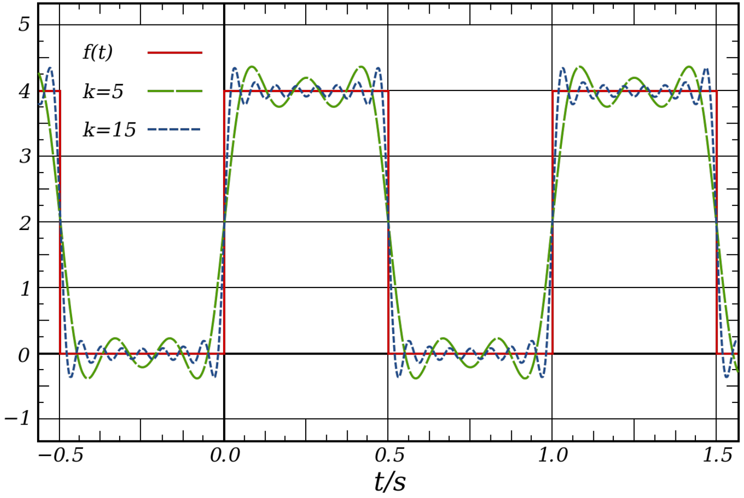


$$v = v' + u?$$

$$v = \frac{v'u}{1 + \frac{v'u}{c^2}}!$$

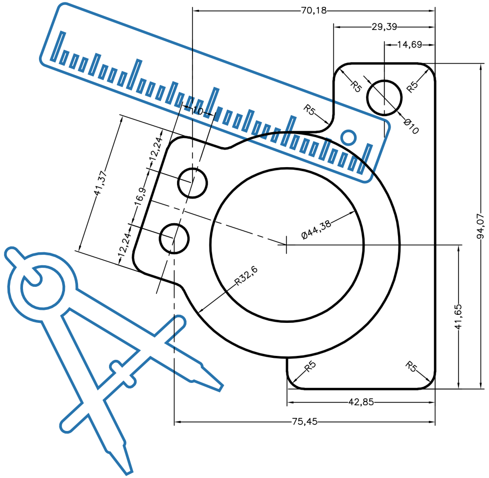


Model error, Physic laws, etc..



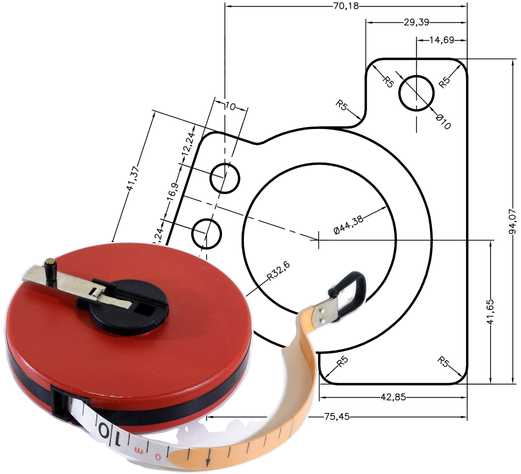
Function approximations, Neural networks

Will the model misspecification affect decisions?



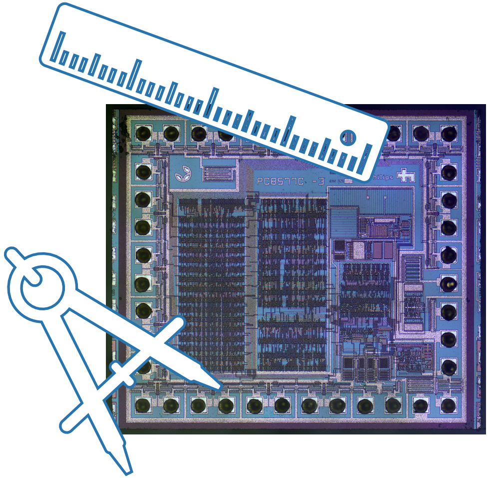
Draw a blueprint with ruler and compass?

Good match! 🤓



Draw a blueprint with a tape measure?

Too large misspecification! 🤔

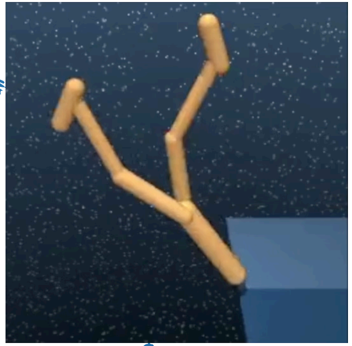


Design a IC with ruler and compass?

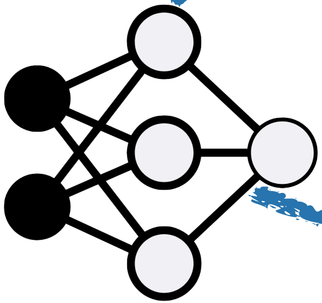
Very high precision! 🤔

Model misspecification in Reinforcement Learning

Environment



Exploration
Learn a robust estimator



Exploitation
Use the estimator properly

Misspecified function $f(x)$

🤔 What's the relationship between misspecification & precision in RL?

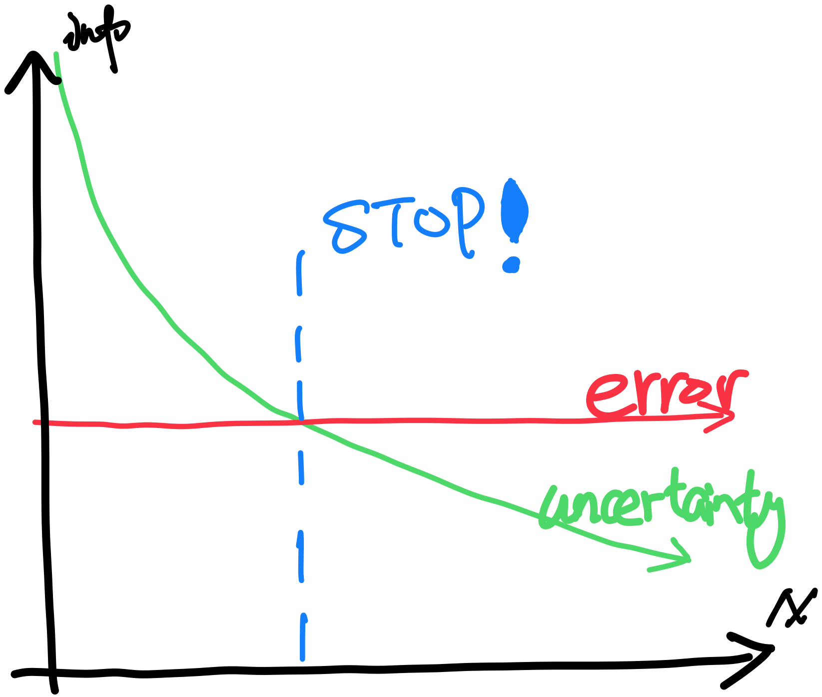
🤓 The interplay between misspecification & "precision"

[ZHFG, ICML'23; ZFHG, 24]

Learning a proper function approximation

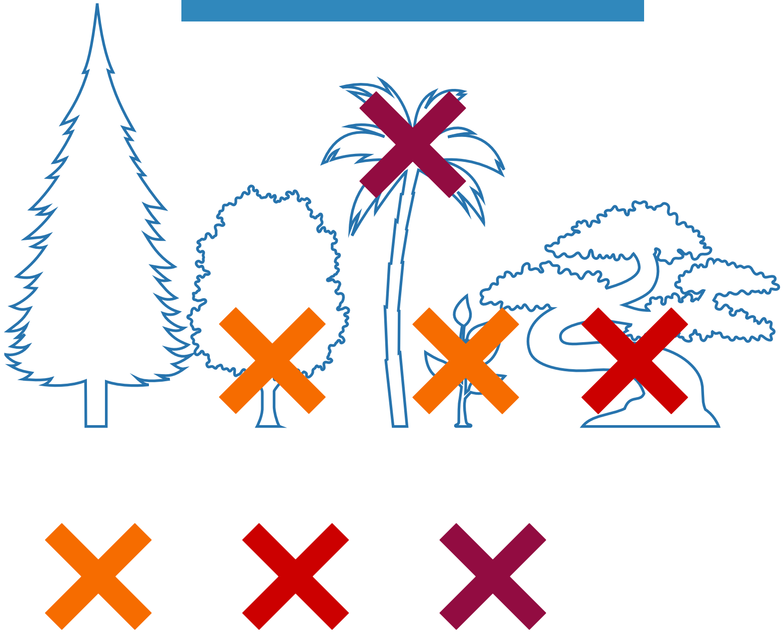
$$r(\mathbf{x}) = \underbrace{f(\mathbf{x})}_{\text{Model}} + \underbrace{u(\mathbf{x})}_{\text{Uncertainty}} + \underbrace{\zeta(\mathbf{x})}_{\text{Error}}$$

- Gain from reducing uncertainty: $\tilde{\mathcal{O}}(1/\sqrt{N})$
- Lost from error: $\tilde{\mathcal{O}}(1)$
  - N: number of data we **used**
- STOP before making mistakes
  - Skip the data $u(\mathbf{x}) \lessapprox \Delta$(desired precision)
  - Learn from the data $u(\mathbf{x}) \gtrapprox \Delta$



When desired precision Δ is not given to us…

Arm Elimination [CLRS' 11]



$\Delta = 4\text{ft}$ $\Delta = 2\text{ft}$ $\Delta = 1\text{ft}$

Theoretical results — Robust Data Selection for RL

Theorem [ZHFG23]. For any $0 < \delta < 1$, let the parameter be properly set, if the misspecification level is bounded by $\sqrt{d}\zeta \lessapprox \Delta$, then with probability at least $1 - \delta$, the cumulative regret is bounded by $\text{Regret}(K) \leq \tilde{\mathcal{O}}(d^2\Delta^{-1}\log(\delta^{-1}))$

Precision v.s. misspecification

Δ: difference between the 1^{st} and 2^{nd} action
ζ: model misspecification

$\text{Regret}(K) = \sum_{k=1}^{K} r_k^* - r(\mathbf{x}_k)$:
(total 'mistakes' for k rounds)

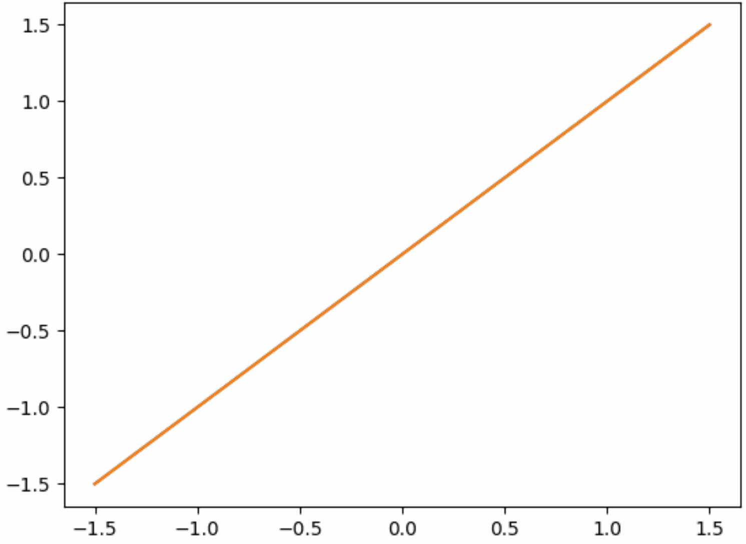
d: dimension of (linear) function approximation
δ: high-probability factor

Interplay between precision and model error

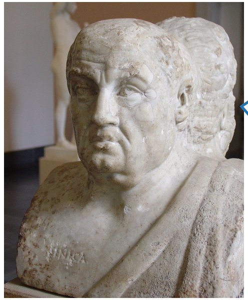
Theorem [ZHFG23]. For any $0 < \delta < 1$, let the parameter be properly set, if the misspecification level is bounded by $\sqrt{d}\zeta \lessapprox \Delta$, then with probability at least $1 - \delta$, the cumulative regret is bounded by $\text{Regret}(K) \leq \tilde{\mathcal{O}}(d^2\Delta^{-1}\log(\delta^{-1}))$

Theorem [ZHFG23]. When $\sqrt{d}\zeta \gtrapprox \Delta$, then there exists some hard case such that $\text{Regret}(K) \approx K\Delta$

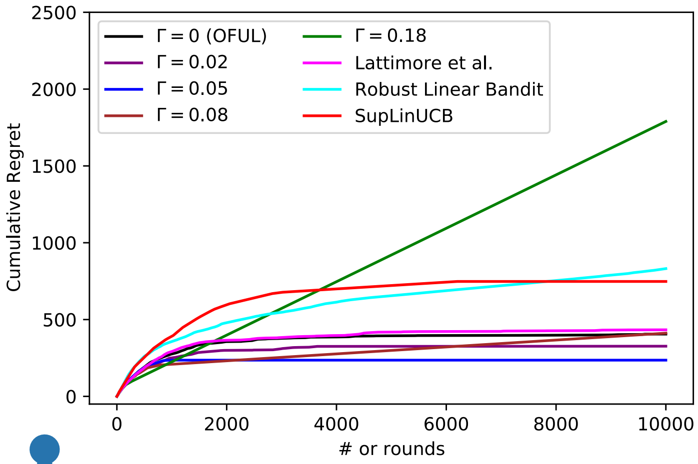
You can never learn a good estimator!



Byproduct: Constant Regret and Finite Mistakes

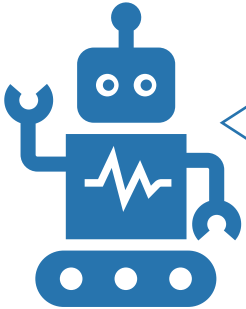


To err is human, but to persist is diabolical
— Seneca the Younger



$\text{Regret}(K) \leq \tilde{O}(d^2\Delta^{-1}\log(\delta^{-1}))$ not grow with K!

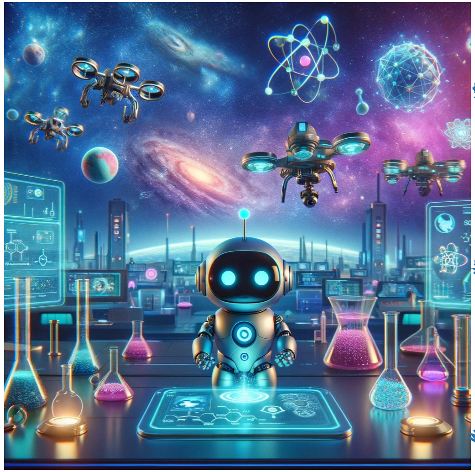
- First constant regret without prior assumption
  - Does not violate the $\Omega(\log K)$ bound
    - Easily match with $\delta = 1/K$
- Only finite budget (error) to learn the task!



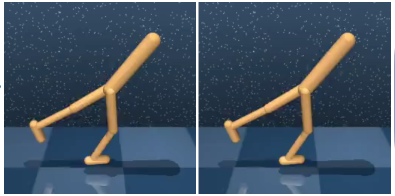
I make mistakes, but only finite mistakes
— RL with data selection

Uncertainty-aware data selection helps control model misspecification

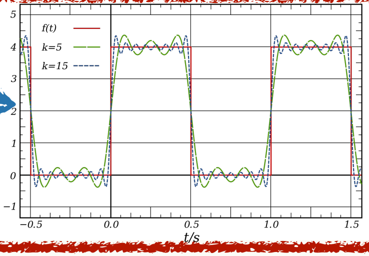
Next-step Decision Making for Science



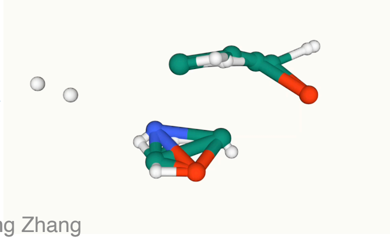
Endeavor: Decision making for scientific discoveries



Unsupervised data collection and exploration in reinforcement learning
[NeurIPS’21; ICML’23, ‘24]



Robust reinforcement learning under model error / misspecification
[ICML’23]



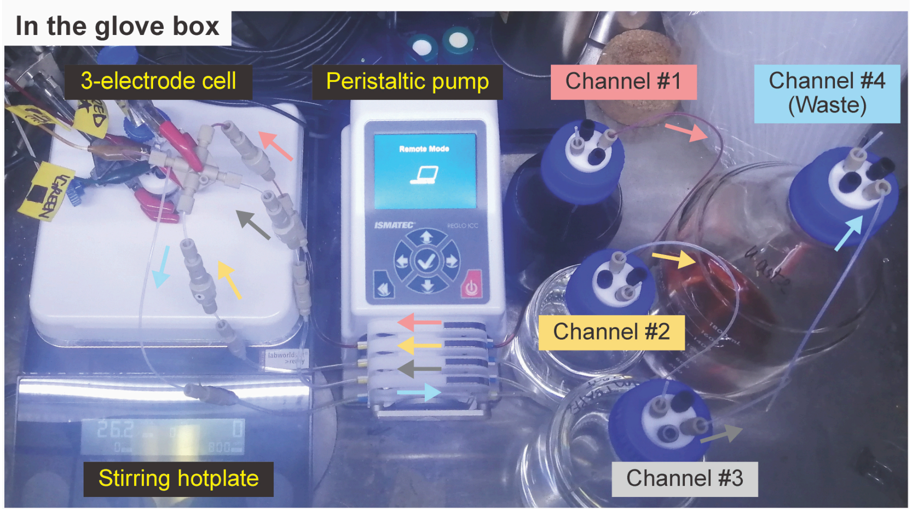
Frontier models / decision making for scientific tasks and drug design
[ACS Meas.Au’22, Nat. Comm.’24, etc.]

Next-step Decision Making for Science

OTHER WORKS AND FUTURE DIRECTION

Reinforcement learning for chemical analysis

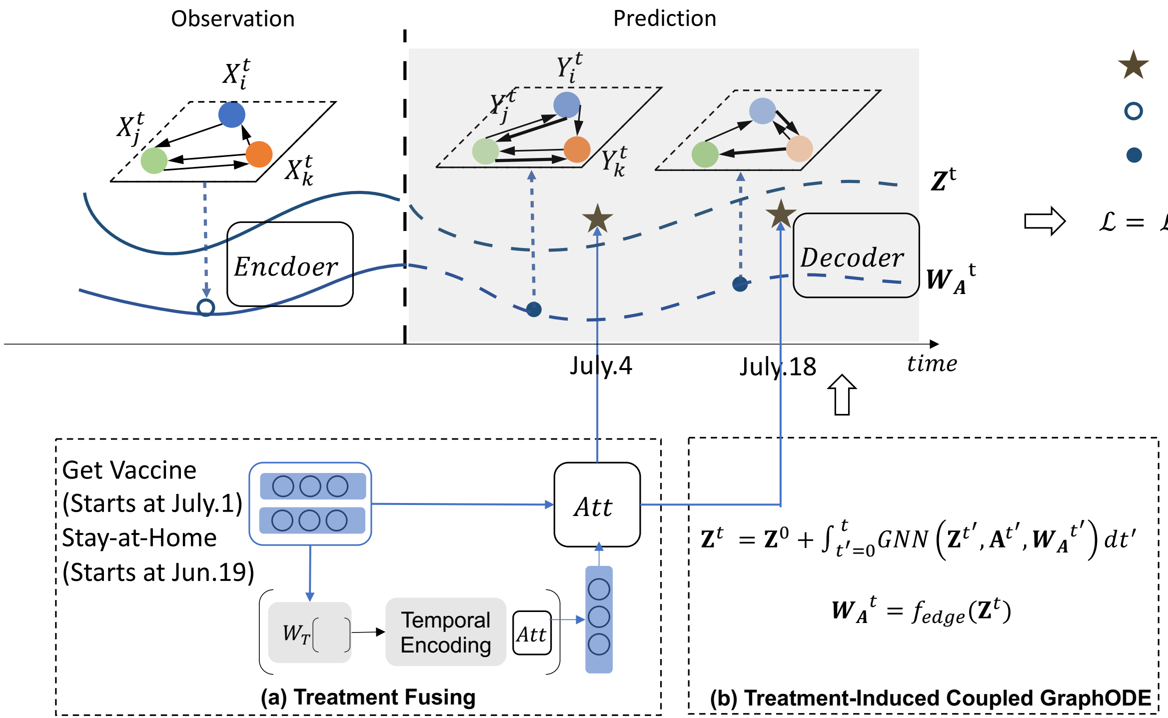
- Robotic systems:
  - 600 hrs wet lab -> 55 robot hrs
- Future directions:
  - Understanding the foundation of
    - Chemical reactions
    - Molecule science
  - Explainable RL for robust reaction analysis



UCLA College | Physical Sciences Chemistry & Biochemistry

[HZ+, ACS Meas. Au'22]
[SS+, Nat. Comm.'24]

Pandemic control using causal inference



★ Treatments

○ Observations

● Latent Representations

$$\Rightarrow \quad \mathcal{L} = \mathcal{L}^{<Y>} + \lambda\mathcal{L}^{<W>} + \alpha\mathcal{L}^{<A>} + \beta\mathcal{L}^{<G>} + \gamma\mathcal{L}_{KL}$$

How many people will get infected by COVID?

Will it be better if we all get vaccine?

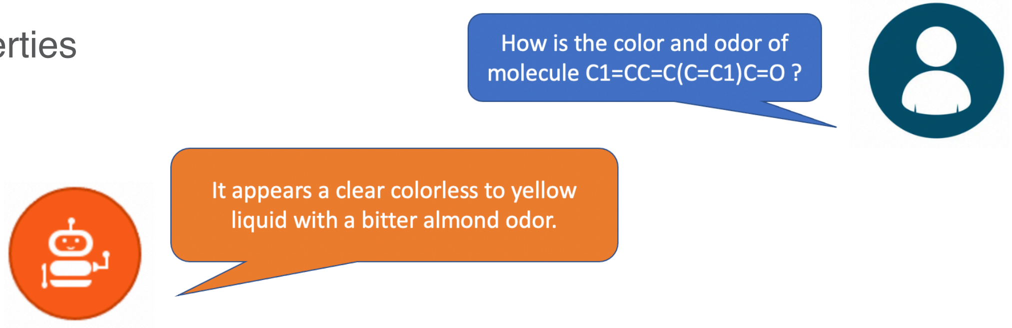
Future work: physical informed NNs for epidemic models (SEIR)

CNN → Seq2Seq → Atari RL
Diffusion → LLM → 🔥

🚀 Better decision making empowered by foundation models

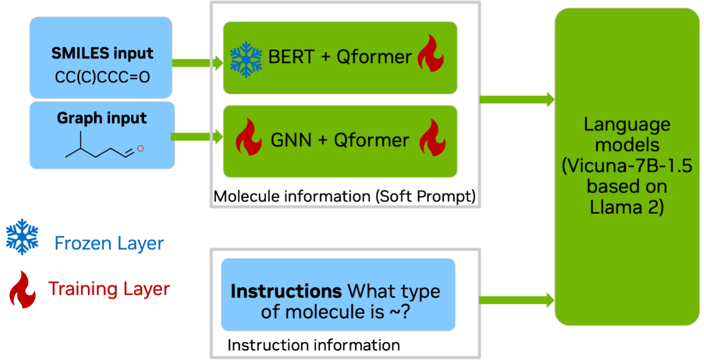
Multi-modality LLM for molecule prediction

- LLM to describe the molecule properties
- New dataset for evaluation



Future work

- Reasoning for scientific tasks
  - E.g. Why this molecule is toxic?
    - (Similar molecule is toxic?)
    - (Some structure is toxic?)
- Foundation model of RL + LLM
  - E.g. RLHF, self-supervised learning,
  - E.g. LLM agent with RL



Drug discovery using diffusion models

- Equivariant model (Rotation, translation)
  - Theoretical framework

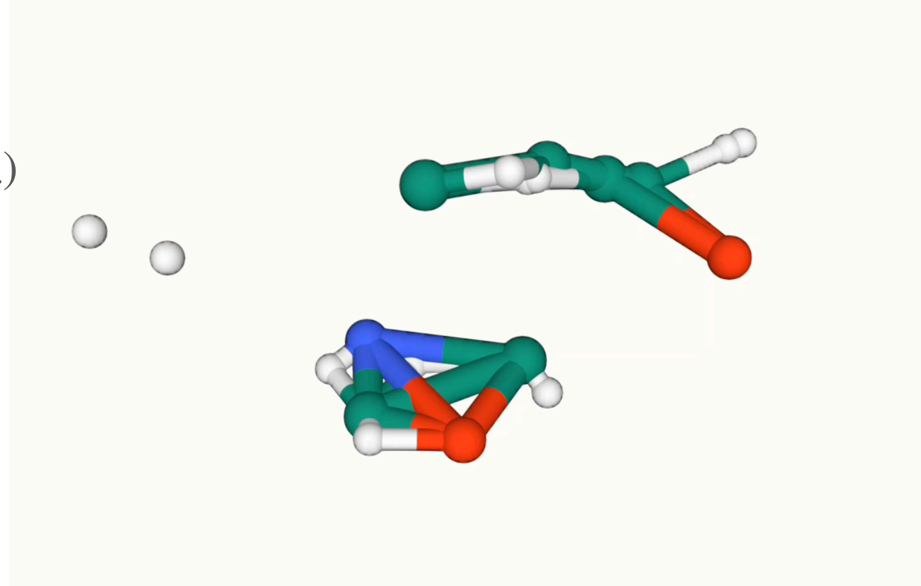
$$\Pr\left(\{\vec{x}_n\}\right) = \Pr\left(\{\vec{x}_n - \vec{x}_c\}\right)\Pr(\vec{x}_c)$$

- Discrete generative model for atom type

$\Rightarrow$ Stable, higher quality generation

Future work

- RL + diffusion model $\Rightarrow$ trial and error!
- Protein / Ligand generation



Last 300 step in reverse (denoising) process

Decision making for scientific discoveries and healthcare

- Exploration for scientific tasks
- Automated systems research
- Field research in public health

Interdisciplinary collaborations for decision making

Advanced decision making algorithms

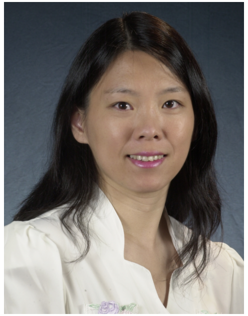
- Unsupervised RL / Exploration
- Robust RL / Adversarial RL
- Multi-agent RL

Decision making with foundation models

- LLM agent / RLHF
- Diffusion RL
- Self-supervised exploration

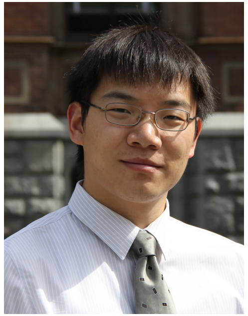
Acknowledgements


Advisor: Quanquan Gu

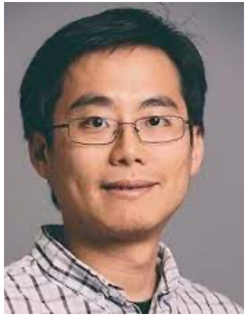

Prof. Wei Wang


Prof. Amy Zhang (UT Austin)


Prof. Yizhou Sun


Prof. Chong Liu (UCLA Chem)


Prof. Dominik Wodarz (UCSD, BioScience)


Dr. Lihong Li (Amazon)


Dr. Joe Eaton (Nvidia)

Dr. Bradley Rees (Nvidia)

SCIENCE HUB FOR HUMANITY AND ARTIFICIAL INTELLIGENCE
UCLA amazon

NVIDIA

Thank You

Image Credits I

- Decision making for scientific discoveries and healthcare: GPT4 (https://chat.openai.com/)
- CDC website: https://web.archive.org/web/20200618014344/https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html
- PCH table, Pearl's book: https://crl.causalai.net/crl-icml20.pdf
- Spherical harmonics: https://en.wikipedia.org/wiki/Spherical_harmonics
- Maze: https://www.mazegenerator.net/
- Unsupervised RL: https://bair.berkeley.edu/blog/2021/12/15/unsupervised-rl/
- Google search: https://www.google.com
- RL demonstration: https://commons.wikimedia.org/w/index.php?curid=57895741
- Go game: https://commons.wikimedia.org/w/index.php?curid=15223468
- Mouse-Maze solving: https://www.youtube.com/watch?v=ZMQbHMgK2rw

Image Credits II

- Issac Newton: https://en.wikipedia.org/wiki/Isaac_Newton
- Albert Einstein: https://en.wikipedia.org/wiki/Albert_Einstein
- IC: https://en.wikipedia.org/wiki/Integrated_circuit
- Tape ruler: https://en.m.wikipedia.org/wiki/File:Retractable_twenty_meter_tape_measure_2.jpg
- Vernier calipers: https://i.ebayimg.com/images/g/FD8AAOSwex1kR9WN/s-l1600.jpg
- Orange selection video: https://www.youtube.com/watch?v=2J_SxL7FvM0
- Seneca the Younger: https://en.wikipedia.org/wiki/Seneca_the_Younger
- Square wave: https://commons.wikimedia.org/wiki/File:Square_Wave_Fourier_Series.svg

References I

- [CR+, PNAS'22]: Cramer, Estee Y., et al. "Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States." *Proceedings of the National Academy of Sciences* 119.15 (2022): e2113561119.
- [SB+, PNAS'23]: Shea, Katriona, et al. "Multiple models for outbreak decision support in the face of uncertainty." *Proceedings of the National Academy of Sciences* 120.18 (2023): e2207537120.
- [HH+, WWW'24]: Huang, Zijie, et al. "Causal Graph ODE: Continuous Treatment Effect Modeling in Multi-agent Dynamical Systems." *The Symbiosis of Deep Learning and Differential Equations III*. 2023
- [HSW'21]: Huang, Zijie, Yizhou Sun, and Wei Wang. "Coupled graph ode for learning interacting system dynamics." *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021.
- [SI+'22]: Seedat, Nabeel, et al. "Continuous-time modeling of counterfactual outcomes using neural controlled differential equations." *arXiv preprint arXiv:2206.08311* (2022).
- [MD+'22]: Ma, Jing, et al. "Assessing the causal impact of COVID-19 related policies on outbreak dynamics: A case study in the US." *Proceedings of the ACM Web Conference 2022*. 2022.
- [WVL'22]: Weltz, Justin, Alex Volfovsky, and Eric B. Laber. "Reinforcement learning methods in public health." *Clinical therapeutics* 44.1 (2022): 139-154.

References II

- [ZZG, NeurIPS'21]: Zhang, Weitong, Dongruo Zhou, and Quanquan Gu. "Reward-free model-based reinforcement learning with linear function approximation." *Advances in Neural Information Processing Systems* 34 (2021): 1582-1593.
- [ZZG, ICML'23]: Zhang, Junkai, Weitong Zhang, and Quanquan Gu. "Optimal Horizon-Free Reward-Free Exploration for Linear Mixture MDPs." *International Conference on Machine Learning. PMLR. 2023.*
- [PA+17]: Pathak, Deepak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. "Curiosity-driven exploration by self-supervised prediction." *International conference on machine learning, PMLR, 2017,* pp. 2778-2787.
- [PG+19]: Pathak, Deepak, Dhiraj Gandhi, and Abhinav Gupta. "Self-supervised exploration via disagreement." *International conference on machine learning. PMLR, 2019,* pp. 5062-5071.
- [BE+18]: Burda, Yuri, Harrison Edwards, Amos Storkey, and Oleg Klimov. "Exploration by random network distillation." *International Conference on Learning Representations. 2018.*
- [SS+, Nat. Comm.'24 (in press)]: Sheng, Hongyuan, et al. "Autonomous closed-loop mechanistic investigation of molecular electrochemistry via automation." (2023).
- [HZ+, ACS Meas. Au'22]: Hoar, Benjamin B., et al. "Electrochemical mechanistic analysis from cyclic voltammograms based on deep learning." *ACS Measurement Science Au* 2.6 (2022): 595-604.