On the Science of Reinforcement Learning and Reinforcement Learning for Science

Weitong Zhang
weightzero@ucla.edu
Department of Computer Science
University of California, Los Angeles

Witnessing recent achievement in machine learning especially reinforcement learning (RL), my overarching research ambition revolves around crafting **reinforcement learning agents** that can push the boundaries of **scientific discovery**. Consider the paradigm of drug design: traditional methods typically necessitate manual drug design, synthesis, and testing. Modern machine learning has elevated this process by giving pre-laboratory drug properties prediction and generating potential drug candidates. With the capabilities of RL, agents can autonomously explore the space of unforeseen drug compounds and optimize the drug structures given empirical feedback, significantly enhancing the efficiency of the discovery process in the field of *chemistry*, *bioengineering* and *biology*.

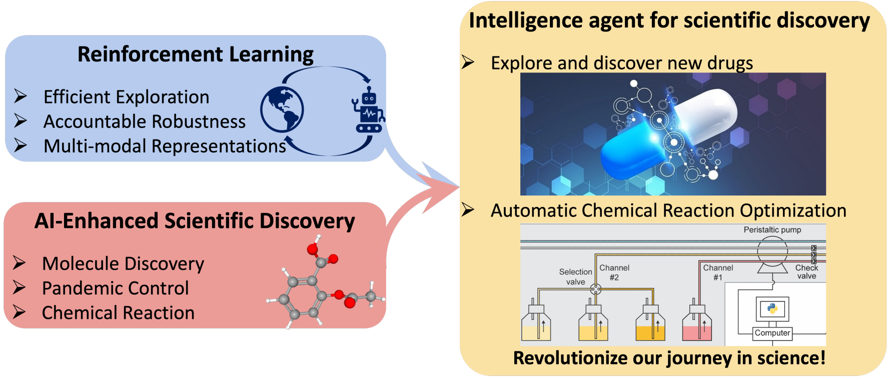

Reinforcement Learning
Efficient Exploration
Accountable Robustness
Multi-modal Representations
AI-Enhanced Scientific Discovery
Molecule Discovery
Pandemic Control
Chemical Reaction
Intelligence agent for scientific discovery
Explore and discover new drugs
Automatic Chemical Reaction Optimization
Peristaltic pump
Selection valve
Channel #2
Channel #1
Check valve
Computer
Revolutionize our journey in science!

Figure 1: Overview of my research interests and achievements

However, applying RL to cutting-edge scientific discovery tasks demands rigorous justifications. One of the cornerstones of scientific research is reliability. The RL algorithms should be *robust* against the perturbation or adversarial attacks. It is also important to justify that the agent's behavior is not just a one-off occurrence but can be replicated across various environments and conditions. In addition, scientific endeavors often entail *multifaceted representations*, spanning molecular structures, protein sequences, and molecular dynamics. Thus, RL models must adeptly leverage information from these varied sources. Moreover, in scientific contexts, the exploration is very expensive in terms of time and resources. Therefore, we need to design RL algorithms that can *explore* environments efficiently, ensuring that valuable resources are not wasted and that results are achieved in a timely manner.

Therefore, my research contributions divide into two primary thrusts: The first is building the **Foundation of Reinforcement Learning**, providing justification and insights for the aforementioned questions. The second is fostering pioneering applications of **AI-Enpowered Scientific Discovery**. Moving forward, I seek to integrate these two thrusts, elevating RL agents to the forefront of scientific discovery. An overview of my research objectives and achievement is presented in Figure 1.

1 Research Achievements

1.1 Foundations of Reinforcement Learning

Reinforcement learning (RL) has demonstrated profound capabilities in navigating complex environments. In this research thrust, I focus on the foundational aspects of reinforcement learning. The driving force behind this focus is to ensure that RL can be confidently applied to rigorous scientific tasks.

1.1.1 Robustness in Reinforcement Learning

Many real-world RL tasks exhibit vast state and action spaces, making it impractical to represent every potential state-action pair explicitly. For example, a Go-game contains nearly 10^{170} different board settings. Therefore, function approximation is usually applied to address this challenge to make RL more scalable and applicable to a broader range of problems. However, there is a significant concern regarding **robustness** in function approximation for reinforcement learning. For example, in scientific domains like robotic lab experiments, environmental factors such as temperature can introduce perturbation given the theoretical models. Given these complexities, it is crucial to address **how misspecification can affect RL algorithms** and, more importantly, **how to develop RL algorithms resilient to these misspecification**.

To tackle these challenges, my work [16] focuses on contextual bandits, a type of one-step RL task where agents primarily learn the reward function. My findings in this domain are twofold. On the one hand, I showed that beyond a certain level of misspecification, the bandit task becomes difficult to learn. On the other hand, I introduced an algorithm that can actively select data during its learning process to ensure robust estimation. I proved that the proposed algorithm can effectively learn the bandit model when the misspecification level is relatively small. This research provides a clear understanding on **the impact of misspecification**.

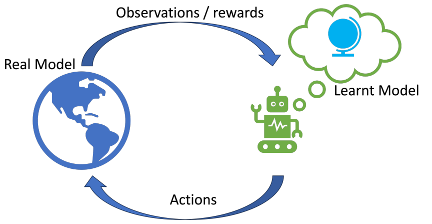


Figure 2: Real-world disturbances often lead to discrepancies between the physical dynamics of the environment and their theoretical formulations.

Expanding on these insights, I delved deeper into more complex RL problems in a followup work. In RL, besides rewards, perturbation can happen on the environmental transition as well. For instance, as shown in Figure 2, real-world disturbances often lead to discrepancies between the physical dynamics of the environment and their theoretical formulations. In response to this, I introduced innovative techniques, notably the "certified estimator" for providing a robust estimation. Harnessing these innovations, I proposed an algorithm that delivers reliable estimations, even in the face of approximation errors. Our results shed light on the maximum misspecification that can be tolerated while still achieving efficient RL.

1.1.2 Representations in Reinforcement Learning

In this research direction, I delve into the fusion of representations from multiple modalities within reinforcement learning. In many real-world scenarios, there are multiple legitimate representations for a given state-action pair. Consider the molecule discovery tasks, the same molecule can be represented by its SMILES string, 2D graph and 3D structured conformers. Each of them has better performance on different tasks like *molecule screening*, *synthesis analysis*, and *reactivity analysis*. This leads to the pressing question: **How can we adeptly fuse various representations within sequential decision systems?**

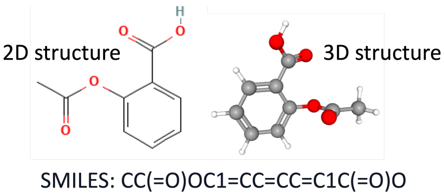


Figure 3: Different molecule modalities and representations (SMILES, 2D graph, 3D graph) can be used separately in drug design tasks.

To answer this question, my work detailed in [17] introduces a method for representation selection in both online and offline RL that integrates representations from diverse modalities. To exemplify this point, consider the aforementioned drug discovery example, the algorithm will intuitively opt for the SMILES representation during *similarity screening* and the 2D graph in the *synthesis analysis* phase. Theoretically, I showed that our proposed algorithm, across both online and offline contexts, achieves lower regret and smaller sub-optimality gap compared with baseline algorithms. Empirical experiments further demonstrate the superiority of our algorithm on synthetic datasets, affirming the practical implications of our theoretical insights.

1.1.3 Unsupervised Reinforcement Learning

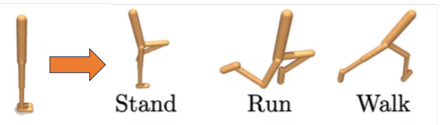


Figure 4: Initially (left), the robot is trained without a defined goal. Subsequently (right), it undergoes fine-tuning for specific tasks such as standing, running or walking.

In this research avenue, I explore the area of *unsupervised reinforcement learning*, often termed *reward-free RL*. Traditional *supervised RL* approaches frequently face data inefficiency challenges. To illustrate, as presented in Figure 4, when integrating RL into robotics tasks, supervised RL typically requires a predetermined goal (e.g., jumping, walking), limiting the exploratory range within the related motions. This constraint becomes evident when, upon altering the goal, the RL agent must re-explore the environment to find the related actions again, which is costly and time-inefficient. In contrast, by *unsupervised reinforcement learning* or *reward-free exploration*, we hope the agent can explore the environment by itself without any predefined goal and then refine its actions by either planning its behavior offline or finetuning it with minimal resources, once given a desired goal. Clearly, this paradigm offers significant resource savings, particularly when the goals are subject to frequent changes, as seen in the trial-and-error processes of rapid iterative design.

My research in this area aims to craft efficient *reward-free exploration* algorithms with theoretical performance guarantees. As a starting point, my work [12] introduces a model-based algorithm for *reward-free exploration* coupled with function approximation. I framed the reward-free exploration concept within the PAC framework, leading to the design of a tailored exploration algorithm. Central to this algorithm is the introduction of the **pseudo reward**—an intrinsic metric guiding the agent's exploration. This *pseudo reward* is designed to prompt the agent to explore the environment. After the exploration of the environment without reward signals, the agent can then commit a near-optimal policy once given the reward function. In a followup work [11], I extended this result to a *horizon-free* context, which further validates the optimality of the proposed algorithm.

1.2 AI-Empowred Scientific Discovery

Alongside foundational reinforcement learning, I have also applied modern machine learning to scientific discovery. While theoretical research forms the backbone of AI, advancing the frontier of machine learning-guided scientific discovery is also a fast growing area that promises to foster cutting-edge applications.

1.2.1 Molecule Design via Modern Machine Learning

My work in this area aims to push the boundaries of molecular discoveries, particularly in the realms of **molecule generation** and **molecule property prediction**, utilizing state-of-the-art machine learning techniques.

To begin, in [13], I applied diffusion models to generate 3D molecular conformations. A significant challenge in using diffusion models for 3D molecules is maintaining the **equivariance** of molecular structures. As illustrated in Figure 5, the generation process should remain unaffected by molecular translations and rotations. Theoretically, I provided a decomposition between the distribution of the center of the molecule and the distribution of the relative position of each atom in the molecule. Through this, I provided theoretical foundation for the existing algorithm [5]. Besides the theoretical contribution, I improved the existing method using discrete diffusion processes for generating atom types in the molecule, leading to improved empirical results in molecule generation.

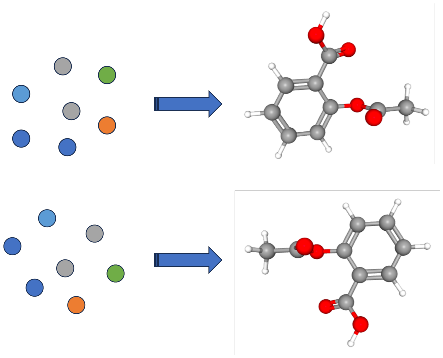
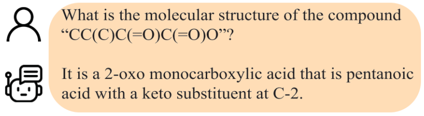


Figure 5: The equivariant diffusion process, which is not affected by the rotation and translation of the molecule.



What is the molecular structure of the compound "CC(C)C(=O)C(=O)O"?

It is a 2-oxo monocarboxylic acid that is pentanoic acid with a keto substituent at C-2.

Figure 6: One sample of the instruction and response of the LLM. The molecule information is extracted by pretrained GNNs or BERTs.

In another endeavor [14], I tackled molecule property prediction. Most existing models primarily focus on closed-set predictions on manually crafted property attributes, like heat capacity, polarizability, or binding affinity, which can hardly be adapted to new molecular properties. Given the rich history of advancements in chemistry and biology, utilizing textual data—such as annotations from past research articles—can enhance a model's ability to make open-set predictions. This approach allows for easy generalization to new tasks, drawing upon the vast body of knowledge encapsulated in previous literature. Building on that insight, I fuse the molecular structure information into large language models (LLMs) that generate textual descriptions for molecule structures based on provided instructions. As displayed in Figure 6, the model produces results reminiscent of human-generated content and demonstrates enhanced performance across several downstream tasks.

1.2.2 Categorizing Chemistry Reactions with Machine Learning

In collaboration with UCLA's Department of Chemistry, I have also been employing machine learning to analyze electrochemical mechanisms. Traditionally, these mechanisms required extensive manual investigation. In a pioneering effort [4], I introduced deep learning methods to *categorize* chemical reactions. Following the initial reaction predictions, we employed online optimization techniques like Bayesian optimization to autonomously refine reaction formulas [9]. My subsequent work has extended this approach, which simultaneously *detects and categorizes* mechanisms from a broad spectrum of cyclic voltammogram data in chemical reactions.

1.2.3 Leveraging AI in Pandemic Control

I have been passionate about harnessing machine learning for societal benefit. Beginning in March 2020, our team launched a project[1] [18] to model and forecast the COVID-19 spread using AI-driven epidemic models. These predictions found application in federal and local government agencies, including the US CDC. However, mere predictions were insufficient for designing public health policies. In subsequent research [6], I integrated *causal inferences* to estimate the effect of interventions like "mask mandates" and "social distancing" policies, which provide interpretable guidance for policy makers to combat the pandemics.

[1] Website: `covid19.uclaml.org` – note: updates ceased in 2022

2 Future Research Plan

2.1 Reinforcement Learning Agents with Large Language Models

Large Language Models (LLMs) have demonstrated remarkable reasoning capabilities, showing the possibility to deployment in real-world tasks such as scientific discoveries. Yet, the transition from reasoning to actions in real environments presents a huge challenge. More specifically, there is a clear challenge in directing the LLM to effectively understand and act in real-world situations. I aim to address these issues from two aspects:

First, my focus will be on developing unsupervised RL methodologies that enable LLMs to effectively explore complex environments. Take the search engine empowered by LLMs for example, can we guide the LLM to autonomously traverse online resources and subsequently, when presented with a search directive, formulate a search strategy? I aim to apply theoretical insights from my prior works [11, 12] to craft this unsupervised paradigm and provide both practical algorithms and theoretical understandings with *In Context Learning* or *Chain of Thoughts*. My prior engagements with LLM-centric topics [14] will undoubtedly be instrumental in this endeavor.

Second, my efforts will pivot towards engineering robust decision-making algorithms for LLM-based RL agents. For example, *hallucination* issues in LLMs [8] can be perceived as a misspecification between the real-world environment and the environment construed by the LLM's underlying transformer structure. Leveraging theoretical insights from my past research [16], I am optimistic about rectifying such hallucinations during the decision-making phase. Furthermore, this approach might also pave the way for enhancing the robustness of LLM responses, particularly against adversarial attacks.

2.2 Reinforcement Learning for Trustworthy Scientific Discovery

In this direction, I plan to leverage Reinforcement Learning (RL) to drive pivotal scientific advancements, ensuring the delivery of trustworthy and accountable discoveries. My primary focus is on two scientific domains: (1) RL for chemical reactions and (2)RL for drug discovery. In pursuing this goal, I will build interdisciplinary collaborations with researchers from academia and industry to push the frontier of this area.

Starting with RL for chemistry, my earlier research [7, 15] has developed efficient exploration algorithms using neural networks. Building on my experience with deep neural networks in analyzing electrochemical reactions, as shown in Figure 7, the optimization can be formulated as finding the best combination of substance 'A', 'B' and 'C', where the performance of the reaction is predicted by neural networks [4]. I anticipate that exploration methods in [7, 15] will enhance the process of optimizing these reactions with accountable performance guarantee. For this interdisciplinary research goal, I plan to partner with experts in electrochemistry and computational chemistry to craft more tailored exploration algorithms.

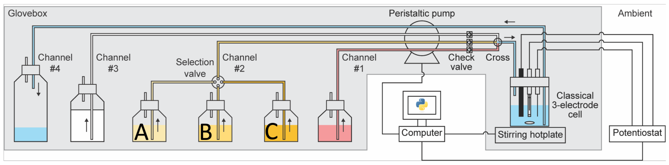


Figure 7: An illustration of applying RL to optimizing chemical reactions, the RL agent explore the ratio of substance 'A', 'B' and 'C' to maximize the reaction performance. The performance is estimated by some neural networks ('computer') in the figure.

Turning to the RL application in protein ligand generation, I plan to expand on my prior work [13], which employs diffusion techniques for generating protein ligand structures. For this venture, I intend to closely collaborate with researchers in NVIDIA and leading pharmaceutical and biotech companies to pioneer advancements in AI-powered drug design. Notably, our initial findings in computational chemistry indicate that ligands produced by current state-of-the-art models [1, 3] are not always viable for drug use: these generated ligands might require 10 to 100 times more folding energy to achieve the desired conformation necessary for effectively binding to target proteins, compared to ligands already used in therapies. This limitation reduces the reactivity of the generated ligands, often down to less than 1% of active drugs. Given the comprehensive research on predicting the energy of proteins/molecules [2, 10], my proposal involves using RL to refine the generation phase of the diffusion model. With this strategy, I aim to lower the free energy of the produced molecules, making the resulting ligands more suitable for real-world pharmaceutical applications.

2.3 Explainable AI for Scientific Applications

Beyond establishing foundations for reinforcement learning, I am keen on crafting explainable techniques for scientific pursuits using *physics-informed neural networks*. As an illustration, my prior research [13] devised a neural network that captures the equivariance inherent in the generation process of diffusion models. Recent progress in the field includes the use of spherical harmonics as a basis function to approximate density functional theories in quantum chemistry. With this research direction, I aim to delve deeper into discovering more physics-informed features, such as diverse basis functions, or incorporating physics-informed layers tailored for distinct scientific domains. In the meantime, I am dedicated to elucidating the theoretical advantages in these specialized designs.

Future Collaboration and Funding Resources

I intend to expand my existing collaboration with the Department of Chemistry at UCLA and establish interdisciplinary partnerships with academic researchers in chemistry, biology, and bio-engineering, tailored to the specific requirements of my research agenda. Building on my existing collaborations with NVIDIA and Amazon, I aim to forge robust industrial connections. A long-term collaboration with NVIDIA Research is on the horizon. To translate my research into real-world applications, I will capitalize on my existing partnerships with the NVIDIA cuGraph and BioNeMo engineering teams, as well as the AstraZeneca research team, Amazon DGL Life Science team. Furthermore, I am keen on extending my collaborations with entities like IBM Research AI, Google Research, Facebook AI Research (FAIR), Microsoft Research, OpenAI, and DeepMind.

I anticipate potential support from federal funding agencies such as the National Science Foundation (NSF), National Institutes of Health (NIH), Defense Advanced Research Projects Agency (DARPA), Department of Engineering (DoE), Alfred P. Sloan Foundation, Robert Wood Johnson Foundation, Simons Foundation, Wellcome Trust, Bill & Melinda Gates Foundation, and the Allen Institute for Artificial Intelligence (AI2).

References

[1] Gabriele Corso et al. "Diffdock: Diffusion steps, twists, and turns for molecular docking". In: *arXiv preprint arXiv:2210.01776* (2022).

[2] Johannes Gasteiger, Janek Groß, and Stephan Günnemann. "Directional message passing for molecular graphs". In: *arXiv preprint arXiv:2003.03123* (2020).

[3] Jiaqi Guan et al. "DecompDiff: Diffusion Models with Decomposed Priors for Structure-Based Drug Design". In: (2023).

[4] Benjamin B Hoar et al. "Electrochemical mechanistic analysis from cyclic voltammograms based on deep learning". In: *ACS Measurement Science Au* (2022).

[5] Emiel Hoogeboom et al. "Equivariant diffusion for molecule generation in 3d". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 8867–8887.

[6] Zijie Huang et al. "Causal Graph ODE: Continuous Treatment Effect Modeling in Multi-agent Dynamical Systems". In: *NeurIPS 2023 Workshop on The Symbiosis of Deep Learning and Differential Equations III*. 2023.

[7] Yiling Jia et al. "Learning neural contextual bandits through perturbed rewards". In: *International Conference on Learning Representations*. 2021.

[8] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: `2303.08774 [cs.CL]`.

[9] Hongyuan Sheng et al. "Autonomous closed-loop mechanistic investigation of molecular electrochemistry via automation". In: *ChemRxiv preprint* (2023).

[10] Philipp Thölke and Gianni De Fabritiis. "Torchmd-net: equivariant transformers for neural network based molecular potentials". In: *arXiv preprint arXiv:2202.02541* (2022).

[11] Junkai Zhang, **Weitong Zhang**, and Quanquan Gu. "Optimal Horizon-Free Reward-Free Exploration for Linear Mixture MDPs". In: *International Conference on Machine Learning*. PMLR. 2023.

[12] **Weitong Zhang**, Dongruo Zhou, and Quanquan Gu. "Reward-Free Model-Based Reinforcement Learning with Linear Function Approximation". In: *Advances in neural information processing systems* (2021).

[13] **Weitong Zhang** et al. "DiffMol: 3D Structured Molecule Generation with Discrete Denoising Diffusion Probabilistic Models". In: *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*. 2023.

[14] **Weitong Zhang** et al. "MoleculeGPT: Instruction Following Large Language Models for Molecular Property Prediction". In: *NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development*. 2023.

[15] **Weitong Zhang** et al. "Neural Thompson Sampling". In: *International Conference on Learning Representations*. 2020.

[16] **Weitong Zhang** et al. "On the Interplay Between Misspecification and Sub-optimality Gap in Linear Contextual Bandits". In: *International Conference on Machine Learning*. PMLR. 2023.

[17] **Weitong Zhang*** et al. "Provably Efficient Representation Learning in Low-rank Markov Decision Processes". In: *UAI*. 2023.

[18] Difan Zou et al. "Epidemic model guided machine learning for COVID-19 forecasts in the United States". In: *medRxiv* (2020).