

FinRL-Podracr: High Performance and Scalable Deep Reinforcement Learning for Quantitative Finance

Zechu Li
zl2993@columbia.edu
Columbia University

Xiao-Yang Liu*
xl2427@columbia.edu
Columbia University

Jiahao Zheng
jh.zheng@siat.ac.cn
Shenzhen Inst. of Advanced Tech.

Zhaoran Wang
zhaoranwang@gmail.com
Northwestern University

Anwar Walid†
anwar.i.walid@gmail.com
Amazon & Columbia University

Jian Guo‡
guojian@idea.edu.cn
IDEA Research

ABSTRACT

Machine learning techniques are playing more and more important roles in finance market investment. However, finance quantitative modeling with conventional supervised learning approaches has a number of limitations, including the difficulty in defining appropriate labels, lack of consistency in modeling and trading execution, and lack of modeling the dynamic nature of the finance market. The development of deep reinforcement learning techniques is partially addressing these issues. Unfortunately, the steep learning curve and the difficulty in quick modeling and agile development are impeding finance researchers from using deep reinforcement learning in quantitative trading. In this paper, we propose an *RLOps in finance* paradigm and present a *FinRL-Podracr* framework to accelerate the development pipeline of deep reinforcement learning (DRL)-driven trading strategy and to improve both trading performance and training efficiency. FinRL-Podracr is a cloud solution that features high performance and high scalability and promises *continuous training*, *continuous integration*, and *continuous delivery* of DRL-driven trading strategies, facilitating a rapid transformation from algorithmic innovations into a profitable trading strategy. First, we propose a generational evolution mechanism with an ensemble strategy to improve the trading performance of a DRL agent, and schedule the training of a DRL algorithm onto a GPU cloud via multi-level mapping. Then, we carry out the training of DRL components with high-performance optimizations on GPUs. Finally, we evaluate the FinRL-Podracr framework for a stock trend prediction task on an NVIDIA DGX SuperPOD cloud. FinRL-Podracr outperforms three popular DRL libraries *Ray RLlib*, *Stable Baseline 3* and *FinRL*, i.e., 12% ~ 35% improvements in annual return, 0.1 ~ 0.6 improvements in Sharpe ratio and $3\times \sim 7\times$ speed-up in training time. We show the high scalability by training a trading agent in 10 minutes with 80 A100 GPUs, on NASDAQ-100 constituent stocks with minute-level data over 10 years.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning; Neural networks; Markov decision processes; Reinforcement learning.**

*Equal contribution.

†A. Walid finished this project at Bell labs, before joining Amazon.

‡Corresponding author.

KEYWORDS

RLOps in finance, deep reinforcement learning, stock trend prediction, scalability, GPU cloud

1 INTRODUCTION

Algorithmic trading is increasingly deployed in the financial investment process. A conventional supervised learning pipeline consists of five stages [31, 40], as shown in Fig. 1 (left), namely data pre-process, modeling and trading signal generation, portfolio optimization, trade execution, and post-trade analysis. Recently, deep reinforcement learning (DRL) [34, 36, 37] has been recognized as a promising alternative for quantitative finance [2, 7, 16, 17], since it has the potential to overcome some important limitations of supervised learning, such as the difficulty in label specification and the gap between modeling, positioning and order execution. We advocate extending the principle of *MLOps* [1] ¹ to the *RLOps in finance* paradigm that implements and automates the continuous training (CT), continuous integration (CI), and continuous delivery (CD) for trading strategies. We argue that such a paradigm has vast profits potential from a broadened horizon and fast speed, which is critical for wider DRL adoption in real-world financial tasks.

The *RLOps in finance* paradigm, as shown in Fig. 1 (right), integrates middle stages (i.e., modeling and trading signal generation, portfolio optimization, and trade execution) into a DRL agent. Such a paradigm aims to help quantitative traders develop an end-to-end trading strategy with a high degree of automation, which removes the latency between stages and results in a compact software stack. The major benefit is that it can explore the vast potential profits behind the large-scale financial data, exceeding the capacity of human traders; thus, the trading horizon is lifted into a potentially new dimension. Also, it allows traders to continuously update trading strategies, which equips traders with an edge in a highly volatile market. However, the large-scale financial data and fast iteration of trading strategies bring imperative challenges in terms of computing power.

Existing works are not satisfactory with respect to the usage of large-scale financial data and the efficiency of agent training. For DRL strategy design, existing works studied algorithmic trading on a daily time-frame [2, 25, 26, 42–45] or hourly time-frame [14], which is hard to fully capture the dynamics of a highly volatile market. For DRL library/package development, existing works may not

¹MLOps is an ML engineering culture and practice that aims at unifying ML system development (Dev) and ML system operation (Ops).

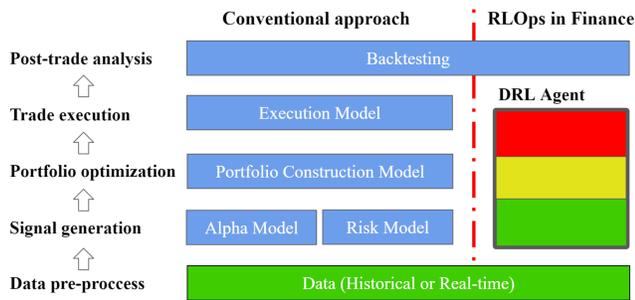


Figure 1: Software stack for an algorithmic trading process: conventional approach vs. RLOps in finance.

be able to meet the intensive computing requirement of relatively high frequency trading tasks, large-scale financial data processing and tick-level trade execution. We evaluate the training time of three popular DRL libraries, FinRL [25, 26], RLlib [19] and Stable Baseline3 [9], on NASDAQ-100 constituent stocks with minute-level data (described in Section 5.2) in Table 1, which shows that it is difficult for them to effectively train a profitable trading agent in a short cycle time.

In recent years, distributed DRL frameworks and massively parallel simulations have been recognized as the critical software development for the RLOps paradigm [12, 19, 20, 27]. It is promising to utilize extensive computing resources, e.g., a GPU cloud, to accelerate the development of trading strategies, including both financial simulation and model training. Therefore, we investigate a candidate solution on a GPU cloud, an NVIDIA DGX SuperPOD cloud [38] that is the most powerful AI infrastructure for enterprise deployments.

In this paper, we propose a *FinRL-Podcracer* framework as a high-performance and scalable solution for *RLOps in finance*. At a high level, FinRL-Podcracer schedules the training process through a multi-level mapping and employs a generational evolution mechanism with an ensemble strategy. Such a design guarantees scalability on a cloud platform. At a low level, FinRL-Podcracer realizes hardware-oriented optimizations, including parallelism encapsulation, GPU acceleration, and storage optimization, thus achieving high performance. As a result, FinRL-Podcracer can effectively exploit the supercomputing resources of a GPU cloud, which provides an opportunity to automatically design DRL trading strategies with fast and flexible development, deployment and production.

Our contributions can be summarized as follows

- We propose a *FinRL-Podcracer* framework built on two previous projects, FinRL [25, 26] and ElegantRL [23]², to initiate a paradigm shift from conventional supervised learning approaches to *RLOps in finance*.
- FinRL-Podcracer employs a generational evolution mechanism during the training of DRL agents and provides high-performance optimizations for financial tasks.
- We show the training efficiency by obtaining a trading agent in 10 minutes on an NVIDIA DGX SuperPOD cloud [38] with 80 A100 GPUs, for a stock trend prediction task on NASDAQ-100 constituent stocks with minute-level data over 10 years.

²ElegantRL is a scalable and elastic deep reinforcement learning library. It supports general robotic and game playing applications.

	Sharpe ratio	Max dropdown	Training time
RLlib [19]	1.67	-23.248%	110 min
SB3 [13]	1.82	-23.750%	150 min
FinRL [44]	1.35	-27.267%	345 min
QQQ	1.25	-28.559%	-

Table 1: Evaluations of existing DRL libraries on an NVIDIA DGX A100 server [5]. We evaluate on NASDAQ-100 constituent stocks with minute-level data by training from 01/01/2009 to 05/12/2019 and backtesting from 05/13/2019 to 05/26/2021. Invesco QQQ ETF is a market benchmark.

We evaluate the trained agent for one year and show its trading performance outperforms three DRL libraries, *FinRL* [25, 26], *Ray RLlib* [19] and *Stable Baseline3* [13], i.e., 12% ~ 35% improvements in annual return, 0.1 ~ 0.6 improvements in Sharpe ratio and 3× ~ 7× speed-up in training time.

The remainder of this paper is organized as follows. Section 2 describes related works. Section 3 models a typical stock trend prediction task as a Markov Decision Process. In Section 4, we present the FinRL-Podcracer framework and describe its evolution and training layers, respectively. In Section 5, we describe the experimental setup for the trading task and present experimental results. We conclude this paper and discuss future directions in Section 6.

2 RELATED WORKS

This section summarizes related works from two aspects: DRL applications in quantitative finance and the MLOps development.

2.1 DRL in Finance

With the successes of DRL in game playing, e.g., Atari games [30] and GO games [35], more and more finance researchers show their interests in this area, and they have done some early attempts to applying DRL in quantitative finance investment. In this paper, we take the stock trend prediction (STP) task as an example to introduce existing works and show great potentials of DRL in finance area.

Stock trend prediction task is often considered a challenging application of machine learning in finance due to its noisy and volatile nature. Traditionally, the STP task is formulated as a supervised learning problem, where features of stocks are extracted from a past time window of technical indices, fundamental data and alternative data [4], and labels are usually extracted from a future time window of concerned criteria such as rise/fall, returns, excess returns or Sharpe ratios. Recently, deep reinforcement learning has been applied to solving STP tasks. Zhang *et al.* [45] constructed a trading agent using three DRL algorithms, DQN, PG, and A2C, for both discrete and continuous action spaces. Yang *et al.* [44] used an ensemble strategy to integrate different DRL algorithms, A2C, DDPG, and PPO based on the Sharpe ratio. They applied the idea of a rolling window, where the best algorithm is picked to trade in the following period. Recently, many researchers provide more DRL solutions for STP tasks [3, 8, 18]. However, most existing works are based on several assumptions, which limits the practicality. For example, the backtesting has no impacts on the market; there are almost no other complex actions besides buying, holding, and selling; only one stock type is supported for each agent.

2.2 Principle of MLOps

Recently, Google trends put Machine Learning Operations (MLOps) as one of the most promisingly increasing trends [41]. MLOps is a practice in developing and operating large-scale machine learning systems, which facilitates the transformation of machine learning models from development to production [6, 28]. In essence, MLOps entails cloud computing power to integrate and automate a standard machine learning pipeline: 1) data pre-processing; 2) feature engineering; 3) continuous ML model training; 4) continuous ML model deployment; 5) output production, thus building applications that enable developers with limited machine-learning expertise to train high-quality models specific to their domain or data [22, 39].

However, the DRL is quite different from conventional machine learning approaches. For example, training data of DRL is not prepared in advance compared with conventional supervised learning but collected through an agent-environment interaction inside the training process. Such a significant difference requires a re-integration of the automated pipeline and a re-scheduling of the cloud computing resources with respect to the conventional MLOps principle. Therefore, we advocate extending the principle of MLOps to the *RLOps in finance* paradigm to seek an opportunity for the wider DRL adoption in production-level financial services.

3 STOCK TREND PREDICTION TASK

We describe the problem formulation of a typical financial task, stock trend prediction, which locates at the task layer of Fig. 2. Our setup follows a similar setting in [26, 44].

A stock trend prediction task is modeled as a Markov Decision Process (MDP): given state $s_t \in \mathcal{S}$ at time t , an agent takes an action $a_t \in \mathcal{A}$ according to policy $\pi_\theta(s_t)$, transitions to the next state s_{t+1} and receives an immediate reward $r(s_t, a_t, s_{t+1}) \in \mathbb{R}$. The policy $\pi_\theta(s)$ with parameter θ is a function that maps a state to an action vector over n stocks. The objective is to find an optimal policy π_θ^* (a policy network parameterized by θ) that maximizes the expected return (the fitness score used in Fig. 2) over T times slots

$$\pi_\theta^* = \operatorname{argmax}_\theta J(\pi_\theta), \text{ where } J(\pi_\theta) = \mathbb{E} \left[\sum_{t=0}^T \gamma^t r(s_t, a_t, s_{t+1}) \right], \quad (1)$$

where $\gamma \in (0, 1]$ is a discount factor.

Then, for the stock trend prediction task with n stocks, we specify the state space \mathcal{S} , action space \mathcal{A} , reward function $r(s_t, a_t, s_{t+1})$, and the state transition, as in [26][44].

State space \mathcal{S} describes an agent's perception of a market environment. We summarize various features that are used by human trader and use them to construct the state space:

- Balance $b_t \in \mathbb{R}_+$: the account balance at time t .
- Shares $\mathbf{h}_t \in \mathbb{Z}_+^n$: the number of shares for n stocks at t .
- Closing price $\mathbf{p}_t \in \mathbb{R}_+^n$: the closing prices of n stocks at t .
- Technical indicators help the agent make decisions. Users can select existing indicators or add new indicators. E.g., Moving Average Convergence Divergence (MACD) $\mathbf{M}_t \in \mathbb{R}^n$, Relative Strength Index (RSI) $\mathbf{R}_t \in \mathbb{R}_+^n$, Commodity Channel Index (CCI) $\mathbf{C}_t \in \mathbb{R}_+^n$, etc.

Action space \mathcal{A} describes the allowed actions an agent can take at states s_t , $t = 1, \dots, T$. For one stock, action is $a \in \{-k, \dots, -1, 0, 1, \dots, k\}$,

where $k \in \mathbb{Z}$ or $-k \in \mathbb{Z}$ denotes the number of shares to buy or sell, respectively, and $a = 0$ means to hold. Users can set a maximum number of shares h_{\max} for a transaction, i.e., $k \leq h_{\max}$, or set a maximum ratio of capital to allocate on each stock.

Reward r_t for taking action a_t at state s_t and arriving at state s_{t+1} . Reward is the incentive for an agent to improve its policy for the sake of getting higher rewards. A relatively simple reward can be defined as the change of the account value, i.e.,

$$r_t = (b_{t+1} + \mathbf{p}_{t+1}^T \mathbf{h}_{t+1}) - (b_t + \mathbf{p}_t^T \mathbf{h}_t) - c_t, \quad (2)$$

where the first and second terms are the account values at s_{t+1} and s_t , and c_t denotes the transaction cost (market friction).

Transition (s_t, a_t, r_t, s_{t+1}) . Taking action a_t at state s_t , the environment steps forward and arrives at state s_{t+1} . A transition involves the change of balance, number of shares, and the stock prices due to the market changes. We split the stocks into three sets: selling set S , buying set B and holding set H , respectively. The new balance is

$$b_{t+1} = b_t + (\mathbf{p}_t^S)^T \mathbf{k}_t^S - (\mathbf{p}_t^B)^T \mathbf{k}_t^B, \quad (3)$$

where $\mathbf{p}^S \in \mathbb{R}^n$ and $\mathbf{k}^S \in \mathbb{R}^n$ are the vectors of prices and number of selling shares for the selling stocks, and $\mathbf{p}^B \in \mathbb{R}^n$ and $\mathbf{k}^B \in \mathbb{R}^n$ for the buying stocks. The number of shares becomes

$$\mathbf{h}_{t+1} = \mathbf{h}_t - \mathbf{k}_t^S + \mathbf{k}_t^B \geq \mathbf{0}. \quad (4)$$

The supercomputing power is necessary to achieve the massively parallel simulations for an STP task. During the training, a DRL agent keeps observing and trading on the historical market data to sample trajectories (one trajectory is a series of transitions). However, the historical data has to be significantly large in order to provide a broadened horizon. For example, the historical data could scale up in two dimensions: the **data volume** and **data type** [16]:

- The data volume varies with respect to:
 - **the length of data period**: from several months up to more than ten years.
 - **the time granularity**: from daily-level to minute-level, second-level or microsecond-level.
 - **the number of stocks**: from thirty (Dow 30) to hundreds (NASDAQ 100 or S&P 500), or even covers the whole market.
- The data type varies with respect to:
 - **the raw market data** includes data points of open-high-low-close-volume (OHLCV) for each stock, which provides a direct understanding of a stock's market performance.
 - **the alternative data** usually refers to the large-scale collection of both structured and unstructured data, e.g., market news, academic graph data [4], credit card transactions and GPS traffic. The agent could employ different encoders to analyze the insights of investment techniques provided by the alternative data.
 - **the indexes and labels** could be directly given as a kind of powerful technical indicator, which helps the agent make decisions.

In practice, the market simulation, alternative data processing and index analyzing are computationally expensive, therefore, a cloud-level solution is critical to a fast iteration of a trading strategy.

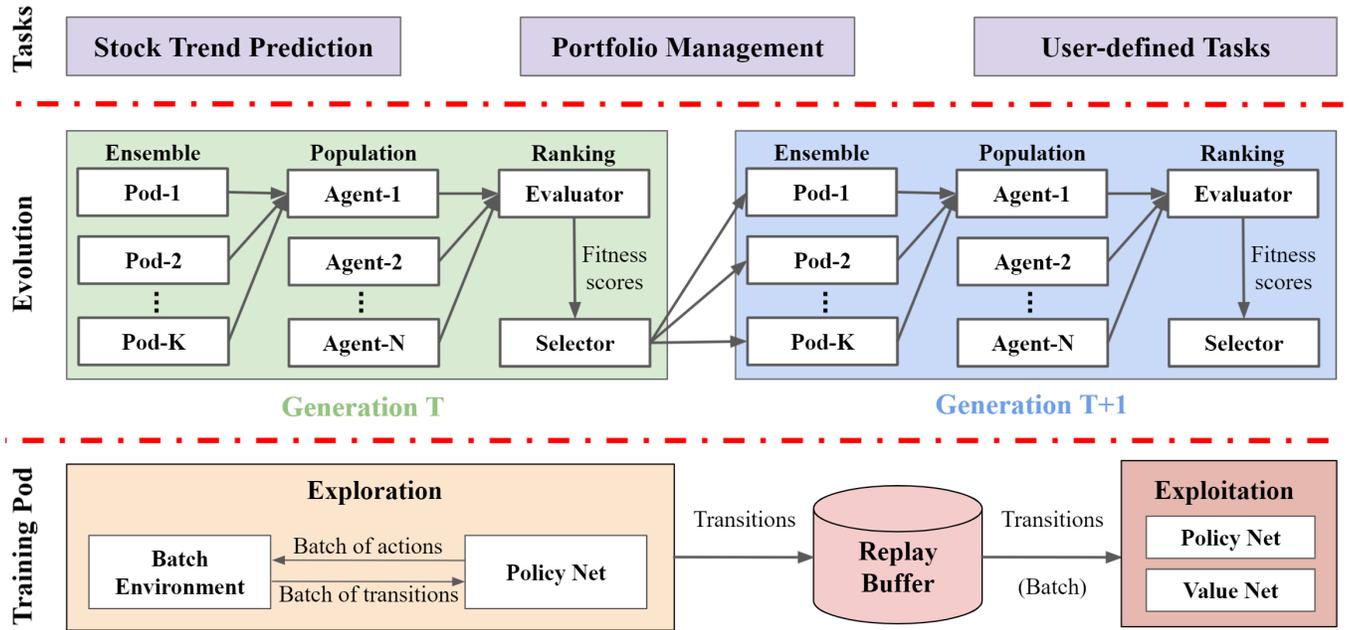


Figure 2: Overview of *FinRL-Podracr* that has three layers: trading task layer, evolution layer and training layer.

4 FINRL-PODRACER FRAMEWORK

We propose a *FinRL-Podracr* framework to utilize the supercomputing power of a GPU cloud for training DRL-driven trading strategies. We first present an overview of *FinRL-Podracr* and then describe its layered architecture.

4.1 Overview

Based on the experiments in Table 1, we found that existing DRL libraries/packages [13, 19, 25, 26] have three major issues that restrict the trading performance and training efficiency:

- There is no criteria to determine **overfitting or underfitting of models (trading strategies)** during the training process. It is critical to overcome underfitting by utilizing more computing power and avoid overfitting that wastes computing power, while both cases would lead to suboptimal models.
- The training process of a trading strategy is **sensitive to hyper-parameters**, which may result in unstable trading performance in backtesting and trading. However, it is tedious for human traders to search for a good combination of hyper-parameters, and thus an automatic hyper-parameter search is favored.
- **Computing power is critical** to effectively explore and exploit large-scale financial data. Sufficient exploration guarantees a good trading performance, and then smart exploitation results in good training efficiency. A strong computing power helps achieve a balance between exploration and exploitation.

Therefore, we provide a high performance and scalable solution on a GPU cloud, *FinRL-Podracr*, to develop a profitable DRL-driven trading strategy within a small time window. To fully utilize a GPU cloud, say an NVIDIA DGX SuperPod cloud [38], we organize *FinRL-Podracr* into a three-layer architecture in Fig. 2, a trading

task layer on the top, an evolution layer in the middle and a training layer at the bottom.

In the evolution layer, we employ a generational evolution mechanism with the ensemble strategy and address the issues of overfitting and hyper-parameter sensitivity through the synergy of an *evaluator* and a *selector*. The evaluator computes the fitness scores $J(\pi_\theta)$ in (1) of a population of N agents and mitigates the performance collapse caused by overfitting. The hyper-parameter search is automatically performed via *agent evolution*, where the selector uses the fitness scores to guide the search direction. An effective cloud-level evolution requires a high-quality and scalable scheduling, therefore we schedule a population of parallel agents through a multi-level mapping.

In the training layer, we realize high-performance GPU-oriented optimizations of a decomposable DRL training pipeline. We locally optimize each component (a container within a pod), namely explorer, replay buffer, and learner, through parallelism encapsulation, GPU acceleration, efficient parameter fusion, and storage optimization. Thus, we maximize the hardware usage and minimize the communication overhead, which allows each component to be efficiently executed on a GPU cloud.

Such an evolution-and-training workflow pipelines the development of a trading strategy on a GPU cloud. It enjoys great performance and scalability, which promotes fast and flexible development, deployment and production of profitable DRL trading strategies.

4.2 Scalable Evolution Layer

FinRL-Podracr exploits a *generational evolution* mechanism with an ensemble strategy to coordinate the parallel agents and to automatically search the best hyper-parameters. For each generation,

it is composed of **model ensemble** and **population ranking**, as shown in the middle layer of Fig. 2. At present, we utilize an *evaluator* and a *selector* to schedule the agent evolution, where more modules can be incorporated, e.g., a monitor, an allocator, etc.

The evaluator evaluates agents and provides their fitness scores as the metric for the future ranking, as shown in the ranking stage of Fig. 2. From our observations, it is difficult for users to use existing libraries [13, 19, 25, 26] to train a profitable trading strategy because the overfitting agent may be treated as the best agent as the training process moves forward. When the dataset scales up, we need to increase the training time/steps to fully explore the large-scale data, making it harder to set appropriate stop criteria, and the resulting agent may hardly be the best one. The evaluator effectively mitigates the performance collapse brought by overfitting: in the course of the training, it evaluates the agent at each iteration, outputs a fitness score, and keeps track of the best agent so far; when the fitness score in (1) drops, the evaluator would stop the training process using the early stopping mechanism and output the best agent as the final agent.

The selector acts as a central controller to perform the selection strategy as in a genetic algorithm (GA) [29]. GA is an optimization algorithm inspired by natural evolution: at every generation, a population of N agents is trained, and the evaluator calculates their fitness scores in (1) based on an objective function; then the selector redistributes the agents with the highest scores to form a new population for the next generation. Since the agents are parallel and replicable, the concept of natural selection scales up well on a GPU cloud. As shown in the evolution layer of Fig. 2, there are N agents with different hyper-parameters in a population. The synergy of the evaluator and selector enables FinRL-Podracr to naturally select the best agent for the future generation and eliminates the potential negative impact from poorly evolved agents, which effectively improves the stability and efficiency of the training.

FinRL-Podracr achieves the ensemble training of an agent by concurrently running K pods (training processes) in parallel and fusing the trained models from K pods at each epoch. All parallel pods for each agent are initialized with same hyper-parameters but different random seeds. Such a design, as shown in the ensemble stage of the evolution layer in Fig. 2, guarantees randomness and stabilizes the learning process of the agent. The experiment results in Section 5 will perform an ablation study of the improvement brought by the generational evolution mechanism.

4.3 Packaging Worker-Learner into Pod

FinRL-Podracr achieves effective and efficient allocation of cloud resources through a multi-level mapping, which follows the principle of *decomposition*-and-*encapsulation*. We employ a worker-learner decomposition [10–12] that splits a DRL training process (pod) into three components (containers):

- **Worker (exploration)**: samples transitions through the actor-environment interactions.
- **Replay buffer**: stores transitions from a worker and feeds a batch of transitions to a learner.
- **Learner (exploitation)**: consumes transitions and trains the neural networks.

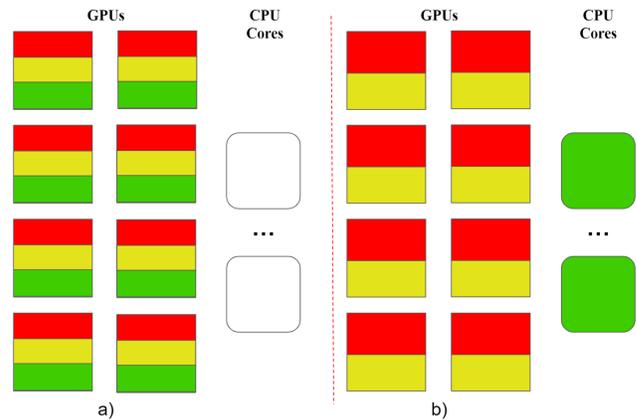


Figure 3: Two implementations of a training process: a) the environment simulation (green), action inference (yellow), and model update (red) are all located on GPUs. b) the environment simulation is executed on CPUs, and action inference and model update are on GPUs. An NVIDIA DGX-A100 server [38][5] contains 8 A100 GPUs.

Each training process of an agent consists of the three types of components, which are packaged into a suite that is mapped into a GPU pod. In addition, We run those components separately where each component is mapped to a GPU container. Such a two-level mapping is natural since a GPU pod consists of multiple containers, while correspondingly a training process of an agent consists of different components.

The above pod-container structure enables scalable allocation of GPU computing resources. We take advantage of a GPU cloud software stack and use the Kubernetes (K8S) software to scalably coordinate pods among servers. Consider a cloud with 10 servers (i.e., 80 A100 GPUs), we encapsulate a package of components into a pod, replicate it 80 times, and send them to a K8S server. Then, K8S distributes these 80 pod replicas to computing nodes that carry out the training process. The pod replication reflects strong parallelism, and it is highly scalable since a GPU cloud can support a large number of pods.

4.4 High Performance Training Layer

The optimization of each component is critical to the overall performance. We describe the hardware-oriented optimizations of components, including parallelism encapsulation, GPU acceleration, efficient parameter fusion and storage optimization.

Market simulation with GPU-acceleration. The market simulation is both computing- and communication-intensive. We propose a batch mode to perform massively parallel simulations, which maximizes the hardware utilization (either CPUs or GPUs). We instantiate multiple independent sub-environments in a batched environment, and a batched environment is exposed to a rollout worker that takes a batch of actions and returns a batch of transitions.

Fig. 3 a) illustrates a GPU-accelerated environment. Environments of financial tasks are highly suitable to GPUs because financial simulations involve "simple" arithmetics, where a GPU with

Hyper-parameters	Value
Total #GPUs	80
#Agent (N)	10
#Pods per agent (K)	8
Optimizer	Adam
Learning rate	2^{-14}
Discount factor	$\gamma = 0.99$
Total steps	2^{20}
Batch size	2^{10}
Repeat times	2^3
Replay buffer Size	2^{12}
Ratio clip (PPO)	0.25
Lambda entropy (PPO)	0.02
Evaluation interval (second)	64

Table 2: Hyper-parameter settings in our experiments.

thousands of cores has the natural advantages of matrix computations and parallelism. Then, financial environments written in CUDA can speed up the simulation. The GPU-accelerated environment also effectively reduces the communication overhead by bypassing CPUs, as supported by a GPU cloud [38]. The output transitions are stored as a tensor in GPU memory, which can be directly fetched by learners, avoiding the data transfer between CPU and GPU back and forth.

Fig. 3 b) presents an environment on CPUs. There are some financial simulations with frequent CPU usage (addressing trading constraints), making it inefficient to compute on GPUs. In our experiments, some environments run much slower on GPUs than CPUs. Thus, we simulate those environments on CPUs.

Replay buffer on GPU. We allocate the replay buffer on the contiguous memory of GPUs, which increases the addressing speed and bypasses CPUs for faster data transfer. As the worker and learner are co-located on GPUs, we store all transitions as tensors on the contiguous memory of GPUs. Since the collected transitions are packed together, the addressing speed increases dramatically well when a learner randomly samples a batch of transitions to update network parameters.

Learner with optimizations. To better support the ensemble training in the evolution layer, we propose a novel and effective way for learners of each pod to communicate, i.e., sending the network parameters rather than the gradients. Most existing libraries [13, 19, 25, 26] send the gradients of learners by following a traditional synchronization approach on supervised learning. Such an approach is inefficient for DRL algorithms since the learner will update the neural network hundreds of times within each training epoch, namely it needs to send gradients hundreds of times. By taking advantage of the soft update [21], we send the model parameters rather than the gradients. The parameter of the models is amenable to communication because model size in DRL is not comparable to that in other deep learning fields. Here, communication happens once at the end of each epoch, which is a significantly lower frequency of communication.

5 PERFORMANCE EVALUATION

We describe the GPU cloud platform, the performance metrics and compared methods, and then evaluate the performance of FinRL-Podracr for a stock trend prediction task.

5.1 GPU Cloud Platform

All experiments were executed on NVIDIA DGX-2 servers [5] in an NVIDIA DGX SuperPOD platform [38], a cloud-native super-computer. We use 256 CPU cores of Dual AMD Rome 7742 running at 2.25GHz for each experiment. An NVIDIA DGX-2 server has 8 A100 GPUs and 320 GB GPU memory [5].

5.2 Performance Metrics

We evaluate the trading performance and training efficiency of the FinRL-Podracr for a stock trend prediction task.

Data pre-processing. We select the NASDAQ-100 constituent stocks as our stock pool, accessed at 05/13/2019 (the starting time of our testing period), and use the datasets with two time granularities: minute-level and daily. The daily dataset is directly downloaded from Yahoo!Finance, while the minute-level dataset is first downloaded as raw data from the Compustat database through the Wharton Research Data Services (WRDS) [33] and then pre-processed to an open-high-low-close-volume (OHLCV) format. We split the datasets into training period and backtesting period: the daily data from 01/01/2009 to 05/12/2019 for training; the minute-level data from 01/01/2016 to 05/12/2019 for training; For both datasets, we backtest on the same period from 05/13/2019 to 05/26/2021.

Evaluation metrics. Six common metrics are used to evaluate the experimental results:

- **Cumulative return:** subtracting the initial value from the final portfolio value, then dividing by the initial value.
- **Annual return and volatility:** geometric average return in a yearly sense, and the corresponding deviation.
- **Sharpe ratio:** the average return earned in excess of the risk-free rate per unit of volatility.
- **Max drawdown:** the maximum observed loss from a historical peak to a trough of a portfolio, before a new peak is achieved. Maximum drawdown is an indicator of downside risk over a time period.
- **Cumulative return vs. training time:** the cumulative return during the testing period, achieved by an agent trained within a certain amount of time.

Compared methods. For trading performance evaluation, we compare FinRL-Podracr and vanilla FinRL-Podracr (without agent evolution) with FinRL [25, 26], RLlib [19], Stable Baseline3 [9], and NASDAQ Composite/Invesco QQQ ETF. We use Proximal Policy Optimization (PPO) [32] as the DRL algorithm in the reported results and fine-tune each library to maximize its performance. Each library is allowed to use up to 80 GPUs.

For training efficiency evaluation, the experiments are conducted on multiple GPUs. We compare with RLlib [19] since it has high performance on distributed infrastructure. However, both FinRL [25] and Stable Baseline 3 [9] do not support the training on multiple GPUs, thus we do not compare with them. We keep hyper-parameters and computing resources the same to guarantee fair

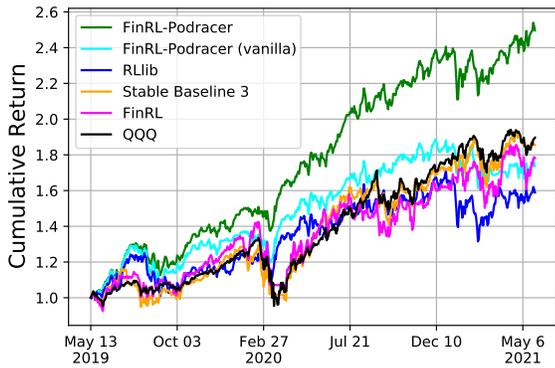


Figure 4: Cumulative returns on daily dataset during 05/13/2019 to 05/26/2021. Initial capital \$1,000,000, transaction cost percentage 0.2%, and Invesco QQQ ETF is a market benchmark.

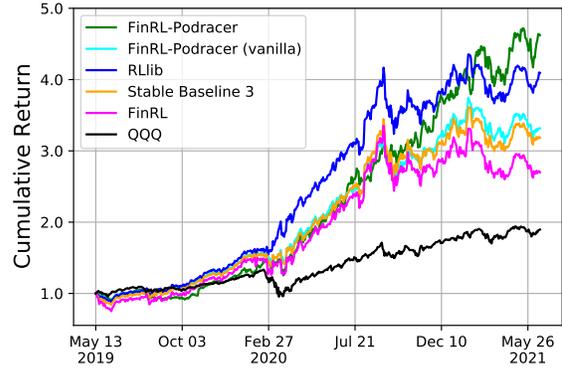


Figure 5: Cumulative returns on minute dataset during 05/13/2019 to 05/26/2021. Initial capital \$1,000,000, transaction cost percentage 0.2%, Invesco QQQ ETF is a market benchmark.

	Cumul. return	Annual return	Annual volatility	Max drawdown	Sharpe ratio
FinRL-Podracer (Ours)	149.553%/362.408%	56.431%/111.549%	22.331%/33.427%	-13.834%/-15.874%	2.12/2.42
FinRL-Podracer (vanilla)	73.546%/231.747%	30.964%/79.821%	23.561%/31.024%	-18.428%/-21.002%	1.27/2.05
RLlib [19]	58.926%/309.54%	25.444%/99.347%	30.009%/31.893%	-23.248%/-22.292%	0.91/2.33
Stable Baseline3 [13]	85.539%/218.531%	35.316%/76.28%	31.592%/34.595%	-24.056%/-23.75%	1.12/1.82
FinRL [25]	78.255%/169.975%	32.691%/62.576%	37.641%/42.908%	-26.774%/-27.267%	0.94/1.35
Invesco QQQ ETF	89.614%	36.763%	28.256%	-28.559%	1.25

Table 3: Performance of stock trading on NASDAQ-100 constituent stocks with daily (red) and minute-level (blue) data.

comparisons, and a general hyper-parameter setting is given in Table 2.

5.3 Trading Performance

We backtest the trading performance from 05/13/2019 to 05/26/2021 on both daily and minute-level datasets. From Fig. 4 and Fig. 5, all DRL agents are able to achieve a better or equal performance than the market in cumulative return, which demonstrates the profit potentials of DRL-driven trading strategies. Comparing Fig. 4 with Fig. 5, we observe that all methods have a much better performance on the minute-level dataset than that on the daily dataset. The trading performance of most agents is almost the same as that of the market on daily dataset, however, all agents significantly outperform the market if they have a larger dataset to explore. With a higher granularity data, the Sharpe ratios are also lifted up to a new level. From Table 3, agents achieve Sharpe ratios of 2.42, 2.05, 2.33, 1.82, 1.35 on the minute-level dataset, which are 0.3, 0.78, 1.42, 0.7, and 0.41 higher than those on the daily dataset. Therefore, we conclude that the capability to process large-scale financial data is critical for the development of a profitable DRL-driven trading strategy since the agent can better capture the volatility and dynamics of the market.

From Table 3, Fig. 4, and Fig. 5, we also observe that our FinRL-Podracer outperforms other baselines on both datasets, in terms of expected return, stability, and Sharpe ratio. As can be seen from Table 3, our FinRL-Podracer achieves the highest cumulative returns

of 149.533% and 362.408%, annual returns of 56.431% and 111.549%, and Sharpe ratios of 2.12 and 2.42, which are much higher than the others. Furthermore, FinRL-Podracer also shows an outstanding stability during the backtesting: it achieves Max drawdown -13.834% and -15.874%, which is much lower than other methods. Consider the vanilla FinRL-Podracer as a direct comparison, we find that vanilla FinRL-Podracer has a similar trading performance with other baseline frameworks, which is in consistent with our expectation since the settings are the same. Such a performance improvement of FinRL-Podracer over vanilla FinRL-Podracer demonstrates the effectiveness of the generational evolution mechanism, as further verified by Fig. 7.

5.4 Training Efficiency

We compare the training efficiency of FinRL-Podracer with RLlib [19] on a varying number of A100 GPUs, i.e., 8, 16, 32, and 80. We store the model snapshots at different training time, say every 100 seconds, then later we use each snapshot model to perform inference on the backtesting dataset and obtain the generalization performance, namely, the cumulative return.

In Fig. 6, as the number of GPUs increases, both FinRL-Podracer and RLlib achieve a higher cumulative return with the same training time (wall-clock time). FinRL-Podracer with 80 GPUs has a much steeper generalization curve than others, e.g., it can achieve a cumulative return of 4.0 at 800s, which means it learns in a much faster speed. However, FinRL-Podracer with 32 GPUs and 16 GPUs need

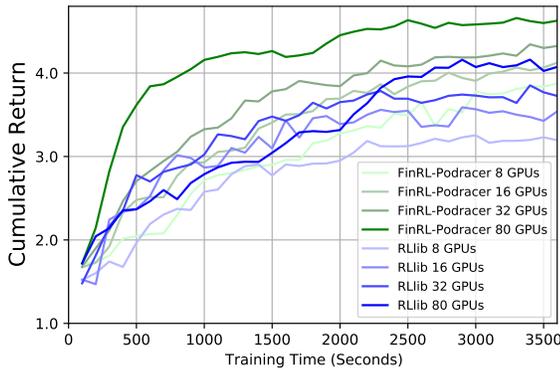


Figure 6: Generalization performance on backtesting dataset, using the model snapshots of FinRL-Podracer and RLib [19] at different training time (wall clock time).

2, 200s and 3, 200s to achieve the same cumulative return, respectively. The generalization curves of RLib with different numbers of GPUs are relatively similar, and we do not observe much speed-up. For example, FinRL-Podracer needs approximately 300s to achieve a cumulative return of 3.5, however, RLib needs 2, 200s to achieve the same cumulative return. FinRL-Podracer is $3\times \sim 7\times$ faster than RLib.

It is counter-intuitive that the increase of GPU resources not only makes FinRL-Podracer have a fast training, but also improves the trading performance over RLib [19]. We know from Fig. 4 and Fig. 5 that the generational evolution mechanism promotes the trading performance of FinRL-Podracer, therefore, we empirically investigate the agent evolution process. Fig. 7 explicitly demonstrates an evolution of $N = 10$ agents, where the selector chooses the best model to train in the next generation every 800s. The inner figure of Fig. 7 depicts the generalization curves of the ten agents in the first generation (without using the agent evolution mechanism). The curve with the evolution mechanism (the thick green curve) is substantially higher than the other ten curves.

6 DISCUSSION AND CONCLUSION

In this paper, we have proposed a high-performance and scalable deep reinforcement learning framework, *FinRL-Podracer*, to initiate a paradigm shift from conventional supervised learning approaches to *RL Ops in finance*. FinRL-Podracer provides a highly automated development pipeline of DRL-driven trading strategies on a GPU cloud, which aims to help finance researchers and quantitative traders overcome the steep learning curve and take advantage of supercomputing power from the cloud platforms.

FinRL-Podracer achieved promising performance on a cloud platform, mainly by following the two principles, *the virtues of nested hierarchies* and *getting smart from dumb things* [15]. For low-level training, FinRL-Podracer realizes nested hierarchies by employing hardware-oriented optimizations, including parallelism encapsulation, GPU acceleration, and storage optimizations. As a high level scheduling, FinRL-Podracer obtains a smart agent from hundreds of weak agents, which is the essence of ensemble methods, by

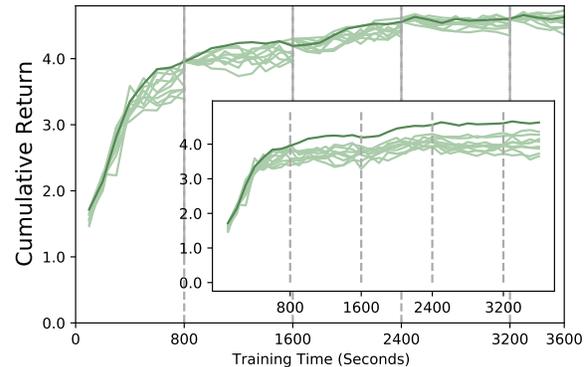


Figure 7: Generalization performance of a model along the agent evolution in the training process of FinRL-Podracer.

employing a generational evolution mechanism. We further investigate the evolution and training layers in a followup work [24] for a cloud-native solution. We believe that ensemble multiple weak agents is preferable to aiming to train one strong agent. Thus we propose a new orchestration mechanism, a tournament-based ensemble training method [24] with asynchronous parallelism, which involves relatively low communication overhead. Also, we observe the great potential of massively parallel simulation, which lifts the exploration capability up into a potentially new dimension.

FinRL-Podracer is our first step from building a standard DRL pipeline of financial tasks to using DRL agents to understand the dynamics of the markets. We believe that FinRL-Podracer is critical for the ecosystem of the FinRL community [25, 26] because it offers opportunities for many future directions. First, FinRL-Podracer provides a way to take advantage of large-scale financial data. It is possible to allow DRL agents to work in second or microsecond level and cover all stocks in the market, which is meaningful for the exploration and understanding of the dynamics of the market. Moreover, training on the cloud makes DRL agents adapt to much more complex financial simulations and neural networks, thus achieving wider DRL applications to various financial tasks, e.g., portfolio allocation, fraud detection, DRL-driven insights for yield improvement and optimization. Furthermore, the low-level optimizations in FinRL-Podracer could be also useful for the future development of financial simulators, such as using GPU-accelerated techniques to reduce latency.

ACKNOWLEDGEMENT

This research used computational resources of the GPU cloud platform [38] provided by the IDEA Research institute.

REFERENCES

- [1] Sridhar Alla and Suman Kalyan Adari. 2021. What Is MLOps? In *Beginning MLOps with MLFlow*. Springer, 79–124.
- [2] H. Buehler, L. Gonon, J. Teichmann, and B. Wood. 2019. Deep hedging. *Quantitative Finance* 19 (2019), 1271 – 1291.
- [3] L. Chen and Qiang Gao. 2019. Application of deep reinforcement learning on automated stock trading. *IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)* (2019), 29–33.

- [4] Qian Chen and Xiao-Yang Liu. 2020. Quantifying ESG alpha using scholar big data: an automated machine learning approach. *Proceedings of the First ACM International Conference on AI in Finance* (2020).
- [5] Jack Choquette, Wishwesh Gandhi, Olivier Giroux, Nick Stam, and Ronny Krashinsky. 2021. NVIDIA A100 tensor core GPU: Performance and innovation. *IEEE Micro* 41, 2 (2021), 29–35.
- [6] Google Cloud. 2020. MLOps: Continuous delivery and automation pipelines in machine learning. https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning#mlops_level_0_manual_process. Google Cloud, Jan. 07, 2020.
- [7] Giacomo Corbo, Oliver Flemin, and Nicolas Hohn. 2021. It's time for businesses to chart a course for reinforcement learning. <https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/its-time-for-businesses-to-chart-a-course-for-reinforcement-learning>. *McKinsey Analytics*, April, 01, 2021.
- [8] Quang-Vinh Dang. 2019. Reinforcement learning in stock trading. *ICCSAMA* (2019).
- [9] DLR-RM. 2021. Stable-baseline 3. <https://github.com/DLR-RM/stable-baselines3>.
- [10] Lasse Espeholt, Raphaël Marinier, Piotr Stanczyk, Ke Wang, and Marcin Michalski. 2020. SEED RL: scalable and efficient deep-RL with accelerated central inference. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- [11] Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Volodymyr Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. 2018. IMPALA: scalable distributed deep-RL with importance weighted actor-learner architectures. In *Proceedings of the International Conference on Machine Learning (ICML)*.
- [12] Matteo Hessel, Manuel Kroiss, Aidan Clark, Iurii Kemaev, John Quan, Thomas Keck, Fabio Viola, and Hado van Hasselt. 2021. Podracer architectures for scalable reinforcement learning. *ArXiv abs/2104.06272* (2021).
- [13] Ashley Hill, Antonin Raffin, Maximilian Ernestus, Adam Gleave, Anssi Kanervisto, Rene Traore, Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, and Yuhuai Wu. 2018. Stable baselines. <https://github.com/hill-a/stable-baselines>.
- [14] Zhengyao Jiang, Dixing Xu, and Jinjun Liang. 2017. A deep reinforcement learning framework for the financial portfolio management problem. *ArXiv abs/1706.10059* (2017).
- [15] Kevin Kelly. 1994. *Out of control: The rise of neo-biological civilization*. Addison-Wesley Longman Publishing Co., Inc.
- [16] Marko Kolanovic and Rajesh T. Krishnamachari. 2017. Big data and AI strategies: machine learning and alternative data approach to investing. <https://www.cognitivefinance.ai/single-post/big-data-and-ai-strategies>. *J.P. Morgan Securities LLC*, May, 18, 2017.
- [17] Petter N. Kolm and G. Ritter. 2019. Modern perspectives on reinforcement learning in finance. *Econometrics: Mathematical Methods & Programming eJournal* (2019).
- [18] Xinyi Li, Yinchuan Li, Yuancheng Zhan, and Xiao-Yang Liu. 2019. Optimistic bull or pessimistic bear: Adaptive deep reinforcement learning for stock portfolio allocation. *ICML Workshop on Applications and Infrastructure for Multi-Agent Learning* (2019).
- [19] Eric Liang, Richard Liaw, Robert Nishihara, Philipp Moritz, Roy Fox, Joseph Gonzalez, Ken Goldberg, and Ion Stoica. 2017. Ray RLLib: a composable and scalable reinforcement learning library. *ArXiv abs/1712.09381* (2017).
- [20] Jacky Liang, Viktor Makovychuk, A. Handa, N. Chentanez, M. Macklin, and D. Fox. 2018. GPU-accelerated robotic simulation for distributed reinforcement learning. In *Conf. on Robot Learning (CoRL)*.
- [21] T. Lillicrap, Jonathan J. Hunt, A. Pritzel, N. Heess, T. Erez, Yuval Tassa, D. Silver, and Daan Wierstra. 2016. Continuous control with deep reinforcement learning. *CoRR abs/1509.02971* (2016).
- [22] Paul Lipton, Derek Palma, Matt Rutkowski, and Damian Andrew Tamburri. 2018. TOSCA solves big problems in the cloud and beyond! *IEEE Cloud Computing* (2018), 1–1. <https://doi.org/10.1109/MCC.2018.111121612>
- [23] Xiao-Yang Liu, Zechu Li, Zhaoran Wang, and Jiahao Zheng. 2021. ElegantRL: A Scalable and Elastic Deep Reinforcement Learning Library. <https://github.com/AI4Finance-Foundation/ElegantRL>.
- [24] Xiao-Yang Liu, Zechu Li, Zhuoran Yang, Jiahao Zheng, Zhaoran Wang, Anwar Walid, Jiang Guo, and Michael Jordan. 2021. ElegantRL-Podracr: Scalable and elastic library for cloud-native deep reinforcement learning. *Deep Reinforcement Learning Workshop at NeurIPS* (2021).
- [25] Xiao-Yang Liu, Hongyang Yang, Qian Chen, Runjia Zhang, Liuqing Yang, Bowen Xiao, and Christina Dan Wang. 2020. FinRL: a deep reinforcement learning library for automated stock trading in quantitative finance. *Deep Reinforcement Learning Workshop at NeurIPS* (2020).
- [26] Xiao-Yang Liu, Hongyang Yang, Jiechao Gao, and Christina Dan Wang. 2021. FinRL: Deep reinforcement learning framework to automate trading in quantitative finance. *ACM International Conference on AI in Finance (ICAIF)* (2021).
- [27] Viktor Makovychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. 2021. Isaac Gym: High performance GPU-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470* (2021).
- [28] Rick Merritt. 2020. What Is MLOps? <https://blogs.nvidia.com/blog/2020/09/03/what-is-mlops/>. NVIDIA, Sep. 03, 2020.
- [29] Melanie Mitchell. 1996. An introduction to genetic algorithms.
- [30] V Mnih, K Kavukcuoglu, D Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540) (2015), 529–533.
- [31] G. Nuti, Mahnoosh Mirghaemi, P. Treleaven, and Chaiyakorn Yingsaeree. 2011. Algorithmic trading. *Computer* 44 (2011), 61–69.
- [32] John Schulman, F. Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *ArXiv abs/1707.06347* (2017).
- [33] Wharton Research Data Service. 2015. Standard & poor's compustat. Data retrieved from Wharton Research Data Service.
- [34] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484–489.
- [35] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529(7587) (2016), 484–489.
- [36] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. *nature* 550, 7676 (2017), 354–359.
- [37] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.
- [38] NVIDIA DGX A100 system reference architecture. 2020. *NVIDIA DGX SuperPOD: Scalable infrastructure for AI leadership*. NVIDIA Corporation.
- [39] Damian A. Tamburri. 2020. Sustainable MLOps: Trends and challenges. In *2020 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*. 17–23.
- [40] P. Treleaven, M. Galas, and V. Lalchand. 2013. Algorithmic trading review. *Commun. ACM* 56 (2013), 76–85.
- [41] Google Trends. 2021. What Is MLOps? <https://www.google.com/trends>.
- [42] Jia Wu, Chen Wang, Lidong Xiong, and Hongyong Sun. 2019. Quantitative trading on stock market based on deep reinforcement learning. In *2019 International Joint Conference on Neural Networks (IJCNN)*. 1–8.
- [43] Zhuoran Xiong, Xiao-Yang Liu, Shan Zhong, Hongyang Yang, and Anwar Walid. 2018. Practical deep reinforcement learning approach for stock trading. *NeurIPS Workshop* (2018).
- [44] Hongyang Yang, Xiao-Yang Liu, Shan Zhong, and Anwar Walid. 2020. Deep reinforcement learning for automated stock trading: An ensemble strategy. *ACM International Conference on AI in Finance (ICAIF)* (2020).
- [45] Zihao Zhang, Stefan Zohren, and Stephen Roberts. 2020. Deep reinforcement learning for trading. *The Journal of Financial Data Science* 2(2) (2020), 25–40.