EMPIRICAL STEP REWARD FOR RL-BASED SYMBOLIC REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

TD-based advantage estimation (For example, one step forward return $Q^\pi(s_t, a_t) = E_\pi[r_t + \gamma Q(s_{t+1}, a_{t+1})]$) can be biased when symbolic regression, because the optimal policy is non-ergodic, and there is only terminal reward($Q^\pi(s_0, a_0) = E_\pi[\gamma^\tau r_\tau]$). For example, dataset choice can be important, and wrong symbol choice can be not recoverable in symbolic regression. On the other hand, Monte-Carlo sampling $\gamma^{\tau-t} r_\tau - V(s_t)$ shows high variance which makes training difficult. Tail distribution can be used(Petersen et al., 2021) to partially mitigate this limitation. In our research, intermediate fitness of the equation is obtained for each step, which enables calculation of empirical reward for each step, defined by improvement of the fitness $\phi(s_{t+1}) - \phi(s_t)$. Our method is based on virtual stack machine, which maps each symbol to stack state update. Our empirical step reward can be seen as finding useful heuristic expressions, which is inspired by traditional genetic programming. Our method was verified using financial factor search task, with adjustment for multiple testing bias. Increased diversity of the symbolic expression set is observed in mean-variance optimized portfolio return.

1 INTRODUCTION

Interpretable concepts are required to explain the observations in a scientific way. Traditional regression analysis requires prior knowledge or assumptions about the relationship of the observations, usually in the form of equations. On the other hand, the Symbolic Regression tries to automatically discover useful combinations of operators and data that explain the observations.

Symbolic regression provides information on symbolic relevance (La Cava et al., 2021), and it provides interpretability for both humans and machines. In other words, symbolic regression try to see the system as a compact symbolic representation like a game(Schaeffer et al., 1996; Schaeffer, 2001). From this point of view, the challenge of symbolic regression is the size of the symbolic combination space, which makes the substantive rationality (Simon, 1976) difficult to achieve using exhaustive search.

Genetic Programming(GP)-based Symbolic Regression (Koza, 1994) reduces the search dimension by projecting combination of symbols (genotype space) into property of the equations (phenotype space), based on heuristic rules such as mutation and crossover. These rules are based on the procedure of building pools of equations for each task. Each equation in a pool can be seen as a particle in Particle Swarm Optimization(PSO), the solution diversity in the pool is well-defined and maintained (Blackwell, 2005; Blackwell et al., 2008). One shortcoming of GP-based symbolic regression is that it can be time consuming to build a pool for each different search space, and the heuristics need careful design.

Neural net methods can embed data into non-symbolic numeric latent space (Bishop et al., 1995), and the neural net can be used to approximate the policy function, combined with efficient sampling and estimations methods or policy gradient methods. (Sutton et al., 1998) Neural net based symbolic can be purely based on numeric embedding of symbols and dataset distribution ('end-to-end' SR by Kamienny et al. (2022)), or it can be more estimation of non-symbolic policy under symbolic environment ('deep SR' by Petersen et al. (2021)). 'End-to-end' SR (Kamienny et al., 2022) inference is extremely efficient, trained to fit the in-sample distribution. The reason why this architecture can map symbolic understanding into the latent vector in general is still under research. (Geirhos

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2023; Hewitt et al., 2025; Adebayo et al., 2018) 'Deep' SR (Petersen et al., 2021) try to see equation writing process as a Markov Chain, and optimize the state embedding and the policy using the policy gradient. As a result symbolic space can be efficiently sampled without heuristic rules, and the out-sample performance is relatively higher.

Symbolic regression(SR) has been used for factor research in the field of empirical asset pricing, and a possible example is popularized by Kakushadze (2016); Tulchinsky (2019). SR enables human researcher to check if the equation is interpretable as a financial factors, by providing information on symbolic relevance. For example, subtracting sector average can be interpreted as neutralizing the industry risk, while selecting from a day of week is usually noise overfitting without portfolio diversification. In this case, symbolic relevance reveals true diversity better than method based on historical return, by providing data and operator usage.

In contrast to symbolic regression for scientific application(La Cava et al., 2021), diversity is one of the main qualities of financial research rather than accuracy and simplicity. For GP based SR, the diversity of the expression pool can be defined as the diameter of the solution set in the genotype space, similarly to the solution diversity in PSO (Blackwell, 2005). For a dynamic optimization problem (it is different from dynamic programming; Blackwell (2005)), maintaining solution diversity helps out-sample robustness of the solution, by providing multiple local optima as candidates that can work after unpredictable change of the system. Sharpe ratio ($\frac{E[r]}{\text{std(r)}}$) is a common measure of performance where r is excess return. It formularizes the effect of diversification as a trade-off relationship between profitability fitness and robustness enhancement. Sharpe ratio estimations on the historical backtests are often used for portfolio optimization, and out-sample Sharpe ratio can be used for evaluation purposes.

In our research RL based symbolic regression for financial anomaly discovery is investigated. Financial market was chosen as our research subject because well-known alpha decay phenomenon Zhou & Lin (2017) is a clear example of a dynamic optimization problem (Blackwell, 2005) because the significance of market anomaly changes over time. Please note that our research is discovery of financial anomaly that is not explained by well-known risk premiums detected by Fama-Macbeth 2-pass regression. Anomalies discovered in our research may be non-tradable. For example, non-tradable anomalies include: shorting delisted instrument, trading financial restatement, split, dividend, high frequency patterns smaller than the transaction fee. Multiple-testing bias adjustments (Giglio et al., 2021) to test the significance of the market anomaly, to ensure that the equation is not a pure noise or a common factor.

Application of RL based SR with episodic reward (Petersen et al., 2021) for financial anomaly discovery resulted convergence of policy to a single pattern. We identify this problem as a credit assignment problem caused by the sparse reward computed only for the finished equation. To improve the diversity of equation that can be sampled from the policy, we propose step reward based on fitness improvement of intermediate signals in the stack machine. Result shows the improved diversity of equations. The result shows higher alpha significance than sampling the baseline policy in high temperature.

2 Related Works and Background

2.1 Solution Diversity in Genetic Programming

The problem of outdated memory and diversity loss in Particle Swarm Optimization(PSO) for dynamic environment was identified by Blackwell et al. (2008). In similar passion, Genetic Programming for Symbolic Regression, which is widely used in empirical asset pricing (Kakushadze, 2016), can experience alpha decay(Pénasse, 2022) due to market environment that dynamically adapts to public information.

Diversity measure for particle swarm can be defined as maximum distance between two solutions in a solution set(Blackwell, 2005), where the distance is Euclidean distance in a embedding space. Similarly, Diversity in symbolic regression was defined using edit distance, and its correlation with fitness was studied(Burke et al., 2004) to promote diversity in the domain of fitness function by ensuring diversity in the symbolic dimension. On the other hand, portfolio diversification effect in Kakushadze (2016) is based on the covariance matrix of the return time series, which is the domain

of the fitness function. In this case, symbolic regression can provide useful information about the diversity such as such as data usage of equation, which can be useful for empirical application.

2.2 Policy Diversity in Reinforcement Learning(RL)

RL based symbolic regression estimates the value function that maps symbolic sequence to the expected fitness using Monte Carlo sampling. On the other hand, risk-seeking policy gradient(Petersen et al., 2021) estimates $1 - \epsilon$ quantile instead of expectation value, which results efficient search of global optima. Other ML based methods for Symbolic Regression are provided in Section 1.

Independently, there are discussions(Parker-Holder et al., 2020) on policy diversity in multi-agent environment. They define policy distance in behavior embedding, which is action distribution matrix conditioned by probe states. Direct application of behavior embedding for symbolic regression is not well-known, and one possible reason is the difficulty of defining multi-agent environment for symbolic regression task. It is also worth noting that policy diversity, or 'algorithm herding' gained attention among financial researchers (Klein, 2021; Dou et al., 2025) especially for high-frequency trading. Relatively robust Q-table is often used in finance research, and the policy is directly compared.

In this research, the diversity for financial application was verified based on empirical knowledge. There are two main diversity in stock market signals : technical indicators and fundamental ratios. They exhibit different fitness characteristics, lookback periods, and data usage. Due to different fitness characteristics between these two classes, baseline estimation using elitist selection among batch sample can limit exploration of diverse classes. In simple words, we sample both technical indicators and fundamental ratios using our single trained policy, to verify the solution diversity.

2.3 Advantage Estimation

The bias-variance tradeoff relation between Monte Carlo estimation and the TD estimation for the policy gradient algorithm is well known. Generalized Advantage Estimation(GAE) (Schulman et al., 2015) balances this tradeoff relationship by controlling the discount rate γ and λ of TD(λ) estimation of the value function.

The GAE(Schulman et al., 2015) and the reward shaping (Ng et al., 1999) both mitigate the credit assignment problem, but the GAE let the response function to decompose the advantage function across the steps, while the reward shaping the reward against the barrier in the fitness landscape without altering the optimal policy. Our proposed method is similar to reward shaping, and we use empirical fitness of intermediate expression as the potential.

Our proposed method efficiently samples non-ergodic policy. Finding useful combinations from multiple financial dataset is highly non-ergodic, because the factors are almost orthogonal, and the fitness of the factor depends on the early stage of the equation writing (data choice, ratio combination, concepts) rather than the later stage (risk neutralization, selection, minor optimization).

2.4 Multiple Testing Bias in Empirical Asset Pricing

In the field of empirical asset pricing, there have been constant concerns about data snooping and multiple testing bias (White, 2000; Harvey et al., 2016). Misusing usual significance cutoffs to the "zoo of new factors" generated by Symbolic Regression, can introduce such problems. It means that in-sample portfolio performance is not a valid measure of optimality over the fitness function, when the portfolio is combination of equations generated by symbolic regression.

Similar to the idea of elitist selection(Petersen et al., 2021), False Discovery Rate (FDR) controlled hypothesis testing(Giglio et al., 2021) adjusts p-value based on cross-sectional information in the batch, to adjust the multiple testing effect. In this "taming the factor zoo"(Feng et al., 2020; Giglio et al., 2021) approach, Fama-MacBeth 2-pass regression is used to estimate the t-statistics of alpha(residual term of the factor model) $t = \frac{\hat{\alpha}}{\text{SE}(\hat{\alpha})}$ by removing the effect of latent or predefined factors.

We agree that our simple symbolic regression model lacks understanding of the factor structure of the financial market. That is, we expect the equations set from our model to be noisy rather

than diverse in the financial context. Therefore, we apply alpha selection based on FDR controlled hypothesis testing to the solution set, before evaluating the solution diversity.

3 METHODOLOGY

3.1 NETWORK ARCHITECTURE OF POLICY AND VALUE NETWORK

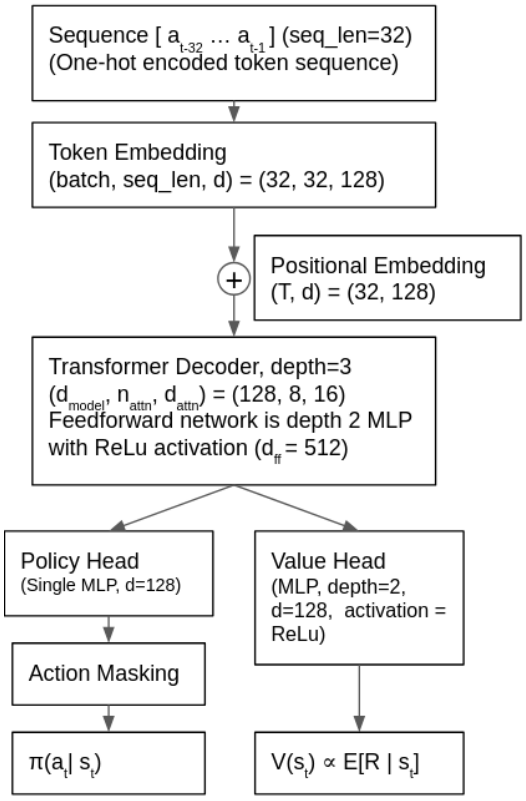

Figure 1: Dual-head causal Transformer architecture for symbolic policy optimization.

Our network architecture for the policy and value estimation is presented in Figure. 1. 2-layer Transformer decoder was used to approximate the autoregressive policy and the value function. The choice of architecture is based on recent wide application of transformer architecture, to the general machine learning including symbolic regression. The policy and value network takes shared transformer output as hidden representation. Output of the policy head is action logits, and the action masking (Petersen et al., 2021) adds $-\infty$ for invalid(type mismatch and stack underflow) actions, so the complete sequence are always syntactically valid. Output of value head estimates the value of current state. Policy network was trained using Proximal Policy Optimization (PPO), and the value network was trained to minimize Mean Squared error (MSE) loss.

3.2 ENVIRONMENT FOR REINFORCEMENT LEARNING ALGORITHM

Action is defined by symbols that can be appended to the equation sequence. Common operators for financial expressions (Yu et al., 2023), candlestick data for technical indicators(Tulip Indicators), commonly used financial accounts data(CFA Institute, 2024), various conventional lookback period constants in the unit of working days is included. For simplicity, the minimal action space for most frequent symbol was used in our financial benchmark. The action space is described in Eq. 1, and an example of action sequence is represented in Fig. 2.

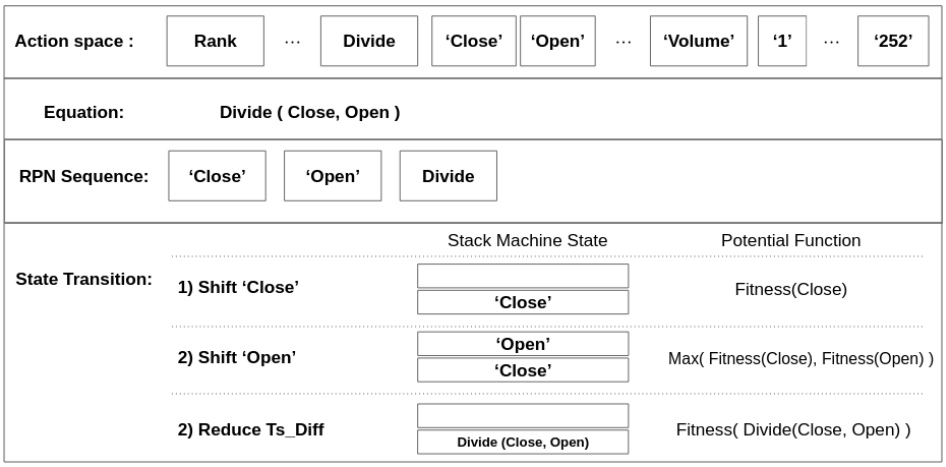

Figure 2: **Our proposed virtual stack machine environment for symbolic regression** Action space of our environment is symbol set. Symbol sequence in reverse Polish notation represents the equation. Bottom part of the figure illustrates the state transition of the virtual stack.

$$
\begin{aligned}
&[\texttt{done}, 1, 2, 5, 21, 64, 252,\\
&\texttt{add(x, y)}, \texttt{subtract(x, y)}, \texttt{divide(x, y)}, \texttt{cs_rank(x)}, \texttt{delta(x, t)}, \texttt{ts_mean(x, t)},\\
&\texttt{Close}, \texttt{High}, \texttt{Low}, \texttt{Volume}, \texttt{Close to Close return},\\
&\texttt{Revenue}, \texttt{Income}, \texttt{Asset}, \texttt{Long term Debt}]
\end{aligned}
\tag{1}
$$

State transition of our virtual stack based environment is illustrated in Fig. 2. Each actions are mapped to addition of operands or the reduction into sub-expression result. Compared to the conventional polish notation or parenthesized equations, our reverse polish notation can evaluate intermediate sub-expressions, which are useful for estimating potential function $\phi(s_t)$.

3.3 Potential-based Reward Shaping

Standard Proximal Policy Optimization(PPO) is our choice of the policy training algorithm, by following common choice. One minor modification is very weak clipping condition of $(0, e^2]$, which preserves the risk-seeking(Petersen et al., 2021) aspect of the policy gradient.

The reward function $r(s_t, a_t, s_{t+1})$ is defined as follows in Eq. 2, to mitigate the credit assignment problem of episodic reward for Symbolic Regression. It is similar to potential based reward shapingNg et al. (1999), where ϕ is the potential function.

$$
r(s_t, a_t, s_{t+1}) = \gamma\phi(s_{t+1}) - \phi(s_t) - c_{\text{step}} - c_{\text{unfinished}} \cdot \mathbf{1}_{\{t=T_{\max} \wedge s_{t+1} \notin \mathcal{S}_{\text{terminal}}\}} \tag{2}
$$

A small step penalty of $c_{\text{step}} = 0.02$ was introduced to prevent patterns without effect (for example, repeated `rank`), and a penalty for unfinished equation $c_{\text{unfinished}} = 10.0$ was introduced to prevent repeated data introduction.

The potential function $\phi(s)$ was defined as follows in Eq. 3, where s is stack state represented in evaluated sub-expression set, and Fitness is Sharpe ratio ($\frac{\mathbb{E}[r]}{\text{stdev}(r)}$, Annualized) of the evaluated return sequence. The choice of $\max$ function is inspired by common practice of factor research, that applies weak signal to enhance main signal.

$$
\phi(s) = \begin{cases} \max\limits_{e \in s} \text{Fitness}(e), & \text{if } s \neq \emptyset \\ 0, & \text{otherwise} \end{cases} \tag{3}
$$

3.4 BENCHMARK OF THE MODEL ON FINANCIAL DATASET

As noted in Sec. 2.4, in-sample portfolio fitness of multiple expression combination requires careful significance testing. In addition, it is difficult to distinguish significant alpha from noisy signal without proper alpha screening. In simple words, improvement in diversity is not just a higher temperature sampling.

Fully acknowledging the concerns on data snooping by symbolic regression, we interpret backtested return as a noisy statistics for researchers who are looking for inefficient pattern in the market. It should not be interpreted as a expected capital change. Our backtesting universe have survivorship bias by selecting high coverage stock instrument in NASDAQ and NYSE listings. Our data period is February 2015 to November 2021, which do not include significant drawdown period and hence imposes selection bias, was chosen based on our data availability.

Our backtesting return is based on close-to-close return with one day execution lag ($t + 1$ to $t + 2$). For the portfolio performance calculation, linear combination of the equation performance was used and book size constraint was not applied. Also, transaction cost and slippage was ignored. Again, we emphasize that our backtest return is an embedding of price pattern, without a financial significance.

Diversity of equation for financial dataset is evaluated using common knowledge(Tulip Indicators; CFA Institute, 2024) of empirical asset pricing. Equation for technical indicator is usually composed of the symbols listed in Eq. 4

$$\begin{aligned}&[\texttt{1}, \texttt{2}, \texttt{5}, \texttt{add(x, y)}, \texttt{subtract(x, y)}, \texttt{delta(x, t)},\\ &\texttt{Close}, \texttt{High}, \texttt{Low}, \texttt{Volume}, \texttt{Close to Close return}]\end{aligned} \tag{4}$$

On the other hand, equation for financial ratio is usually composed of the symbols listed in Eq. 5

$$[\texttt{64}, \texttt{252}, \texttt{Revenue}, \texttt{Income}, \texttt{Asset}, \texttt{Long term Debt}] \tag{5}$$

We apply alpha screening to the equation set generated by symbolic regression, by controlling False Discovery Rate (FDR) using the Benjamini-Hochberg (Benjamini & Hochberg, 1995) procedure. Target FDR rate was set to $q = 0.05$ by following previous work of Giglio et al. (2021). Resulting subset of equation after alpha screening was checked to verify that the agent is capable of generating meaningful technical indicators and financial ratios in a single policy.

To check factor robustness, portfolio return was calculated based on denoising covariance estimation (De Prado, 2020) with Mean-Variance optimization with constraint of $0 <= \text{weight} < 0.1$. skfolio (Delatte et al., 2024) package was used for portfolio optimization. Due to dataset noise, in-sample return was clipped to $(\mu - \frac{\sigma}{2}, \mu + \frac{\sigma}{2})$. Please note that our portfolio performance is not robust, and depends heavily on optimization parameter.

4 RESULTS

Our symbolic regression agent was trained in two configurations. First, the policy was optimized using episodic reward. Second, the policy was optimized using our potential based reward. Our proposed method utilizes information about fitness of all valid subexpressions of the final expression. Details about the Transformer decoder architecture and the definition of the reward is explained in Section. 3. The network was trained using Adam optimizer, with a learning rate of 1e-4, for 32000 episodes with batch size 32.

Trained agents were sampled 40000 times to generate equations. Since episodic reward baseline policy is almost deterministic, sampling in high temperature ($T = 10$) was added for comparison. Only valid and unique expressions were stored and calculated using our backtester. Please refer Section. 2.4 for our interpretation on the backtesting result. Expressions were screened to select financially meaningful factors, based on the significance of alpha with estimated FDR rate of 0.05, following methods in (Giglio et al., 2021).

Symbolic property of the result is described in Table. 1 and Fig. 3. Alpha significance of our reward shaping based method is higher than episodic reward with high temperature sampling. Most frequent pattern generated by our method is substraction of rank, while repeated rank application is most frequent in the policy trained with the episodic reward.

Fitness related property of the solution sets are shown in 4. Please note that our backtesting ignores transaction cost, delisting return, survivorship bias, dividends, selection bias, optimization parameter sensitivity, effects of risk premium. Therefore, the cumulative return should be recognized as a general robustness check in out-sample period, not a expected financial profit.

	Training algorithm	REINFORCE		REINFORCE		PPO	
	Training Reward	Episode Sharpe		Episode Sharpe		Potential based reward	
	Temperature	1		10		1	
	Screening	No	Yes	No	Yes	No	Yes
Category	Financial Ratio	4	0	28	6	114	26
	Technical indicators	0	0	96	14	380	148
	Mixed	2	0	93	10	418	94
	Total	6	0	217	30	912	268

Table 1: Number of unique equations generated using symbolic regression. Baseline RL episodic reward policy is almost deterministic, and sampling in high temperature shows relatively low alpha screening selection rate. Each policies were sampled 40000 times. Same or invalid equations are not counted. Estimated FDR for the alpha screening is 0.05.

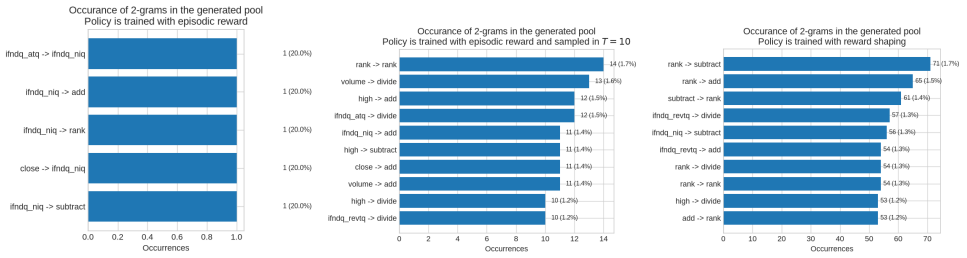

Figure 3: 2-gram frequency of the samples equations. Left: episodic reward agent is almost deterministic, Middle: episodic reward agent sampled in high temperature ($T = 10$), Right: Our reward shaping based agent.

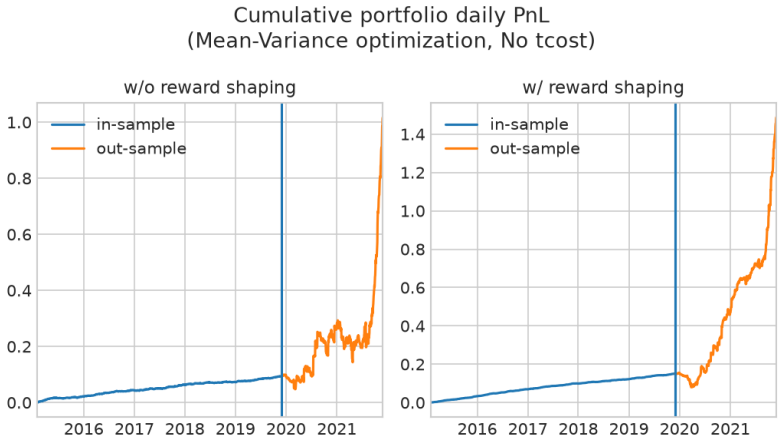

Figure 4: Cumulative portfolio return of expression pool. Annualized Sharpe ratio of expressions generated from policy trained without reward shaping is 3.13 in-sample and 2.40 out-sample. Result with reward shaping shows 4.02 in-sample and 7.81 out-sample. Please note that purpose of this backtest is comparison of factor robustness, not performance comparison. Also, executability is not checked, because transaction cost is being ignored.

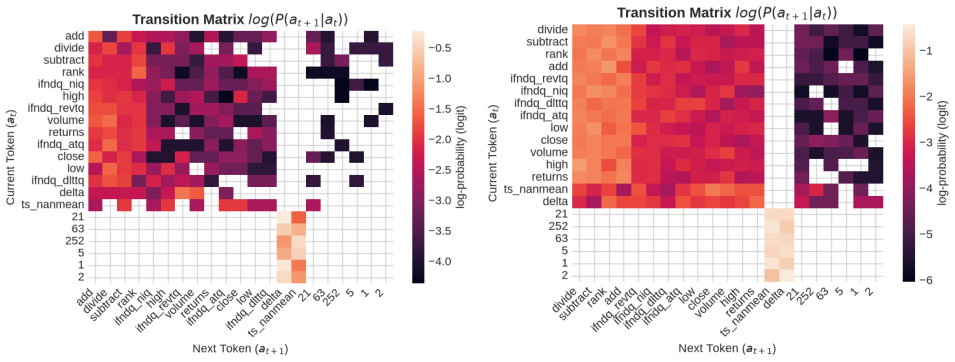

Figure 5: Sampled log-probability of token transition. Left : Policy trained on episodic reward (in high temperature $T = 10$), Right : Policy trained on our potential based reward shaping. ($T = 1$)

5 Discussion and Conclusion

In this paper, the potential based reward shaping with small length penalty with Proximal Policy Optimiaztion(PPO) was proposed for RL-based symbolic regression. Simply speaking, we try to find fragmented symbol sequence that improves potential fitness, rather than directly optimizing for the complete equation fitness which is difficult due to credit assignment problem.

More diverse sampling of equations is observed in Table. 1. Our reward shaping method produces both financial ratios and technical indicators, while the episodic method produced only financial ratios. Our reward shaping method is distinguished from high temperature sampling by relatively higher alpha significance, which was measured using FDR-controlled significance testing.

Acknowledging limited financial realism of our backtester, portfolio return shows similarly positive for both episodic reward case and our potential based reward shaping method. One difference is the number of expression found. Episodic reward based method only 30 equations due to low selection rate in FDR-controlled significance test. Our reward shaping generated 268 equations with relatively high alpha significance.

To summarize, reward shaping with for RL based symbolic regression is possible with our novel, virtual stack based environment. More diverse behavior of a policy (Fig. 5) is observed, which implies successful variance control of expected return and less credit assignment problem. As a result, equation diversity was improved under same sampling trial condition.

6 AI usage disclosure

In this work, we used generative AI tools for edit software codes(transformer, virtual stack machine, FDR controlled significance testing), looking for related works(Policy diversity in RL (Sec. 2.2)), assist with translation. We have not used generative AI tools for data or developing theoretical models in this work. Additionally, we used generative AI tools for writing drafts and understanding existing works. We checked AI-generated codes by inspection by an author. Draft generated by the LLM was not directly used in the paper, and fully rewritten by human author. The author who inspected take responsibility for the AI-generated codes and final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Christopher M Bishop et al. *Neural networks for pattern recognition*, volume 482. Clarendon press Oxford, 1995.

Tim Blackwell, Jürgen Branke, and Xiaodong Li. Particle swarms for dynamic optimization problems. In *Swarm Intelligence: Introduction and Applications*, pp. 193–217. Springer, 2008.

TM Blackwell. Particle swarms and population diversity. *Soft Computing*, 9(11):793–802, 2005.

Edmund K Burke, Steven Gustafson, and Graham Kendall. Diversity in genetic programming: An analysis of measures and correlation with fitness. *IEEE Transactions on Evolutionary Computation*, 8(1):47–62, 2004.

CFA Institute. Cfa financial ratio list, 2024. URL https://www.cfainstitute.org/sites/default/files/-/media/documents/support/programs/cfa/cfa_program_level_ii_financial_ratio_list.pdf.

Marcos M López De Prado. *Machine learning for asset managers*. Cambridge University Press, 2020.

Hugo Delatte, Carlo Nicolini, and Matteo Manzi. skfolio, 2024. URL https://doi.org/10.5281/zenodo.16148630.

Winston Wei Dou, Itay Goldstein, Zigang Li, and Liyan Yang. Human edge, machine limits: Ai-human competition in financial markets. *The Wharton School Research Paper*, 2025.

Guanhao Feng, Stefano Giglio, and Dacheng Xiu. Taming the factor zoo: A test of new factors. *The Journal of Finance*, 75(3):1327–1370, 2020.

Robert Geirhos, Roland S Zimmermann, Blair Bilodeau, Wieland Brendel, and Been Kim. Don't trust your eyes: on the (un) reliability of feature visualizations. *arXiv preprint arXiv:2306.04719*, 2023.

Stefano Giglio, Yuan Liao, and Dacheng Xiu. Thousands of alpha tests. *The Review of Financial Studies*, 34(7):3456–3496, 2021.

Campbell R Harvey, Yan Liu, and Heqing Zhu. ... and the cross-section of expected returns. *The Review of financial studies*, 29(1):5–68, 2016.

John Hewitt, Robert Geirhos, and Been Kim. We can't understand ai using our existing vocabulary. *arXiv preprint arXiv:2502.07586*, 2025.

Zura Kakushadze. 101 formulaic alphas. *Wilmott*, 2016(84):72–81, 2016.

Pierre-Alexandre Kamienny, Stéphane d'Ascoli, Guillaume Lample, and François Charton. End-to-end symbolic regression with transformers. *Advances in Neural Information Processing Systems*, 35:10269–10281, 2022.

Timo Klein. Autonomous algorithmic collusion: Q-learning under sequential pricing. *The RAND Journal of Economics*, 52(3):538–558, 2021.

John R Koza. Genetic programming as a means for programming computers by natural selection. *Statistics and computing*, 4(2):87–112, 1994.

William La Cava, Bogdan Burlacu, Marco Virgolin, Michael Kommenda, Patryk Orzechowski, Fabrício Olivetti de França, Ying Jin, and Jason H Moore. Contemporary symbolic regression methods and their relative performance. *Advances in neural information processing systems*, 2021(DB1):1, 2021.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pp. 278–287. Citeseer, 1999.

Jack Parker-Holder, Aldo Pacchiano, Krzysztof M Choromanski, and Stephen J Roberts. Effective diversity in population based reinforcement learning. *Advances in Neural Information Processing Systems*, 33:18050–18062, 2020.

Julien Pénasse. Understanding alpha decay. *Management Science*, 68(5):3966–3973, 2022.

Brenden K. Petersen, Mikel Landajuela, T. Nathan Mundhenk, Cláudio Prata Santiago, Sookyung Kim, and Joanne Taery Kim. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

Jonathan Schaeffer. A gamut of games. *AI Magazine*, 22(3):29–29, 2001.

Jonathan Schaeffer, Robert Lake, Paul Lu, and Martin Bryant. Chinook: The world man-machine checkers champion. *AI magazine*, 17(1):21–29, 1996.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

Herbert A Simon. From substantive to procedural rationality. In *25 years of economic theory: Retrospect and prospect*, pp. 65–86. Springer, 1976.

Richard S Sutton, Andrew G Barto, and Andrew Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Igor Tulchinsky. *Finding Alphas: A quantitative approach to building trading strategies*. John Wiley & Sons, 2019.

Tulip Indicators. Tulip indicators - open-source technical analysis library. `https://tulipindicators.org/`.

Halbert White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000.

Shuo Yu, Hongyan Xue, Xiang Ao, Feiyang Pan, Jia He, Dandan Tu, and Qing He. Generating synergistic formulaic alpha collections via reinforcement learning. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 5476–5486, 2023.

Yi Zhou and Jianwu Lin. The alpha life cycle of quantitative strategy. In *2017 IEEE/SICE International Symposium on System Integration (SII)*, pp. 53–59. IEEE, 2017.