AlphaX: A Tri-Agent Framework for Microstructure-Constrained Symbolic Alpha Discovery in Crypto Futures

Jiaqi Wang[1] and Shuo Yin[2(✉)]

[1] School of Finance Engineering, Southeast University, Nanjing 211100, China
[2] Tsinghua University, Beijing 100190, China
213230851@seu.edu.cn
yins25@mails.tsinghua.edu.cn

Abstract. We study symbolic alpha discovery for cryptocurrency perpetual-futures markets, where standard equity-oriented factor pipelines transfer only partially because the data include derivatives-native fields and stronger microstructure effects. AlphaX uses a two-phase process. In Phase I, three agents generate, repair, and evaluate symbolic expressions under a seven-gate validity cascade, removing 65.7% of degenerate candidates while updating later search from report-based observations and cross-round feedback. In Phase II, the retained Phase I top-50 library is orthogonalized, reducing mean absolute pairwise correlation to 0.26. On the 2024 evaluation split, the retained library reaches mean factor IC 0.145. On the 2022–2024 full panel, HRP allocation on the retained library reaches Sharpe 1.739 with 56.0% annualized return net of costs, ahead of the reproduced baselines under matched execution settings.

Keywords: Large Language Models, Multi-Agent Systems, Cryptocurrency Derivatives, Symbolic Alpha Discovery, Redundancy Control.

1 Introduction

Symbolic alpha discovery has moved through several stages, from genetic programming [1, 2] and deep predictors [3, 4, 5] to recent LLM-based synthesis through natural-language prompts [6, 7, 8, 9]. Much of this work, however, is still evaluated with assumptions inherited from equity cross-sectional factor research. These assumptions do not transfer cleanly to cryptocurrency derivatives. In perpetual-futures markets, price discovery is fragmented [10, 11], and useful signals often depend on funding-rate dynamics, position ratios, and persistent temporal structure [12, 13, 14], which are outside standard equity factor libraries.

Existing pipelines fail along three dimensions. First, sparse derivatives-native fields and noisy high-frequency inputs can exhaust evaluation budgets when candidate symbolic expressions are sent directly to backtesting without deterministic pre-screening. Second, stateless generation cannot shift toward productive regions and may repeatedly revisit semantically homogeneous candidates [6, 15, 16]. Third, unstructured factor

assembly yields residual collinearity that can translate into concentration and co-drawdowns under leverage.

AlphaX addresses these failures with a two-phase discovery pipeline for cryptocurrency perpetual futures. Phase I uses three agents to separate idea generation, symbolic implementation, and evaluation. This design avoids manual seed formulas, repairs candidates at the failed gate when possible, and updates later sampling through trajectory feedback. Phase II evaluates the retained library after redundancy reduction. Our contributions are:

1. **Validity-constrained tri-agent discovery.** In Phase I, AlphaX combines role-separated tri-agent generation, seven deterministic pre-backtest quality gates ($O(n)$ complexity), and a failure-localized repair protocol, increasing the pass rate by 2.3 × relative to unconditional regeneration.
2. **Feedback-driven search from market reports.** In Phase I, AlphaX replaces stateless prompting with a search process that begins from report-derived observations, preserves breadth through a diversity mechanism, and updates direction scores from cumulative IC feedback.
3. **Fixed-library evaluation after redundancy reduction.** In Phase II, AlphaX evaluates the retained factor library after orthogonalization under aligned execution settings, so redundancy is handled before downstream routing rather than only after factor generation.

2 Related Work

Cryptocurrency Microstructure and Signal Discovery. Symbolic discovery in quantitative finance has traditionally relied on genetic programming and heuristic expression search over handcrafted operator sets [1, 2]. More recent statistical learning architectures often improve predictive performance, but they also reduce transparency and direct executability relative to symbolic factor expressions [3, 4, 5]. In cryptocurrency markets, the difference is not only the asset class but also the market structure. Prior work reports fragmented price discovery, strong order-flow effects, and pronounced temporal dependence in crypto trading [10, 11, 13, 14]. Other recent evidence finds interpretable regularities in cryptocurrency microstructure data, including patterns related to order-book dynamics and trading-flow imbalances [17]. While benchmark infrastructures such as Qlib help standardize equity-factor evaluation workflows [18], they are not designed for derivatives-native signals such as funding rates, open interest, and long-short positioning ratios.

LLMs, Agentic Finance, and Financial Decision Support. Recent large language models have broadened the scope of automated research and program synthesis. Advances in instruction following and reasoning [8, 9, 19, 20] support executable code generation, intermediate reasoning traces, and interaction with external tools. In finance, domain-adapted models such as BloombergGPT and FinBERT show the growing use of language models in financial text understanding and decision support [21, 22]. Related trading studies have also explored the combination of LLM guidance and reinforcement learning in noisy markets. One line of work uses agentic reinforcement-

learning pipelines for stock trading [23], whereas another studies goal-aware portfolio exploration or reinforcement-learning decision rules under noisy market conditions [24, 25]. These papers are relevant to trading support, but they do not directly study the task considered here, namely the generation of symbolic alpha expressions that are both valid and backtestable.

Agentic Factor Mining and Symbolic Alpha Generation. Multi-agent coordination is increasingly used in LLM-based research workflows because role separation, reflection, and iterative repair can improve robustness in open-ended coding tasks [26, 27]. In quantitative research, related ideas can be seen across several systems. AlphaGPT and RD-Agent are early examples, while QuantAgent, FactorMiner, and AlphaSAGE extend this line in different directions; more recent work also includes neural-symbolic and evolutionary mining frameworks [6, 7, 15, 16, 28, 29, 30]. Taken together, these studies suggest that LLMs are now being used not only for narrative assistance but also for hypothesis generation, symbolic mining, and search-space exploration. At the same time, most existing frameworks still place more emphasis on search breadth or iterative exploration than on disciplined admission control before backtesting.

Portfolio Robustness and Redundancy Control. Predictive factor discovery is closely connected to portfolio construction when candidate signals are highly collinear. Classical allocation methods such as Black–Litterman and hierarchical risk parity explicitly account for covariance structure and diversification during portfolio construction [31, 32]. This line of work suggests that factor discovery and library construction should not be treated as fully separable problems when redundancy can propagate into concentration and co-drawdowns. In many LLM-based mining pipelines, however, redundancy control is still limited or deferred until after factor generation, even though correlated signals can materially reduce effective breadth.

Comparative Framing. Existing LLM-driven and multi-agent pipelines have broadened the scope of symbolic discovery, but several issues remain for crypto-derivatives mining. Many frameworks still search broadly without lightweight pre-backtest checks tailored to derivatives-native expressions. Their feedback loops are often local to the current attempt and do not clearly redirect later rounds. Redundancy is also often handled only after factor generation. These points motivate the present design, which combines pre-backtest validity checks, cross-round feedback, and redundancy reduction before downstream evaluation.

3 Methodology

3.1 System Overview and Objective

Formally, let $\mathcal{U} = \{s_1, \ldots, s_{12}\}$ denote the 12-instrument USDT perpetual-futures universe. For any instrument $s \in \mathcal{U}$ at time t, the prediction target is the cross-sectionally normalized log-return $y_t^{(s)}$ over $[t, t + 24\text{h}]$. The objective is to extract an optimal factor library $\mathcal{L}^* \subseteq \mathcal{F}$ that maximizes risk-adjusted predictive power while minimizing internal redundancy:

$$\mathcal{L}^* = \underset{\mathcal{L}\subseteq\mathcal{F}}{\operatorname{argmax}}\left[\frac{1}{|\mathcal{L}|}\sum_{f\in\mathcal{L}} \mathrm{ICIR}(f) - \lambda\cdot\bar{\rho}(\mathcal{L})\right] \tag{1}$$

This maximization is constrained by admission thresholds covering standalone information coefficient ($\mathrm{IC}(f) \geq 0.01$), single-factor strategy Sharpe ratio ($\mathrm{SR}(f) \geq 0.1$), maximum drawdown ($|\mathrm{MDD}(f)| \leq 0.6$), and syntactic tree depth ($n_{\mathrm{nodes}}(f) \leq 120$). The library-level collinearity penalty $\bar{\rho}(\mathcal{L})$ is defined as the average absolute Pearson correlation across all factor pairs:

$$\bar{\rho}(\mathcal{L}) = \frac{2}{|\mathcal{L}|(|\mathcal{L}|-1)}\sum_{i<j} |\rho_{ij}| \tag{2}$$

and $\lambda = 0.3$, selected on the 2023 validation split. Except for explicitly reported validation-selected values, remaining control thresholds are fixed before 2024 evaluation. The full phase definitions and data-usage protocol are deferred to Sect. 4.

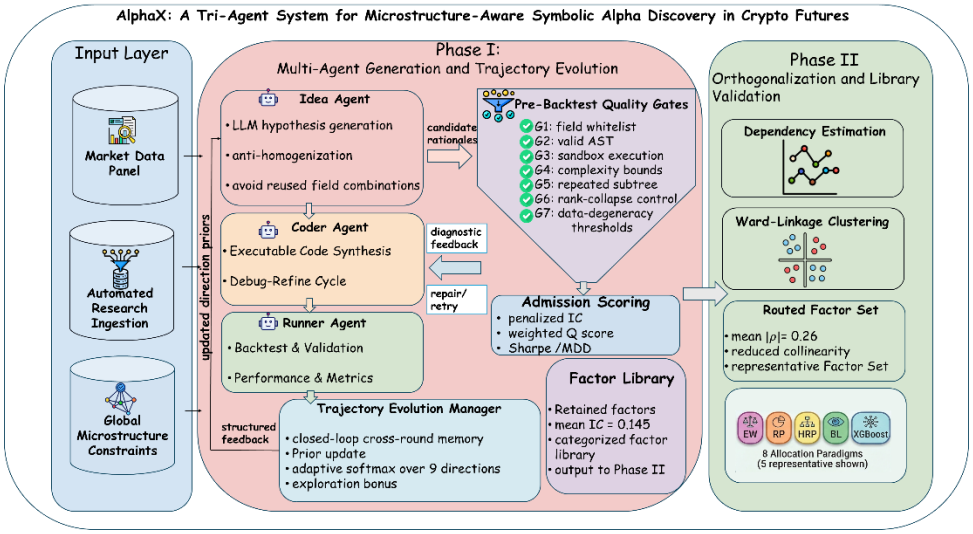

Fig. 1. Two-phase AlphaX pipeline: Phase I removes invalid candidates before backtesting, and Phase II evaluates fixed-library usability through orthogonalization and routing.

3.2 Phase I: Multi-Agent Generation and Trajectory Evolution

Phase I proceeds round by round. In each round, the system starts from one market rationale, turns it into a symbolic expression, checks it with the gate cascade, and sends it to evaluation only after the checks are passed. Information from earlier rounds is then used to adjust which research directions are sampled later.

The search begins with the Idea Agent, which reads report-derived market observations and develops hypotheses in nine crypto-specific directions. The use of these directions is motivated by the structure of perpetual-futures data, where predictive information is often tied to leverage disagreement, derivatives sentiment, and microstructure-driven momentum or volatility effects rather than to standard equity-style factor

families. The first group therefore centers on leverage and positioning, covering smart-money divergence, whale positioning momentum, funding-rate signals, funding-rate extremes, and open-interest dynamics. The second group focuses on trading behavior and cross-market propagation, covering trade microstructure, volatility regimes, cross-asset momentum, and composite sentiment. These directions act as search priors rather than fixed templates, so the Idea Agent can still propose new expressions within each direction while staying anchored to fields that are specific to perpetual-futures data. The Coder Agent then writes the corresponding symbolic expression and checks it against the seven-gate cascade. Expressions that pass are backtested by the Runner Agent, whose feedback is reused in later attempts.

Quality control and admission. Every candidate symbolic expression f must satisfy

$$\mathrm{PASS}(f) = \prod_{g=1}^{7} \mathbf{1}\big[G_g(f)\big] \tag{3}$$

where the seven gates are applied in a fixed order before full backtesting, so that invalid or degenerate expressions can be removed before expensive evaluation. The early checks establish whether the candidate is usable as an executable symbolic expression at all. Accordingly, G_1 enforces the 15-field whitelist in Table 1, G_2 requires a valid AST parse, G_3 requires sandbox execution without runtime errors, G_4 bounds expression complexity by $5 \le n_{\mathrm{nodes}}(f) \le 120$ and $\mathrm{depth}(f) \le 12$, and G_5 rejects repeated internal subtrees with at least three AST nodes. Passing these checks, however, does not yet guarantee that the resulting series is informative on crypto data. For that reason, G_6 is used to prevent rank collapse by disallowing bare RANK() and requiring TS_RANK_PERCENTILE windows to be at least 100; bare ranks are repaired to TS_RANK_PERCENTILE$(\cdot, 168)$ before rechecking. G_7 then screens for data degeneracy by requiring NaN ratio $\le 25\%$, zero ratio $\le 50\%$, and unique-value ratio ≥ 0.001. Surviving expressions receive the composite score

$$\mathcal{Q}(f) = \sum_{m \in \mathcal{M}} w_m q_m(f) \tag{4}$$

where $\mathcal{M} = \{\mathrm{IC}, \mathrm{RankIC}, \mathrm{ICIR}, \mathrm{RankICIR}, \mathrm{bonus}, \mathrm{AnnRet}\}$ and the corresponding weights are $(0.35, 0.15, 0.25, 0.15, 0.05, 0.05)$. Each component is mapped to 0–100; admission requires $\mathcal{Q}(f) \ge 80$ within at most five repair attempts. If a candidate fails a gate, the failed gate and the error message are passed back to the Coder Agent. The next attempt repairs the local failure instead of regenerating the whole expression from scratch.

Trajectory evolution and cross-round adaptation. The trajectory module keeps a score for each research direction and updates it after admitted expressions are evaluated. The delayed reward is

$$\mathrm{IC}_{\mathrm{pen}}(f) = \mathrm{IC}(f) - \lambda_c \log\big(1 + n_{\mathrm{nodes}}(f)\big) \tag{5}$$

where $\lambda_c = 0.003$ and $n_{\mathrm{nodes}}(f)$ is the AST node count. The penalty discourages long expressions when two candidates have similar IC. Under $p_d \propto \exp\left(\frac{s_d}{T}\right)$ with $T = 5$, the expectation score evolves as

$$s_d \leftarrow s_d + \eta \cdot \mathrm{IC}_{\mathrm{pen}}(f) \tag{6}$$

with learning rate $\eta = 1.0$. An exploration bonus $\epsilon(n_p) = \frac{0.1}{1+0.1n_p}$ decays with the cumulative attempt count n_p, while a five-round rolling admission rate below 0.30 raises the generative temperature to $\tau = 1.2$ until efficiency recovers. Thus, failed attempts affect only the local repair step, whereas direction scores are updated only after admitted expressions are evaluated.

Algorithm 1: AlphaX Phase I: Multi-Agent Validated Generation

Input: Research Directions $\mathcal{D}$, Historical MAB Expectations s_d, Horizon T
Output: Admitted Expression Library $\mathcal{L}$

$\mathcal{L} \leftarrow \emptyset$;
for *round* $t = 1$ **to** T **do**
  Sample direction $d \sim \mathrm{Softmax}(s_d/\tau)$;
  $h \leftarrow \mathrm{IdeaAgent}(d)$;
  for *attempt* $i = 1$ **to** 5 **do**
    $f \leftarrow \mathrm{CoderAgent}(h, \mathrm{feedback})$;
    if $\mathrm{PASS}(f) = 1$ *and* $\mathcal{Q}(f) \geq 80$ **then**
      $\mathcal{L} \leftarrow \mathcal{L} \cup \{f\}$;
      $s_d \leftarrow s_d + \eta \cdot \mathrm{IC}_{\mathrm{pen}}(f)$; // trajectory update
      break;
    else
      feedback $\leftarrow \mathrm{RunnerAgent}(f)$;
    end
  end
  if *rolling_admission* < 0.30 **then**
    $\tau \leftarrow 1.2$; // Inject semantic variance
  end
end
return $\mathcal{L}$;

3.3 Phase II: Orthogonalization and Fixed-Library Evaluation

Phase II operates on the retained Phase I top-50 library and evaluates whether that library remains usable after redundancy reduction under heterogeneous downstream rules. To reduce residual collinearity before downstream evaluation, we apply Ward-linkage clustering using the absolute-correlation distance $d_{ij} = \sqrt{0.5(1 - |\rho_{ij}|)}$. The resulting dendrogram is thresholded to retain the highest-ICIR representative from each semantic cluster.

The reduced library then enters a shared strategy-routing layer evaluated under matched execution settings. To test the retained library across different downstream decision rules, we consider a zero-parameter baseline (EW), classical risk-aware

allocators (MVO-IVS-RP, HRP, Risk Parity, and Black–Litterman), a nonlinear selector based on XGBoost [33], and lightweight adaptive or LLM-assisted selectors.

4 Experimental Setup

4.1 Data and Partitioning

The 2022–2024 hourly panel contains 292,456 instrument-hour observations across 12 Binance USDT perpetual-futures instruments. Instruments are retained based on sample coverage and minimum daily notional volume thresholds. Timestamp-level observations with missing close or volume are discarded without forward-filling (gap rate $<$ 0.03%).

Table 1 summarizes the 15-variable feature space across three domains; the six derivatives-native fields have no counterpart in standard equity libraries such as Alpha158 [18].

Table 1. Feature space by domain.

Domain	Variables (15 total)
Price-Volume	close, high, low, volume, trade_count
Order-Flow Microstructure	taker_buy_vol, taker_sell_vol, trade_size_avg, realized_vol_5m
Derivatives-Native	open_interest, funding_rate, top_ls_account_ratio, global_ls_account_ratio, top_ls_position_ratio, taker_ls_vol_ratio

Two evaluation stages are considered. Phase I uses a chronological 2022/2023/2024 split for factor generation, validation, and single-factor evaluation. Phase II treats the retained top-50 factors as a fixed library and evaluates orthogonalization, allocation, and baseline comparison on the 2022–2024 full panel. Thus, Phase II portfolio metrics represent post-selection fixed-library evaluation, not a fully untouched forward out-of-sample test. These three years cover distinct market regimes, including the 2022 downturn and the subsequent 2023–2024 recovery.

For Phase II comparisons, all baselines are re-implemented under the same data, cost, and execution-budget assumptions. They share the same universe, target horizon, cost model, and rebalancing rule. Genetic Programming uses unguided symbolic search. AlphaGPT uses stateless LLM generation. RD-Agent performs multi-agent discovery without the Phase II orthogonality step. FactorMiner performs symbolic mining without explicit redundancy control. DeepCrypto [34] serves as a neural predictive benchmark.

4.2 Backtest Protocol and Evaluation Metrics

All-in round-trip transaction cost is modeled as 3 bps per side plus an additional 2 bps slippage term per rebalance, i.e., 8 bps in total, with no explicit market-impact model. This serves as a base-case cost assumption for ordinary VIP-tier trading in liquid Binance USDT perpetuals, including exchange fees and moderate slippage, rather than a venue- or capacity-invariant cost level. Short selling is permitted via perpetual-futures mechanics, and funding cash flow is settled on the exchange's 8-hour schedule. Factor values and portfolio weights are refreshed hourly against a 24-hour forward return target using only information available at each rebalance timestamp, so both single-factor and post-routing evaluations follow the same causal execution rule. Initial capital is 100,000 USDT, with a 30% NAV single-instrument cap, an 18% single-factor cap under HRP, and a 0.1% NAV minimum position.

Phase I reports single-factor library metrics on the 2024 evaluation split, including IC, Rank IC, ICIR, Rank ICIR, and the half-life statistic derived from the IC-decay profile. Phase II reports post-routing fixed-library metrics for each downstream strategy under aligned execution settings, covering both predictive metrics (IC, ICIR) and portfolio metrics (Sharpe ratio, annualized return, maximum drawdown, and Calmar ratio). The metric semantics are therefore phase-aligned. In Phase I, predictive metrics are computed from standalone retained library factors on the 2024 evaluation split, whereas in Phase II they are computed from the strategy-level composite signal produced after routing on the retained Phase I top-50 library. For mean-IC significance, statistical inference uses Newey–West HAC standard errors with 24 lags.

Signal persistence is characterized by the half-life $T_{1/2} = \ln 2/\gamma$ derived from the exponential fit

$$\mathrm{IC}(h) = \mathrm{IC}(0) \cdot e^{-\gamma h} \tag{7}$$

estimated over horizons $h \geq 24$h to exclude transitory noise from 8-hour funding settlements.

5 Results and Analysis

5.1 Phase I Library Characteristics

The retained Phase I top-50 library spans seven factor categories (Table 2). On the 2024 evaluation split, it achieves mean IC 0.145 and ICIR 0.376, and the mean IC is above zero under a Newey–West HAC test with 24 lags ($t = 12.8, p < 0.001$). The test uses the retained-library IC time series over the 2024 evaluation split. As a discovery-stage output, the library is not concentrated in a single semantic family. The reported half-life statistics stay positive, and all top-50 library factors use at least one derivatives-native field absent from standard equity libraries.

Table 2. Phase I top-50 library by category on the 2024 evaluation split.

Category	num	IC_{avg}	ICIR_{avg}	$\text{Rank IC}_{\text{avg}}$	$\text{Rank ICIR}_{\text{avg}}$	$T_{\frac{1}{2}}$ (h)
Microstructure	3	**0.259**	**0.596**	**0.252**	**0.449**	56.0
TakerFlow	9	0.160	0.410	0.154	0.366	41.0
Momentum	12	0.130	0.349	0.114	0.255	50.0
Funding / OI Dynamics	14	0.128	0.334	0.105	0.215	44.6
SmartMoney	9	0.123	0.335	0.108	0.297	**63.0**
Volatility	2	0.190	0.460	0.173	0.435	48.0
Sentiment	1	0.202	0.545	0.183	0.545	48.0
Overall	**50**	**0.145**	**0.376**	**0.129**	**0.296**	**49.2**

$$\text{IC}_{\text{max}} = 0.352,\ \text{ICIR}_{\text{max}} = 0.732,\ \text{Rank IC}_{\text{max}} = 0.355,\ Q_{\text{max}} = 98.0$$

Microstructure factors show the largest mean IC and ICIR, while SmartMoney factors have the longest half-life. The discovery loop therefore does not collapse onto a single semantic family. Ward-linkage clustering lowers mean $|\rho|$ from 0.48 to 0.26. The decay profiles in Fig. 2 stay positive through 72h, and half-life estimates use horizons with $h \geq 24$h because the 6h IC is likely affected by the 8h funding cycle.

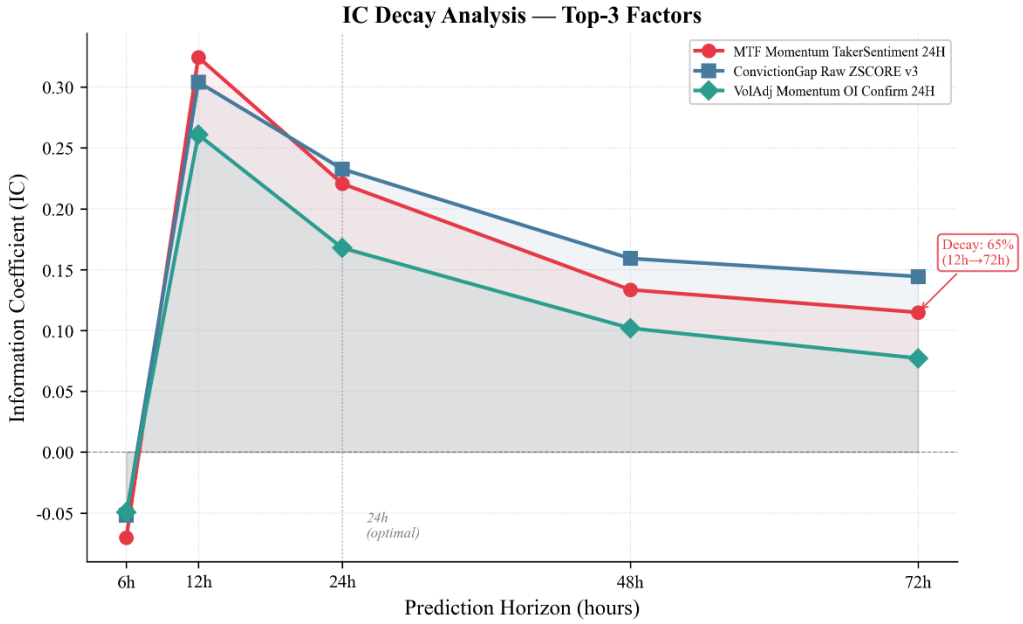

Fig. 2. IC decay profiles for three representative factors at horizons 6h--72h.

5.2 Phase II Fixed-Library Evaluation Across Routing Rules

Table 3 compares routing rules on the retained Phase I top-50 library. Across this set, HRP delivers the strongest risk-adjusted outcome, while XGBoost attains the best post-routing predictive accuracy and the shallowest drawdown.

Table 3. Phase II fixed-library evaluation on the retained Phase I top-50 library.

Method	IC	ICIR	Sharpe	Ann.Ret	MaxDD	Calmar
EW	0.304	0.669	0.653	21.3%	-34.0%	0.627
MVO-IVS-RP	0.372	0.736	1.481	48.1%	-31.7%	1.518
Risk Parity	0.370	0.734	1.220	39.4%	-32.9%	1.200
Black-Litterman	0.374	0.735	0.870	29.0%	-37.0%	0.784
HRP (Ward)	0.365	0.733	**1.739**	**56.0%**	-30.5%	**1.840**
XGBoost	**0.378**	**0.741**	1.333	43.8%	**-26.1%**	1.676
Adaptive-RP	0.369	0.734	1.023	33.8%	-34.9%	0.968
LLM Selector	0.361	0.721	0.958	31.1%	-35.6%	0.874

Table 4 reports the system-level comparison. The upper block contains five reproduced baseline mining pipelines, and the lower block contains representative Phase II fixed-library results on the retained Phase I top-50 library. Because every method is evaluated through its downstream decision route, the table reports post-routing metrics

rather than the Phase I single-factor metrics in Table 2. EW serves as a zero-parameter baseline, Ward-HRP captures correlation-aware risk diversification, and XGBoost-based gain selection tests robustness under a nonlinear selector. For reproduced baselines, the allocator follows the original method or its standard implementation whenever possible, under aligned data, cost, and execution settings.

Table 4. System-level comparison under reproduced system settings: baseline pipelines and AlphaX end-to-end outcomes.

Method	Allocation	IC	ICIR	Rank IC	Sharpe	Ann.Ret	MaxDD	Calmar
Comparison Methods								
Genetic Prog.	EW	0.223	0.382	0.205	0.450	14.2%	-34.1%	0.420
AlphaGPT	IC-Weight	0.261	0.451	0.240	0.610	18.3%	-31.3%	0.590
DeepCrypto	DeepCrypto	0.289	0.503	0.266	0.740	22.1%	-30.3%	0.730
RD-Agent	LightGBM	0.312	0.571	0.287	0.880	27.1%	-31.2%	0.870
FactorMiner	IC-Weight	0.341	0.640	0.314	1.050	32.3%	-30.1%	1.070
AlphaX								
AlphaX (EW)	Equal-Weight	0.304	0.669	0.274	0.653	21.3%	-34.0%	0.630
AlphaX (HRP)	Ward-HRP	0.365	0.733	0.339	**1.739**	**56.0%**	-30.5%	**1.840**
AlphaX (XGBoost)	Gain-Selection	**0.378**	**0.741**	**0.352**	1.333	43.8%	**-26.1%**	1.676

Among the reproduced baselines, FactorMiner attains the highest IC and ICIR, at 0.341 and 0.640, respectively. AlphaX with HRP achieves the highest Sharpe and annualized return, while AlphaX with XGBoost gives the highest IC and the shallowest drawdown. Fig. 3 shows the same ordering over the 2022–2024 full panel. The right panel reports relative excess return against an equal-weight buy-and-hold benchmark over the trading universe.

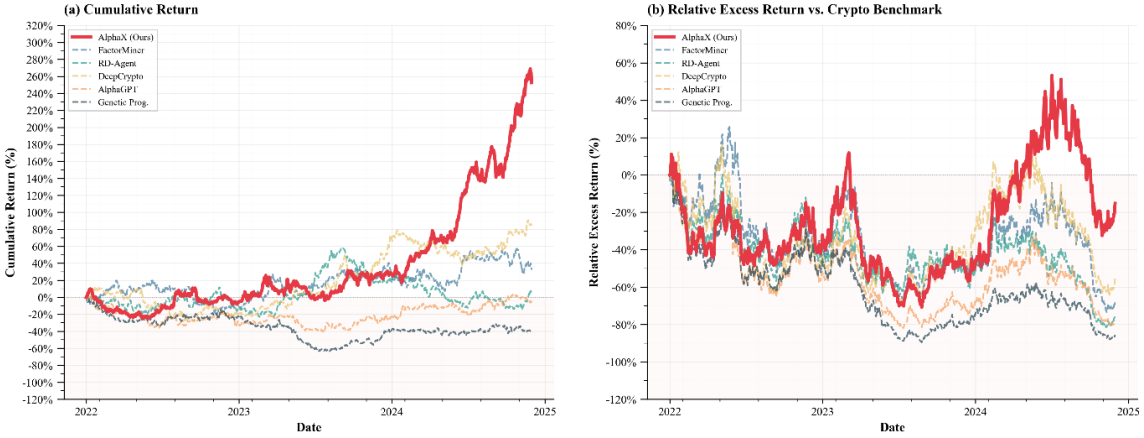

Fig. 3. Phase II cumulative-return comparison and benchmark-relative excess returns under aligned execution settings over the 2022--2024 full panel, comparing AlphaX with Genetic Programming, AlphaGPT, DeepCrypto, RD-Agent, and FactorMiner.

5.3 Ablation Analysis

A sequential ablation in Table 5 quantifies the contribution of the AlphaX-specific data-aware filters (G_4–G_7) and higher-level architectural modules, while keeping the basic executable-hygiene checks (G_1–G_3) fixed. All ablation configurations follow the same data split, cost model, and execution assumptions as the full-system reference.

Table 5. Sequential ablations. IC uses the Phase II XGBoost configuration; Sharpe uses the Phase II HRP configuration as the full-system reference.

Configuration	IC	Sharpe	ΔIC	ΔSharpe	Mean $\|\rho\|$
Full AlphaX (Phase I & II)	**0.378**	**1.739**	—	—	**0.26**
w/o Independent Coder Agent	0.271	1.152	-28.4%	-33.8%	0.28
w/o Data-Aware Gates (G_4–G_7)	0.287	1.198	-23.9%	-31.1%	0.31
w/o Trajectory Evolution	0.355	1.598	-6.0%	-8.1%	0.29
w/o Redundancy Control(Phase I+II)	0.367	0.870	-3.0%	-50.0%	>0.50
Single LLM Agent Baseline	0.237	0.987	-37.3%	-43.2%	>0.50

The largest degradation comes from removing the independent Coder Agent, lowering IC by 28.4% and Sharpe from 1.739 to 1.152. A similarly large decline appears when the data-aware gate set (G_4–G_7) is disabled. This drop is mainly associated with weaker complexity control, less effective rank-collapse prevention, and the loss of data-degeneracy filtering. The pattern is different for Phase II redundancy control. IC falls by only 3.0%, whereas Sharpe is cut by roughly half. The result places its contribution mainly at the fixed-library stage rather than in standalone predictive strength. Trajectory evolution has a smaller effect overall, and the single-LLM baseline remains weaker than the full system.

6 Conclusion

This paper presents AlphaX for symbolic factor discovery in cryptocurrency perpetual futures. The framework uses market-report observations for factor proposal, applies fixed admission gates to the generated expressions, and reduces redundancy before portfolio evaluation. The symbolic form is also advantageous in this setting because the resulting signals remain directly executable and open to analyst inspection.

Several limitations of the current evaluation should be noted. We use a fixed transaction-cost model and do not model market impact beyond the stated slippage term. The tested universe contains liquid perpetual contracts, so results on thinner instruments may differ. The sample also stops at the end of 2024; post-2024 forward testing and rolling-window evaluation are needed to check stability in later regimes.Finally, the comparison focuses on reproduced discovery and predictive baselines, and a broader test against hand-crafted and published cryptocurrency alpha factors remains future work.

References

1. Koza, J.R.: Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, MA (1992).
2. Peters, M., et al.: Building Market Timing Signals with Genetic Programming. The Journal of Portfolio Management **27**(2), 56–65 (2001).
3. Krauss, C., Do, X.A., Huck, N.: Deep Neural Networks, Gradient-Boosted Trees, Random Forests: Statistical Arbitrage on the S&P 500. European Journal of Operational Research **259**(2), 689–702 (2017).
4. Gu, S., Kelly, B., Xiu, D.: Empirical Asset Pricing via Machine Learning. The Review of Financial Studies **33**(5), 2223–2273 (2020).
5. Feng, G., Polson, N.G., Xu, J.: Deep Learning in Characteristics-Sorted Factor Models. Journal of Financial and Quantitative Analysis **59**(7), 3058–3084 (2024).
6. Yu, X., Zhang, H., Pan, Y., et al.: AlphaGPT: Human-AI Interactive Alpha Mining for Quantitative Investment. arXiv:2308.00016 (2023).
7. Chen, X., Li, Y., Gao, Y., et al.: RD-Agent: Automating LLM-based Research and Development. In: NeurIPS 2024 Workshop on Foundation Models for Decision Making (2024).
8. Brown, T., Mann, B., Ryder, N., et al.: Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems **33**, 1877–1901 (2020).
9. Ouyang, L., Wu, J., Jiang, X., et al.: Training Language Models to Follow Instructions with Human Feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022).
10. Makarov, I., Schoar, A.: Trading and Arbitrage in Cryptocurrency Markets. Journal of Financial Economics **135**(2), 293–319 (2020).
11. Alexander, C., Heck, D.F.: Price Discovery in Bitcoin: The Impact of Unregulated Markets. Journal of Financial Stability **50**, 100776 (2020).
12. Corbet, S., Lucey, B., Urquhart, A., Yarovaya, L.: Cryptocurrency Reaction to FOMC Announcements. Journal of Financial Stability **46**, 100706 (2019).
13. Bouri, E., Gupta, R., Roubaud, D.: Trading Volume and the Predictability of Return and Volatility in the Cryptocurrency Market. Finance Research Letters **29**, 340–346 (2019).

14. Jiang, Y., Nie, H., Ruan, W.: Time-Varying Long-Term Memory in Cryptocurrency Markets. Finance Research Letters **38**, 101442 (2021).
15. Zhang, Y., et al.: FactorMiner: Automated Factor Discovery via LLM-based Symbolic Synthesis. arXiv:2412.04215 (2024).
16. Wang, S., Yuan, H., Ni, L.M., Guo, J.: QuantAgent: Seeking Holy Grail in Trading by Self-Improving Large Language Model. arXiv:2402.03755 (2024).
17. Bieganowski, B., Ślepaczuk, R.: Explainable Patterns in Cryptocurrency Microstructure. arXiv:2602.00776 (2026).
18. Yang, X., Liu, W., Zhou, D., Bian, J., Liu, T.Y.: Qlib: An AI-Oriented Quantitative Investment Platform. arXiv:2009.11189 (2020).
19. Wei, J., Wang, X., Schuurmans, D., et al.: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022).
20. Yao, S., Zhao, J., Yu, D., et al.: ReAct: Synergizing Reasoning and Acting in Language Models. In: International Conference on Learning Representations (ICLR) (2023).
21. Wu, S., Irsoy, O., Lu, S., et al.: BloombergGPT: A Large Language Model for Finance. arXiv:2303.17564 (2023).
22. Huang, A.H., Wang, H., Yang, Y.: FinBERT: A Large Language Model for Extracting Information from Financial Text. Contemporary Accounting Research **40**(2), 806–841 (2023).
23. Deng, Z., Wang, J.: AlphaQuanter: An End-to-End Tool-Orchestrated Agentic Reinforcement Learning Framework for Stock Trading. arXiv:2510.14264 (2025).
24. Zhu, Y., et al.: GEMs-LLM: Integrating Large Language Models with Goal-Aware Exploration for RL-Based Portfolio Optimization. In: 21st International Conference on Intelligent Computing (ICIC 2025), pp. 475–486. Springer, Singapore (2025). https://doi.org/10.1007/978-981-96-9949-0_43
25. Shen, Z., et al.: Smart Trading: A Novel Reinforcement Learning Framework for Quantitative Trading in Noisy Markets. In: 20th International Conference on Intelligent Computing (ICIC 2024), Part I, pp. 158–170. Springer, Singapore (2024).
26. Shinn, N., Cassano, F., Gopinath, A., et al.: Reflexion: Language Agents with Verbal Reinforcement Learning. Advances in Neural Information Processing Systems **36**, 8634–8652 (2023).
27. Wang, L., Ma, C., Feng, X., et al.: A Survey on Large Language Model Based Autonomous Agents. Frontiers of Computer Science **18**(6), 1–26 (2024).
28. Chen, B., Liu, L., Ding, H., et al.: AlphaSAGE: Structure-Aware Alpha Mining via GFlowNets for Robust Exploration. arXiv:2509.25055 (2025).
29. Li, Z., Song, R., Sun, C., Xu, W., Yu, Z., Wen, J.-R.: Can Large Language Models Mine Interpretable Financial Factors More Effectively? A Neural-Symbolic Factor Mining Agent Model. In: Findings of the Association for Computational Linguistics: ACL 2024, pp. 3891–3902 (2024).
30. Han, J., et al.: QuantaAlpha: An Evolutionary Framework for LLM-Driven Alpha Mining. arXiv:2602.07085 (2026).
31. Black, F., Litterman, R.: Global Portfolio Optimization. Financial Analysts Journal **48**(5), 28–43 (1992).
32. López de Prado, M.: Building Diversified Portfolios that Outperform. The Journal of Portfolio Management **42**(4), 59–69 (2016).
33. Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. ACM (2016).

34. Gurgul, V., Lessmann, S., Härdle, W.K.: Deep Learning and NLP in Cryptocurrency Forecasting: Integrating Financial, Blockchain, and Social Media Data. International Journal of Forecasting (2024).