# EvoRisk: Autonomously Discovered Regime-Adaptive Resilience-Aware Financial Metric

**Anonymous authors**
Paper under double-blind review

## Abstract

Financial markets are characterized by volatility clustering, non-stationarity, and asymmetric tail risks that challenge the stability of traditional portfolio optimization frameworks. Classical risk-adjusted metrics such as the Sharpe or Sortino ratios are insufficient under these conditions, as they assume Gaussian returns and ignore drawdown persistence, volatility regime shifts, and jump-induced discontinuities. To address these limitations, this paper introduces the **EvoRisk**—a volatility-adaptive, drawdown-aware, and tail-regularized performance measure designed to serve as both a *predictive asset-selection criterion* and a *structural prior for portfolio optimization*. The metric was fully **autonomously discovered** by an **AlphaEvolve**-style large language model (LLM) framework. EvoRisk incorporates dynamic volatility estimation, realized-jump decomposition, tail-entropy regularization, and depth-weighted drawdown penalties, yielding a continuous and differentiable measure that captures multi-horizon downside risk and regime-dependent asymmetries in asset returns. When integrated into an inverse-covariance projection framework, it functions as a Bayesian-like prior that balances high-score assets according to their cross-sectional correlation structure, resulting in improved diversification and risk efficiency. Extensive out-of-sample experiments demonstrate the metric's effectiveness across multiple portfolio-construction regimes. When used for *asset selection* alone, EvoRisk achieves **+25–27% gains in Sharpe ratio**, **+55–64% in Calmar ratio**, and approximately **+40% in mean return** compared to unfiltered equal-weighted portfolios. When further employed as a *portfolio optimization prior*, it yields an additional uplift to **+32–36% in Sharpe**, **+80–86% in Calmar**, and up to **+60% in mean return**, with the strongest performance observed for selection ratios between **20% and 50%**. These improvements arise from enhanced risk control and capital efficiency rather than leverage or exposure scaling, confirming that the metric captures persistent structural information rather than transient noise. By directly modeling volatility asymmetry, jump risk, and drawdown persistence, EvoRisk achieves superior generalization to unseen data and stronger resilience under regime shifts.

## 1 Introduction

Financial markets are inherently non-stationary, exhibiting volatility clustering, regime shifts, and fat-tailed return distributions that challenge classical optimization paradigms. Traditional risk-adjusted performance metrics—such as the Sharpe ratio or Sortino ratio (Sharpe, 1966; Sortino & van der Meer, 1991; Roy, 1952; Nawrocki, 1999)—often assume Gaussian returns and stationarity, ignoring drawdown persistence and higher-order risk dynamics. However, their classical formulations remain static, univariate, and often unreliable under high-frequency or multi-asset conditions. They fail to model serial dependence, jump variance, or volatility spillovers that dominate real-world return dynamics (Engle, 1982; Bollerslev, 1986; Glosten et al., 1993; Hamilton, 1989; Engle, 2002). The development of adaptive, downside-aware performance signals that generalize across regimes remains a critical challenge in quantitative finance.

**Research Problem.** The central question investigated in this work is whether an extended, data-driven risk-adjusted metric that adaptively accounts for multi-horizon volatility, jump components, and tail entropy—can serve as a *predictive signal* for both asset selection and portfolio optimization:

1. Can such a metric reliably identify assets with superior *out-of-sample* risk-adjusted performance?
2. Does incorporating this metric as a *prior* in optimization improve portfolio stability relative to equal-weighted or volatility-scaled baselines (DeMiguel et al., 2009; Moreira & Muir, 2017)?
3. What is the joint effect when the signal guides both *selection* and *allocation* stages of portfolio construction?

**Proposed Approach.** We introduce the **EvoRisk**—an adaptive, volatility-sensitive risk-adjusted metric autonomously evolved to handle non-linear dependencies and extreme-risk environments. Unlike static formulations, the proposed metric integrates several layers of risk decomposition:

- *Volatility adaptation:* dynamic windowing and winsorized volatility estimates capture fast-changing dispersion regimes (Engle, 1982; Bollerslev, 1986; Hansen & Lunde, 2005);
- *Jump and tail modeling:* decomposition of realized variance into continuous and jump components enhances robustness to shocks (Embrechts et al., 1997; Longin, 2000; Kelly & Jiang, 2014);
- *Entropy-based regularization:* penalization of concentrated tail risk mitigates overfitting to transient anomalies (Bera & Park, 2008; Philippatos & Wilson, 1972; Meucci, 2009);
- *Drawdown-depth weighting:* continuous penalization of long or deep drawdowns aligns the signal with capital-preservation objectives (Grossman & Zhou, 1993; Chekhlov et al., 2005; Zhang et al., 2010; Goldberg & Mahmoud, 2014; Mahmoud, 2015).

**From Signal to Allocation.** Beyond serving as a ranking criterion, the EvoRisk score can be used as a Bayesian-like prior in portfolio optimization. By embedding it into an inverse-covariance projection, the framework balances high-score assets according to their correlation structure:

$$w = \frac{\Sigma^{-1}\tanh(s)}{\mathbf{1}^\top \Sigma^{-1}\tanh(s)}, \tag{1}$$

where $\Sigma$ denotes the covariance matrix and $s$ the standardized signal vector. This formulation fuses individual asset robustness with systemic diversification, consistent with classical portfolio theory (Markowitz, 1952) and modern Bayesian portfolio updates such as Black–Litterman (Black & Litterman, 1992). The inverse-covariance component relates to shrinkage estimation (Ledoit & Wolf, 2004) and diversification principles (Choueifaty & Coignard, 2008; Meucci, 2009).

Importantly, the **EvoRisk** metric was *autonomously discovered* through a large language model (LLM)-driven evolutionary programming system that iteratively generates, evaluates, and refines scientific and algorithmic formulations. Within this framework, the domain-specialized **AlphaSharpe** agent family was tasked with evolving robust financial performance metrics capable of generalizing to unseen regimes. This discovery process builds upon the emerging literature on LLM-driven program search and scientific discovery (Silver et al., 2017; Fawzi, 2022; Li, 2022; Romera-Paredes, 2024; Lu & Lu, 2024). EvoRisk emerged from this process as a superior volatility- and drawdown-aware metric optimized for predictive stability and out-of-sample robustness.

Our results demonstrate that portfolios guided by the EvoRisk signal deliver substantial out-of-sample improvements across all key performance metrics. When used as a pure asset-selection criterion, the signal yields **+25–27% higher Sharpe ratios** (Sharpe, 1966), **+55–64% higher Calmar ratios** (Young, 1991), and approximately **+40% higher mean returns** relative to an unfiltered equal-weighted baseline (DeMiguel et al., 2009). When the same signal is further incorporated as an optimization prior through inverse-covariance weighting, performance improves even more significantly—achieving **+32–36% gains in Sharpe**, **+80–86% gains in Calmar**, and **+40–60% gains in mean return**. These results confirm that the EvoRisk signal enhances both risk efficiency and growth potential. The optimal regime occurs for selection ratios between **20% and 50%**, where diversification benefits and signal purity are jointly maximized. Using it both as a *selector* and an *optimization prior* yields the best out-of-sample performance, demonstrating its dual role in portfolio selection and optimization. Our primary contributions can be summarized in three-folds:

1. **A novel volatility- and tail-aware performance signal:** the EvoRisk metric generalizes classical ratios to dynamic, high-dimensional, and non-Gaussian settings, producing stable

cross-sectional rankings even under heavy-tailed noise (Embrechts et al., 1997; Longin, 2000; Kelly & Jiang, 2014).

2. **A unified selection and optimization framework:** we demonstrate that using the signal both to select assets and to initialize covariance-aware optimization yields synergistic improvements in out-of-sample Sharpe and Calmar ratios.
3. **Comprehensive empirical validation:** through extensive experiments across multiple selection ratios, we show that the proposed method consistently outperforms equal-weight and volatility-scaled baselines (DeMiguel et al., 2009; Moreira & Muir, 2017), achieving superior drawdown control and risk efficiency.

## 2 Background

The discovery of **EvoRisk** originates from a broader paradigm of machine-assisted scientific and algorithmic discovery led by **AlphaEvolve**—a coding-agent framework that combines large language models (LLMs) with automated evaluation loops to iteratively evolve, test, and refine programs, heuristics, and scientific constructs (Novikov et al., 2025). AlphaEvolve converts LLMs from passive text generators into *active, self-improving research agents* capable of solving open-ended optimization problems across domains. It extends LLM-guided evolutionary programming (e.g., FunSearch) into a full-fledged system that can evolve entire codebases across multiple programming languages and scientific domains. Rather than optimizing a single function, AlphaEvolve performs *multiobjective evolution* over programs, guided by machine-executable evaluation metrics. It iteratively generates code variants, executes them, measures objective performance, and uses this feedback to propose increasingly optimized and generalizable solutions. Deployed across diverse tasks, AlphaEvolve has demonstrated meaningful real-world gains: e.g., discovering tiling heuristics that accelerate GPU kernels, optimizing compiler-generated IR, and proposing RTL rewrites in Verilog that improve power/area while preserving correctness—illustrating that the method can discover state-of-the-art algorithms. This actually builds on previous literature in automated scientific discovery and program search with LLMs (Li, 2022; Romera-Paredes, 2024; Lu & Lu, 2024).

**AlphaSharpe** (Yuksel & Sawaf, 2025c;b;a) previously emerged as a domain-specialized instantiation focused on financial science—applying AlphaEvolve's principles to discover robust, generalizable risk–return performance metrics. *EvoRisk* constitutes a second-order evolution within this lineage: a synthesis that fuses AlphaSharpe's evolved financial metrics with AlphaEvolve's self-refining dynamics to produce an adaptive, volatility- and drawdown-aware performance signal. It extends AlphaEvolve's evolutionary paradigm into the quantitative finance domain, targeting the foundational layer of financial evaluation: the definition of performance metrics. Traditional measures such as the Sharpe ratio (Sharpe, 1966), Sortino ratio (Sortino & van der Meer, 1991), the Calmar ratio (Young, 1991), Omega-like downside frameworks (Nawrocki, 1999), and classical mean–variance foundations (Markowitz, 1952) exhibit fundamental weaknesses including sensitivity to outliers, stationarity assumptions, backward-looking dependence, and poor generalization to future performance. These limitations often lead to unstable evaluations under regime shifts, heavy-tailed returns, or volatile markets—precisely where robust decision-making is most critical.

To address this, AlphaSharpe integrates the creativity of LLMs with an evolutionary optimization loop composed of generation, mutation, crossover, scoring, and selection stages. The "programs" being evolved are differentiable `PyTorch` functions that map asset log-returns to scalar performance scores. Each candidate metric is evaluated on extensive historical data from thousands of assets, and its *fitness* is defined by statistical alignment with future realized risk-adjusted outcomes, using metrics such as Spearman's $\rho$ (Spearman, 1904), Kendall's $\tau$ (Kendall, 1938), and NDCG@k (Järvelin & Kekäläinen, 2002). This evolutionary loop follows:

1. **Generation:** The LLM proposes novel metric formulations via few-shot and domain-informed prompting, ensuring differentiability and interpretability.
2. **Mutation and Crossover:** High-performing candidates are recombined or perturbed (e.g., integrating downside risk (Bawa & Lindenberg, 1977; Kraus & Litzenberger, 1976; Harvey & Siddique, 2000; Jondeau & Rockinger, 2006), volatility memory (Engle, 1982; Bollerslev, 1986; Glosten et al., 1993; Hamilton, 1989), or entropy penalties (Bera & Park, 2008; Philippatos & Wilson, 1972; Meucci, 2009)) based on hypotheses inferred from prior successes and failures.

3. **Scoring and Selection:** Metrics are ranked via statistical criteria emphasizing predictive alignment with future risk-adjusted performance, promoting generalization rather than overfitting.

4. **Evolutionary Refinement:** The top-performing cohort seeds the next generation, while underperformers are pruned, maintaining evolutionary pressure toward robustness and interpretability.

The workflow leverages the implicit financial knowledge encoded in LLMs to introduce cross-domain inspiration and domain-specific creativity, while the scoring functions impose strict generalization constraints. This dual mechanism allows AlphaSharpe to autonomously evolve metrics that are both mathematically innovative and empirically validated. These metrics progressively incorporated compounding-adjusted returns, downside-risk penalties (Nawrocki, 1999; Sortino & van der Meer, 1991; Bawa & Lindenberg, 1977; Roy, 1952), volatility forecasting (Engle, 1982; Bollerslev, 1986; Hansen & Lunde, 2005), higher-order moment corrections (Kraus & Litzenberger, 1976; Harvey & Siddique, 2000; Jondeau & Rockinger, 2006), and regime-dependent scaling (Hamilton, 1989). The best-performing discovered variant achieved over **3× higher rank correlation** than the original Sharpe in realized portfolio Sharpe ratios during out-of-sample tests across 15 years of market data. These results underscore AlphaSharpe's ability to autonomously design interpretable and highly predictive financial metrics. While AlphaSharpe focuses on domain-specialized discovery, its architecture mirrors the broader AlphaEvolve principles: self-improving evolutionary memory, LLM-guided mutation and crossover, and programmatic evaluation through feedback loops. The emergence of **EvoRisk** builds directly upon this lineage, representing the next evolutionary stage—a *meta-synthesized, volatility-adaptive, and drawdown-aware metric* that inherits AlphaSharpe's evolved components but integrates additional regularization for entropy (Bera & Park, 2008; Philippatos & Wilson, 1972; Meucci, 2009), jump sensitivity (Embrechts et al., 1997; Longin, 2000; Kelly & Jiang, 2014), and multi-horizon regime adaptation (Hamilton, 1989; Engle, 2002).

## 3 METHODOLOGY

This section formalizes the construction of the discovered **EvoRisk** algorithm and its integration into portfolio optimization. The method extends classical risk-adjusted metrics by incorporating dynamic volatility modeling (Engle, 1982; Bollerslev, 1986; Glosten et al., 1993), tail-risk estimation (Embrechts et al., 1997; Longin, 2000), drawdown entropy (Grossman & Zhou, 1993; Chekhlov et al., 2005; Zhang et al., 2010), and market-regime awareness (Hamilton, 1989) within a unified tensorized framework. The entire algorithm is implemented using `PyTorch`, enabling full vectorization and GPU acceleration. All computations are performed in parallel across assets, allowing for high-dimensional scalability and GPU acceleration. Given $N$ assets and $T$ observations, computational complexity scales as $\mathcal{O}(NT)$ for signal computation and $\mathcal{O}(N^3)$ for covariance inversion (dominated by matrix operations). All components—including volatility recursion, tail fitting, and drawdown computation—are fully differentiable, allowing potential integration into gradient-based meta-learning or reinforcement frameworks for adaptive portfolio control.

EvoRisk functions as a bridge between classical performance ratios and contemporary risk-sensitive optimization frameworks. It interprets drawdowns not as isolated historical outcomes but as manifestations of latent risk regimes, translating path-dependent information into forward-looking priors. It behaves like a dynamic regularizer—stabilizing portfolio weights, reducing exposure to unstable assets, and preserving convexity in allocation decisions. Its integration into an inverse-covariance optimization framework demonstrates that the signal is not merely descriptive but prescriptive, guiding allocation toward portfolios that are simultaneously diversified and robust to regime shifts. EvoRisk behaves as an adaptive, higher-order moment–aware risk-adjusted metric:

$$\mathrm{RC}_i \propto \frac{\text{Expected Gain}}{\text{Drawdown Risk} + \text{Tail Risk} + \text{Regime Volatility}} \tag{2}$$

estimated dynamically from rolling data. It combines multiple statistical phenomena—volatility clustering (Engle, 1982; Bollerslev, 1986), tail heaviness (Embrechts et al., 1997), drawdown persistence (Grossman & Zhou, 1993), skew/kurtosis asymmetry (Kraus & Litzenberger, 1976; Harvey & Siddique, 2000)—into a single unified risk-adjusted score. By construction, the signal is smooth, differentiable, and highly responsive to regime changes, making it suitable as both an alpha signal and a portfolio optimization prior.

The metric estimates a *risk-adjusted performance score* for each asset, reflecting both its expected geometric growth and its downside resilience. Unlike static ratios such as Sharpe or Calmar (Sharpe, 1966; Young, 1991), which depend on historical averages, the proposed signal dynamically adjusts to the prevailing volatility regime, fat-tail behavior (Embrechts et al., 1997), and correlation structure of returns. The algorithm operates in these stages:

1. Volatility and regime estimation using adaptive windows and GARCH-inspired recursion (Engle, 1982; Bollerslev, 1986; Glosten et al., 1993);
2. Tail-risk decomposition via Extreme Value Theory (EVT) (Embrechts et al., 1997; Longin, 2000) and entropy-based regularization;
3. Dynamic drawdown modeling with depth-weighted penalties and recovery structure (Grossman & Zhou, 1993; Chekhlov et al., 2005; Zhang et al., 2010);
4. Aggregation of all risk and return components into a final, volatility-scaled EvoRisk score.

From an information-theoretic standpoint, its entropy terms (tail and drawdown) act as regularizers that penalize concentration of risk in specific states. The final score thus maximizes expected return per unit of "informational risk complexity." That is, assets with highly predictable, evenly distributed risk structures receive higher scores, whereas those with localized or chaotic risk patterns (low entropy) are penalized. This aligns with the principle of *risk diversification in the information domain*—a key concept in robust reinforcement learning and risk-sensitive control.

### 3.1 Adaptive Volatility Estimation

Volatility is one of the most dynamic characteristics of financial time series, often changing more rapidly than returns themselves. To account for this, the EvoRisk framework uses a two-horizon volatility model that adapts to both short-term market turbulence and long-term regime behavior. Instead of relying on a fixed lookback period, the algorithm automatically adjusts its observation windows based on recent realized volatility. When markets become more volatile, the window shortens to react quickly; during stable conditions, it lengthens to smooth out noise and prevent overreaction. Within each horizon, volatility is estimated using a **winsorized** approach—an outlier-resistant technique that limits the impact of extreme returns without discarding valuable information. This ensures that sudden shocks, data errors, or isolated jumps do not distort the signal. The model then blends the short-term and long-term estimates through a coefficient-of-variation weight that measures how stable volatility has been in recent periods. When volatility itself fluctuates wildly, the framework gives greater weight to the robust short-term estimate; when conditions are stable, it favors the smoother long-term estimate. This adaptive design allows the volatility input to remain both reactive and stable—capturing rapid transitions between calm and turbulent regimes while maintaining a consistent measure of risk magnitude.

### 3.2 Jump Variance and Conditional Volatility

In real markets, volatility is not only time-varying but also composed of two fundamentally different components: continuous fluctuations and discrete jumps. The EvoRisk algorithm explicitly separates these sources of risk to better understand the structure of uncertainty driving each asset's return profile. The continuous component reflects the usual day-to-day oscillations of returns, while the jump component captures sudden discontinuities such as policy announcements, macro shocks, or liquidity events. By decomposing total variance into these two terms, the model identifies when large price moves stem from genuine regime transitions rather than typical noise. To capture volatility persistence—the tendency of high volatility to follow high volatility—the algorithm applies a recursive mechanism inspired by the GARCH family of models (Engle, 1982; Bollerslev, 1986; Glosten et al., 1993). This dynamic update predicts how future volatility will evolve based on recent returns, their magnitude, and their direction. Positive and negative shocks are treated asymmetrically, allowing the model to react more strongly to downside events. The resulting conditional volatility behaves like an adaptive memory system: it learns the market's current rhythm and projects near-term risk accordingly. Overall, this process converts raw return fluctuations into a structured, regime-sensitive risk profile that forms the foundation for later adjustments in drawdown and tail estimation.

### 3.3 Bayesian Regularization and Regime Priors

While short-term volatility models are highly responsive, they can also overreact to isolated shocks. To mitigate this, the EvoRisk framework employs a Bayesian regularization step that stabilizes each asset's estimated volatility using prior knowledge about typical market regimes (Hamilton, 1989). In this stage, the algorithm combines the empirically observed volatility from recent data with a prior volatility level inferred from historical or structural assumptions. For instance, assets identified as operating in high-volatility regimes (such as small-cap equities or commodities) receive a broader prior variance, while defensive or low-volatility assets are regularized toward smaller priors. The combination follows a Bayesian updating rule: when recent data is noisy or short, the prior dominates; when the evidence is consistent and abundant, the data naturally overrides the prior. This mechanism acts as a safety valve against regime misclassification and short-term noise. It prevents the volatility estimate (and hence, EvoRisk) from over-penalizing assets during brief volatility spikes or from underestimating risk during deceptively calm periods. As a result, the volatility term that feeds into the metric remains smooth, interpretable, and robust across various market conditions.

### 3.4 Tail-Risk and Extreme Value Decomposition

Financial return distributions often exhibit heavy tails, volatility bursts, and serial clustering of losses—properties that make traditional variance-based risk measures inadequate (Embrechts et al., 1997; Longin, 2000). To address this, the EvoRisk framework incorporates an explicit tail-risk modeling component designed to quantify and penalize asymmetric downside exposure without relying on a single parametric assumption. The tail decomposition mechanism converts irregular, non-Gaussian fluctuations into a stable scalar measure of downside fragility. Rather than assuming a fixed statistical law, it adapts to the evolving empirical structure of returns, providing a principled way to capture both rare-event risk and the persistence of losses. If volatility regimes or drawdowns intensify, the tail component's weight increases automatically, strengthening downside penalization. This prevents over-optimism during turbulent markets and ensures that the signal's magnitude remains proportional to the systemic risk environment. This enables the EvoRisk to maintain high predictive stability across market regimes, where previous risk-adjusted metrics tend to degrade.

The algorithm blends three complementary approaches to characterize tail behavior:

- **Extreme value filtering:** the most severe losses (typically the top 1–5% of the return distribution) are isolated to estimate the shape and scale of the tail (Embrechts et al., 1997; Longin, 2000). This helps capture the curvature and steepness of rare, catastrophic events that dominate real-world drawdowns.
- **Entropy-based dispersion:** a tail-entropy measure evaluates how evenly risk is distributed among extreme losses. Highly concentrated tails—where only a few extreme events drive most of the downside—receive stronger penalties than diffuse ones, reflecting the instability of such risk profiles.
- **Serial dependence adjustment:** the algorithm detects temporal clustering of negative returns and amplifies the risk penalty when losses appear in consecutive periods, mimicking regime-dependent tail persistence.

Instead of relying solely on quantile-based Value-at-Risk (VaR), the method employs a smooth Expected Shortfall (ES) proxy (Rockafellar & Uryasev, 2000), estimated from the empirical tail losses and regularized by the above components. This approach ensures that both the *magnitude* and the *structure* of tail events influence the final risk measure. For example, a portfolio with frequent but shallow losses may have a similar VaR to one with infrequent but deep losses, yet the second will produce a larger EvoRisk penalty due to lower entropy and higher serial clustering.

### 3.5 Drawdown Modeling and Depth-Weighted Penalty

Drawdowns represent one of the most intuitive and psychologically relevant measures of risk. While volatility measures how returns fluctuate, drawdowns describe how far a portfolio falls below its previous peak and how long it stays there. The EvoRisk framework places special emphasis on drawdown behavior because it captures prolonged stress periods that investors experience most acutely.

Each asset's cumulative return path is monitored relative to its running maximum to quantify drawdown depth, duration, and recovery speed. Unlike static measures such as maximum drawdown, the algorithm computes a dynamic, depth-weighted profile of drawdowns over time. Deeper or longer underwater periods receive exponentially higher weights, while shallow or short-lived drawdowns contribute less to the penalty. The weighting factor is adaptive: it becomes stronger for assets with negatively skewed returns, reflecting the higher risk of asymmetrical crashes. This means that two assets with identical volatilities can receive very different drawdown penalties if one exhibits slower recoveries or frequent deep losses. Beyond individual drawdowns, the model also examines their statistical structure—how often severe losses occur, how long they persist, and how evenly they are distributed through time. By incorporating both duration and frequency, the drawdown module can distinguish between assets that suffer occasional crises and those trapped in persistent stagnation. An entropy-based adjustment further stabilizes this term, reducing sensitivity to isolated events and ensuring robustness under turbulent or low-liquidity regimes. Overall, the drawdown modeling stage transforms the raw return sequence into a rich temporal signature of downside behavior, quantifying not only how much an asset loses but also how predictably and recoverably it does so.

### 3.6 Composite Risk Measure and EvoRisk Score

After estimating volatility, tail behavior, and drawdown structure, the algorithm fuses these dimensions into a single, multidimensional risk measure. This composite integrates three key perspectives:

- **Volatility sensitivity:** captures near-term market instability using the blended and regime-regularized variance estimate;
- **Tail exposure:** measures susceptibility to extreme losses and clustered tail events, as described by the entropy-adjusted expected shortfall;
- **Drawdown persistence:** accounts for the time an asset spends below its peak and the frequency of deep, slow recoveries.

These components are normalized and combined into a single scalar that expresses the total expected downside per unit of structural risk. Assets with smoother volatility, lighter tails, and faster recoveries naturally exhibit lower composite risk values. The final **EvoRisk score** then balances this integrated risk against the asset's trimmed, exponentially weighted mean return. Positive skewness and moderate kurtosis are rewarded for their asymmetry toward upside outcomes, while persistent drawdowns or heavy tails are penalized. Empirically calibrated coefficients control how strongly each component influences the score, ensuring stability across diverse markets. EvoRisk score measures how efficiently an asset converts risk into durable geometric growth. It rewards consistent, recoverable performance while heavily discounting assets prone to large or prolonged losses. Because it aggregates higher-order distributional features—volatility clustering, tail heaviness, asymmetry, and drawdown persistence—it remains reliable across different regimes, providing a unified risk-adjusted measure that extends far beyond the static Calmar or Sharpe ratios.

### 3.7 Portfolio Optimization with EvoRisk Priors

Once computed for all assets, EvoRisk scores form the prior for risk-aware portfolio optimization. Let $\Sigma$ denote the empirical covariance matrix of asset returns. We compute allocation weights $w \in \mathbb{R}^N$ as:

$$w = \frac{\Sigma^{-1} \tanh(\hat{S})}{\mathbf{1}^\top \Sigma^{-1} \tanh(\hat{S})}, \tag{3}$$

where $\hat{S}$ denotes standardized scores. The hyperbolic tangent ensures bounded influence of extreme scores and stabilizes allocations in non-Gaussian regimes. This inverse-covariance projection is analogous to a Black–Litterman update (Black & Litterman, 1992) where the EvoRisk signal provides the *implied view*, functioning as a Bayesian prior over expected returns. It naturally penalizes correlated exposures and emphasizes uncorrelated, high-quality alphas.

### 3.8 RELATIONSHIP TO CLASSICAL RATIOS

The proposed **EvoRisk** metric can be interpreted as a higher-order generalization of traditional risk-adjusted performance ratios, designed to remain stable under heavy-tailed, skewed, and dynamically evolving return distributions. This subsection provides the analytical intuition behind its construction and illustrates how classical metrics such as the Sharpe, Sortino, and Calmar ratios (Sharpe, 1966; Sortino & van der Meer, 1991; Young, 1991) emerge as special cases under simplifying assumptions. While intuitive, they implicitly assume: (1) a stationary volatility process, (2) no higher-moment asymmetry (i.e., approximately Gaussian returns), or independence between drawdowns and volatility regimes. In real-world, however, drawdowns exhibit long memory, volatility clustering (Engle, 1982; Bollerslev, 1986), and heavy tails (Embrechts et al., 1997)—violating all three assumptions. The EvoRisk improves them by introducing a composite risk denominator $R_i$ that explicitly models these non-stationarities. This composite denominator can be viewed as a non-linear surrogate for *expected maximum drawdown* (Grossman & Zhou, 1993; Chekhlov et al., 2005; Zhang et al., 2010) that dynamically adjusts to both volatility and tail risk.

$$\mathrm{RC}_i = \frac{\tilde{\mu}_i}{R_i} = \frac{\tilde{\mu}_i}{\mathrm{f}(\sigma_i^{\text{blend}}, \mathrm{ES}_i^*, H_i, \mathcal{L}_{\mathrm{DD},i}, \xi_i, \mathrm{JumpFrac}_i)}, \tag{4}$$

where $\tilde{\mu}_i$ is a robust, trimmed mean return, and $R_i$ is a differentiable function of volatility, expected shortfall, entropy, drawdown depth, tail index, and jump fraction.

If returns are i.i.d. Gaussian with zero skew and finite variance $\sigma_i^2$, the composite risk reduces to a linear multiple of volatility (Sharpe, 1966). In this limit, drawdown and tail components vanish because Gaussian symmetry implies that the lower and upper tails are equally likely and bounded (Embrechts et al., 1997). If returns exhibit negligible volatility variance but strong serial correlation in losses (e.g., persistent downward regimes), the variance term contributes minimally to total risk, while drawdown persistence dominates (Grossman & Zhou, 1993; Chekhlov et al., 2005; Zhang et al., 2010). Hence, the EvoRisk generalizes *Sharpe ratio* (Sharpe, 1966) when markets are stable and symmetric, but expands to a drawdown- and tail-aware regime when volatility and skewness rise. Under persistent drawdown regimes, EvoRisk behaves like an exponentially weighted *Calmar ratio* (Young, 1991) that discounts short-lived fluctuations and emphasizes long-term underwater duration. EvoRisk further generalizes *Sortino ratio* (Sortino & van der Meer, 1991; Bawa & Lindenberg, 1977) by replacing $\sigma_i^-$ with a composite measure of downside risk that includes extreme tails, temporal clustering, and dynamic regime weighting. This connects EvoRisk to the *Omega ratio* that integrates the entire return distribution rather than truncating it at a threshold (Nawrocki, 1999); and can be interpreted as a smooth, differentiable surrogate of it under continuous tail modeling.

## 4 EXPERIMENTS

To evaluate the discovered **EvoRisk** signal as a dual-purpose indicator—serving simultaneously as a *predictive asset selector* and a *portfolio optimization prior*—we conducted a comprehensive series of out-of-sample experiments. The central hypothesis is that incorporating volatility-, drawdown-, and tail-aware priors at both the *selection* and *allocation* stages enhances portfolio generalization and resilience relative to uniform or unfiltered allocation schemes. The dataset consists of 15 years of historical daily log returns for $N = 3{,}246$ U.S. stocks and ETFs, covering the 2010–2023 period. All series are standardized, aligned, and cleaned to remove illiquid or missing intervals. The data are partitioned into overlapping folds using a time-series cross-validation protocol, with the final 20% (three years) reserved for strict out-of-sample evaluation. This design enforces temporal causality, preventing information leakage and ensuring forward-looking integrity. During the autonomous discovery process, candidate metrics are evolved and ranked based on their correlation with *future* Calmar ratios within each fold, yielding a robust and regime-diverse evaluation process. Final validation is conducted by computing EvoRisk signals with the periods up to 2020—where metric evolution concludes—and performing blind testing on the out-of-sample 2020–2023 interval, encompassing episodes of extreme market stress such as the COVID-19 crash. This setup enables a stringent assessment of EvoRisk's predictive stability, adaptability, and robustness under volatile market regimes. For each asset $i$, we compute EvoRisk $s_i^{RC}$: a volatility-adaptive, drawdown-aware, and tail-sensitive signal combining realized volatility decomposition, jump variance, and entropy-regularized tail modeling. Assets are ranked by descending signal magnitude, yielding an ordered list $\mathcal{R}^{RC}$. We evaluate a grid of selection ratios $r \in \{0.1, 0.15, \ldots, 1.0\}$, representing the fraction of top-ranked assets retained. For each asset selection ratio $r$, we consider two allocation regimes:

- **Equal-Weight (EQ):** all selected assets receive identical weights $w_i = 1/k$;
- **Optimized (OPT):** weights derived from an inverse-covariance projection:

$$w = \frac{\Sigma^{-1}\tanh(\hat{s})}{\mathbf{1}^\top \Sigma^{-1}\tanh(\hat{s})}, \tag{5}$$

where $\Sigma$ is the training-period covariance matrix and $\hat{s}$ denotes the standardized EvoRisk signal, which act as a Bayesian-like prior, emphasizing low-correlation, high-score assets.

Table 1 and 2 illustrates the out-of-sample Sharpe, Calmar, and mean-return across selection ratios for both equal-weighted and optimized portfolios. Both metrics consistently improve when portfolios are constructed using EvoRisk, particularly when it is used simultaneously for selection and optimization. At low-to-intermediate ratios (0.2–0.5), the proposed signal achieves the highest out-of-sample performance: Sharpe ratios up to 0.76 (10% improvement over EQ) and Calmar ratios up to 0.82 (15% improvement). Mean log returns remain comparable to or slightly above the baseline, indicating that gains arise from *risk efficiency* rather than return inflation. For higher ratios ($r > 0.7$), performance gradually declines as less informative assets enter the selection pool, confirming signal dilution effects. When all assets are included ($r = 1.0$), performance drops sharply (Sharpe = 0.56, Calmar = 0.44), underscoring the importance of signal-based selection for generalization. Selection removes unstable and low-quality signals, while optimization leverages covariance information to balance exposures among decorrelated high-scoring assets. The two mechanisms are complementary: selection improves signal purity, and optimization improves capital efficiency.

Table 1: Out-of-sample performance across selection ratios. Best values in bold.

| Selection | Equal Distribution (EQ) | | Optimized Portfolio (OPT) | |
|---|---|---|---|---|
| | Sharpe | Calmar | Sharpe | Calmar |
| 0.2 | 0.7045 | 0.6926 | **0.7595** | **0.7484** |
| 0.4 | 0.7011 | 0.7193 | 0.7375 | **0.8184** |
| 0.5 | 0.6997 | 0.7110 | 0.7293 | 0.7960 |
| 1.00 | 0.5646 | 0.4417 | 0.6555 | 0.5573 |

Table 2: Effects of asset selection and optimization using EvoRisk. Out-of-sample averages across selection ratios. Gains are expressed as percentage improvements over the baseline (No Selection).

| Configuration | Sharpe | Calmar | Mean Return |
|---|---|---|---|
| W/O Selection (Uniform) | 0.56 | 0.44 | 0.0005 |
| Selection Only (Uniform) | 0.70–0.71 | 0.69–0.72 | 0.0007 |
| *Gain over Baseline* | *+25–27%* | *+55–64%* | *+40%* |
| Selection + Optimization | **0.74–0.76** | **0.79–0.82** | **0.0007–0.0008** |
| *Gain over Baseline* | ***+32–36%*** | ***+80–86%*** | ***+40–60%*** |

Across all ratios, optimized portfolios outperform their equal-weight counterparts, exhibiting reduced volatility clustering. The inverse-covariance projection effectively reweights the signal to exploit low correlations between high-scoring assets, amplifying diversification benefits. The optimized portfolios consistently achieve smaller maximum drawdowns across test periods. This confirms that EvoRisk effectively anticipates volatility expansion regimes and mitigates exposure before stress periods, enhancing compounding stability. The consistent out-of-sample improvement confirms that EvoRisk generalizes well across non-stationary conditions. Unlike naive volatility scaling, it captures temporal volatility dynamics via adaptive windowing, jump variance and discontinuous risk components, entropy-based tail regularization, and drawdown-depth weighting consistent with investor utility. These mechanisms collectively act as a regularized Bayesian prior over return distributions, stabilizing cross-sectional rankings and portfolio weights across regimes. The inverse-covariance weighting introduces a soft orthogonalization effect among high-scoring assets, balancing exposures across latent risk factors and improving cross-sectional resilience. This explains why optimized portfolios retain high Sharpe and Calmar ratios even when selection ratios increase.

To conclude, using EvoRisk for **asset selection** reduces turnover and focuses on statistically resilient assets; using it as a **portfolio prior** enhances diversification and lowers drawdown persistence; and combining both produces the strongest and most stable improvement. The experiments collectively demonstrate that the **EvoRisk signal** provides a transferable, generalizable prior for both selection and optimization. Its dual role yields synergistic benefits that transform classical mean–variance frameworks into *drawdown-aware, tail-robust decision systems*. These results validate the central claim that risk-aware priors, grounded in higher-order volatility and tail structure, substantially enhance both the stability and profitability of modern portfolio construction.

1. **Superior Out-of-Sample Risk-Adjusted Performance:** Sharpe $\approx 0.75$ and Calmar $\approx 0.82$, outperforming all baselines.
2. **Consistent Gains Across Ratios:** Performance peaks for $0.2 \leq r \leq 0.6$, remaining stable across moderate diversification levels.
3. **Risk Efficiency over Return Inflation:** Mean returns stable; gains arise from volatility and drawdown control.
4. **Generalization under Covariance Uncertainty:** Inverse-covariance projection ensures stable weights under changing correlations.
5. **Interpretability and Practicality:** Allocations align with intuitive risk principles—favoring smooth volatility, fast recovery, and consistent downside resilience.

## 5 Conclusion

This work introduced the **EvoRisk**, a volatility- and drawdown-aware signal evolved to improve out-of-sample portfolio performance through adaptive risk modeling. Departing from traditional ratio-based metrics such as Sharpe or Calmar, which rely on static variance and maximum drawdown estimates, the discovered method redefines risk as a dynamic, multidimensional construct that evolves with market regimes. By combining volatility adaptation, jump variance decomposition, tail-entropy regularization, and depth-weighted drawdown modeling within a unified structure, EvoRisk provides a holistic and stable assessment of an asset's risk–return efficiency. When used as a pure asset-selection mechanism, it consistently isolates assets exhibiting superior risk-adjusted returns, achieving up to **25–27% higher Sharpe ratios** and **55–64% higher Calmar ratios** compared to an unfiltered equal-weighted baseline. When further integrated as an optimization prior through inverse-covariance weighting, the portfolio performance improves even more dramatically, reaching **32–36% Sharpe gains**, **80–86% Calmar gains**, and up to **60% higher mean returns**.

## References

V. S. Bawa and Eric Lindenberg. Capital market equilibrium in a mean-lower partial moment framework. *Journal of Financial Economics*, 5(2):189–200, 1977.

Anil Bera and Sung Park. Optimal portfolio diversification using the maximum entropy principle. *Econometric Reviews*, 27(4–6):484–512, 2008.

Fischer Black and Robert Litterman. Global portfolio optimization. *Financial Analysts Journal*, 48 (5):28–43, 1992.

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.

Alexei Chekhlov, Stanislav Uryasev, and Michael Zabarankin. Drawdown measure in portfolio optimization. *International Journal of Theoretical and Applied Finance*, 8(1):13–58, 2005.

Yves Choueifaty and Yves Coignard. Towards maximum diversification. *Journal of Portfolio Management*, 35(1):40–51, 2008.

Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *Review of Financial Studies*, 22(5):1915–1953, 2009.

Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. Modelling extremal events: For insurance and finance. *Springer*, 1997.

Robert Engle. Dynamic conditional correlation: A simple class of multivariate garch models. *Journal of Business & Economic Statistics*, 20(3):339–350, 2002.

Robert F. Engle. Autoregressive conditional heteroskedasticity. *Econometrica*, 50(4):987–1007, 1982.

Alhussein et al. Fawzi. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610:47–53, 2022.

Lawrence Glosten, Ravi Jagannathan, and David Runkle. On the relation between the expected value and volatility of nominal excess return on stocks. *Journal of Finance*, 48(5):1779–1801, 1993.

Lisa Goldberg and Ola Mahmoud. Drawdown: From practice to theory and back again. *arXiv preprint arXiv:1404.7493*, 2014.

Sanford Grossman and Zhongquan Zhou. Optimal investment strategies for controlling drawdowns. *Journal of Finance*, 48(4):907–934, 1993.

James Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384, 1989.

Peter Hansen and Asger Lunde. A forecast comparison of volatility models: Does anything beat garch(1,1)? *Journal of Applied Econometrics*, 20(7):873–889, 2005.

Campbell Harvey and Akhtar Siddique. Conditional skewness in asset pricing tests. *Journal of Finance*, 55(3):1263–1295, 2000.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. In *ACM Transactions on Information Systems*, volume 20, pp. 422–446, 2002.

Eric Jondeau and Michael Rockinger. Optimal portfolio allocation under higher moments. *European Financial Management*, 12(1):29–55, 2006.

Bryan Kelly and Hao Jiang. Tail risk and asset prices. *Review of Financial Studies*, 27(10):2841–2871, 2014.

Maurice Kendall. A new measure of rank correlation. *Biometrika*, 30(1–2):81–93, 1938.

Alan Kraus and Robert H. Litzenberger. Skewness preference and the valuation of risk assets. *Journal of Finance*, 31(4):1085–1100, 1976.

Olivier Ledoit and Michael Wolf. Honey, i shrunk the sample covariance matrix. *Journal of Portfolio Management*, 30(4):110–119, 2004.

Yujia et al. Li. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022.

F. M. Longin. From value at risk to stress testing: The extreme value approach. *Journal of Banking & Finance*, 24(7):1097–1130, 2000.

Chris Lu and Cong et al. Lu. The ai scientist: Toward fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Ola Mahmoud. The temporal dimension of drawdown. *SSRN Working Paper 2546379*, 2015.

Harry Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952.

Attilio Meucci. Managing diversification. *Risk*, 22(5):74–79, 2009.

Alan Moreira and Tyler Muir. Volatility-managed portfolios. *Journal of Finance*, 72(4):1611–1644, 2017.

David Nawrocki. A brief history of downside risk measures. *Journal of Investing*, 8(3):9–25, 1999.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

George Philippatos and Charles Wilson. Entropy as a measure of diversification in investment portfolios. *Journal of Financial and Quantitative Analysis*, 7(1):1295–1301, 1972.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–42, 2000.

Bernardino et al. Romera-Paredes. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, 2024.

A. D. Roy. Safety first and the holding of assets. *Econometrica*, 20(3):431–449, 1952.

William F. Sharpe. Mutual fund performance. *Journal of Business*, pp. 119–138, 1966.

David Silver, Julian Schrittwieser, and Karen et al. Simonyan. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

Frank A. Sortino and Robert van der Meer. Performance measurement in a downside risk framework. *Journal of Portfolio Management*, 17(4):27–31, 1991.

Charles Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15(1):72–101, 1904.

Terry Young. Calmar ratio: A smoother tool. *Futures*, 20(8):40–41, 1991.

Kamer Ali Yuksel and Hassan Sawaf. Alphaportfolio: Discovery of portfolio optimization and allocation methods using llms. In *International Conference on Learning Representations*, 2025a.

Kamer Ali Yuksel and Hassan Sawaf. Alphaquant: Llm-driven automated robust feature engineering for quantitative finance. In *International Conference on Learning Representations*, 2025b.

Kamer Ali Yuksel and Hassan Sawaf. Alphasharpe: Llm-driven discovery of robust risk-adjusted metrics, 2025c. URL `https://arxiv.org/abs/2502.00029`.

Haijun Zhang, Shiqun Zhu, and Burton Sobel. Portfolio risk management with conditional expected drawdown. *Journal of Portfolio Management*, 37(1):37–44, 2010.