

AN EMPIRICAL EVALUATION OF ARTIFICIAL INTELLIGENCE AND
MACHINE LEARNING-BASED TRADING STRATEGIES
IN CRYPTOCURRENCY MARKETS

by

MOHAMMAD RAFI BANDUBAI

DISSERTATION

Presented to the Swiss School of Business and Management Geneva

In Partial Fulfilment

Of the Requirements

For the Degree

DOCTOR OF BUSINESS ADMINISTRATION

SWISS SCHOOL OF BUSINESS AND MANAGEMENT GENEVA

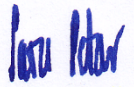
JUNE, 2026

AN EMPIRICAL EVALUATION OF ARTIFICIAL INTELLIGENCE AND
MACHINE LEARNING–BASED TRADING STRATEGIES
IN CRYPTOCURRENCY MARKETS

by

MOHAMMAD RAFI BANDUBAI

APPROVED BY



Prof.dr.sc. Saša Petar, Ph.D., Dissertation chair

RECEIVED/APPROVED BY:



Admissions Director

Acknowledgements

I would like to express my sincere gratitude to my dissertation supervisor, Dr. Apostolos Dasilas, for his guidance, constructive feedback, and academic direction throughout this doctoral research journey. His guidance helped me strengthen the research design, refine the literature positioning, and maintain the level of rigour expected of a doctoral dissertation.

I am grateful to the Swiss School of Business and Management Geneva for the academic structure, the research expectations, and the institutional support that made this thesis possible. This thesis also builds on a wider body of scholarship in financial machine learning, cryptocurrency markets, algorithmic trading, market microstructure, and empirical finance. I acknowledge the researchers whose work provided the theoretical and methodological foundation for this study.

I thank my parents, Allabakash and Emambee, for instilling in me a lifelong learning mindset.

I thank Ali Al Khalifa , Rasha M. Abu Alsaud , Hussein Bin Salem for giving me the opportunity to learn and excel in the professional world. Their leadership abilities and the attitude of ‘Moving Mountains’ in professional world have inspired me to pursue this herculean undertaking.

I am also grateful to the colleagues, peers, and mentors who encouraged me while this research was taking shape. Special thanks to HT and Chalam for their practical insights into real-world trading and financial technology.

Finally, I am especially thankful to my wife Rahee and my children Alisha, Akhila, and Kabir for their patience, encouragement, and understanding throughout this demanding journey. Their support gave me the motivation and discipline required to complete the research, analysis, writing, and revisions.

The responsibility for all interpretations, findings, conclusions, and any remaining errors in this thesis rests with me.

ABSTRACT

AN EMPIRICAL EVALUATION OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING–BASED TRADING STRATEGIES IN CRYPTOCURRENCY MARKETS

MOHAMMAD RAFI BANDUBAI
2026

Dissertation Chair: Dr. Apostolos Dasilas

This DBA thesis investigates whether artificial intelligence (AI) and machine learning (ML) based trading strategies provide statistically significant and economically meaningful benefits over traditional rule-based benchmarks in cryptocurrency markets. The study tests three hypotheses across four Binance perpetual futures contracts, BTCUSDT, ETHUSDT, SOLUSDT and XRPUSDT, using hourly data over a common walk-forward out-of-sample period from 2 April 2022 to 1 January 2026, with 15 non-overlapping folds per asset.

The empirical analysis compares AI/ML strategies with rule-based benchmarks on risk-adjusted return, evaluates whether a regime-aware LightGBM strategy reduces maximum drawdown relative to static strategies, and examines whether ensemble-based AI strategies improve Sharpe stability and cross-asset consistency over single-model approaches. The methodology uses train-validation-test walk-forward folds with embargo controls, train-only feature scaling, transaction-cost-aware backtesting, bootstrap confidence intervals, Wilcoxon and Welch tests, Monte Carlo

block bootstrap, cost-stress testing, leave-one-asset-out analysis, regime breakdown, and independent audit review.

The results do not support any of the three hypotheses on their primary criteria. AI/ML strategies do not achieve statistically significant higher mean Sharpe ratios than rule-based benchmarks. The pooled rule-based advantage is mainly driven by two exceptional bullish outlier folds; after removing them, performance is nearly identical. The regime-aware strategy underperforms the static benchmark, with mean maximum drawdown of 67.25% versus 37.58%. Ensemble strategies also fail the primary Sharpe-stability test, with the pooled median Sharpe delta favouring single-model strategies.

The findings are not entirely negative. AI/ML strategies record higher median Sharpe ratios across all four assets and perform consistently better in bearish regimes. Aggregation of ensembles reduces maximum drawdown for all four assets, indicating that these strategies may play an important role in preserving the capital when the Sharpe is not improved. In summary, AI/ML strategies do not show universal superiority within the tested four-asset, backtest-only crypto universe, but may add value under specific regimes and architectures. The thesis contributes an empirical, auditable framework for testing AI/ML trading claims under realistic and cost-aware conditions.

TABLE OF CONTENTS

LIST OF TABLES	xii
LIST OF FIGURES.....	xiii
LIST OF ABBREVIATIONS	xiv
CHAPTER I: INTRODUCTION	1
1.1 Background to the Study	1
1.2 Research Problem.....	2
1.3 Research Aim	3
1.4 Research Questions and Objectives	3
1.4.1 Research Questions	3
1.4.2 Research Objectives	3
1.5 Scope of the Study.....	4
1.6 Significance of the Study	5
1.7 Structure of the DBA thesis	6
CHAPTER II: REVIEW OF LITERATURE.....	7
2.1 Theoretical Framework	7
2.1.1 Efficient Market Hypothesis	7
2.1.2 Random Walk and Martingale Models	8
2.1.3 Adaptive Markets Hypothesis	8
2.1.4 Behavioural Finance Perspectives.....	9
2.1.5 Financial Machine Learning.....	10
2.1.6 Integration into the Present Study	10
2.2 Conceptual Framework	11
2.3 Cryptocurrency Market Characteristics.....	13
2.4 Efficient Markets and Cryptocurrency Inefficiencies	14

2.5 Algorithmic and Rule-Based Trading Strategies.....	16
2.6 Machine Learning in Financial Prediction	20
2.7 AI/ML Trading in Cryptocurrency Markets.....	23
2.8 Regime-Aware Trading.....	28
2.9 Ensemble Learning for Trading	31
2.10 Backtesting and Robustness in Financial Research	34
2.11. Hypothesis Development	38
2.11.1 Hypothesis H1: Risk-Adjusted Outperformance.....	38
2.11.2 Hypothesis H2: Regime-Aware Drawdown Control	38
2.11.3 Hypothesis H3: Ensemble Stability.....	39
2.11.4 Integrative Claim.....	39
2.11.5 Formal Hypothesis Statements and Mapping Table.....	39
2.12 Research Gap	40
2.13 Summary	41
CHAPTER III: METHODOLOGY.....	42
3.1 Research Philosophy and Design	42
3.2 Asset Universe and Data Sources	43
3.3 Data Preparation and Canonical Schema	44
3.4 Feature Engineering	44
3.5 Label and Signal Framework	48
3.6 Strategy Taxonomy	50
3.7 Backtesting Framework.....	53
3.8 Performance Metrics	55
3.9 Walk-Forward Validation	56
3.10 Statistical Testing.....	58

3.11 Robustness Testing.....	59
3.12 Auditability and Reproducibility	60
3.13 Methodological Limitations	60
3.14 Ethical Considerations.....	61
CHAPTER IV: RESULTS	62
4.1 Introduction	62
4.2 H1: AI/ML versus Rule-Based Strategies	63
4.2.1 H1 Statement and Objective.....	63
4.2.2 H1 Experimental Design	63
4.2.3 Asset-Level Results.....	63
4.2.4 Pooled H1 Results	65
4.2.5 Regime and Cost-Stress Breakdown	67
4.2.6 H1 Audit Observations	70
4.2.7 H1 Hypothesis Decision.....	70
4.3 H2: Regime-Aware versus Static Strategies	70
4.3.1 H2 Statement and Objective.....	70
4.3.2 H2 Experimental Design	71
4.3.3 Asset-Level Results.....	71
4.3.4 Pooled H2 Results	73
4.3.5 H2 Audit Observations	74
4.3.6 H2 Hypothesis Decision.....	75
4.4 H3: Ensemble versus Single-Model Strategies	75
4.4.1 H3 Statement and Objective.....	75
4.4.2 Ensemble Construction and Single-Model Comparators	75
4.4.3 Asset-Level Results.....	76

4.4.4 Pooled H3 Results	79
4.4.5 H3 Audit Observations	83
4.4.6 H3 Hypothesis Decision.....	83
4.5 Cross-Hypothesis Evidence Synthesis	84
4.5.1 Purpose	84
4.5.2 Evidence Hierarchy	84
4.5.3 Overall Verdict Matrix	85
4.5.4 Integrated Findings.....	86
4.5.5 Robustness Across the Three Hypotheses.....	86
4.5.6 Evidence Convergence and Conflicts	87
4.5.7 Final Integrated Research Finding	88
CHAPTER V: DISCUSSION	89
5.1 Interpretation of Results	89
5.2 Comparison with Literature	90
5.3 Why AI/ML May Not Have Outperformed.....	92
5.3.1 Sensitivity to Design Choices	96
5.3.2 Theoretical Implications of the Findings.....	98
5.4 Why Simpler Benchmarks Remain Hard to Beat.....	99
5.4.1 Empirical Frictions in Detail.....	99
5.5 Regime-Aware Strategy Discussion.....	102
5.6 Ensemble Strategy Discussion	102
5.7 Practical Deployment Challenges	103
5.7.1 The Signal-to-Cost Ratio as a Binding Constraint	104
5.7.2 Walk-Forward Performance Decay and Regime Transitions.....	105
5.7.3 Conditional Deployment as a Practical Resolution.....	106

5.8 Implications for Practitioners and Researchers	107
5.8.1 Implications for Practitioners	107
5.8.2 Implications for Researchers	107
5.9 Hypothesis Decisions	108
CHAPTER VI: SUMMARY, IMPLICATIONS, AND RECOMMENDATIONS	110
6.1 Summary of the Research	110
6.2 Answers to Research Questions	110
6.2.1 RQ1: Risk-Adjusted Outperformance	110
6.2.2 RQ2: Regime Performance	111
6.2.3 RQ3: Practical Implementation	111
6.3 Contributions of the Study	112
6.4 Study Limitations	113
6.4.1 Implications by Industry Stakeholder.....	115
6.5 Recommendations	117
6.5.1 For Practitioners	117
6.5.2 For Researchers	117
6.5.3 For Future System Development.....	118
6.6 Future Research.....	118
6.6.1 Rationale for Selected Future Research Items	119
6.7 Final Closing Statement	121
BIBLIOGRAPHY	123
APPENDIX A: FULL STRATEGY INVENTORY	129
APPENDIX B: DATA PIPELINE AND SCHEMA.....	131
APPENDIX C: FEATURE ENGINEERING DETAILS.....	132
APPENDIX D: BACKTEST ENGINE ASSUMPTIONS.....	133

APPENDIX E: H1 DETAILED RESULT TABLES.....	134
APPENDIX F: H2 DETAILED RESULT TABLES	135
APPENDIX G: H3 DETAILED RESULT TABLES	136
APPENDIX H: STATISTICAL TEST OUTPUTS	138
APPENDIX I: AUDIT REPORTS AND CLAIM GOVERNANCE.....	139
APPENDIX J: EXPERIMENT IDs, CONFIGS, AND MANIFESTS	140
APPENDIX K: LIMITATIONS AND METHODOLOGY DISCLOSURES.....	141
APPENDIX L: GLOSSARY OF KEY TERMS.....	142
APPENDIX M: HYPOTHESIS-ASSET-STRATEGY MAPPING	146

LIST OF TABLES

Table 2.1 Research Questions, Hypotheses and Primary Metrics	39
Table 3.1 Asset Universe, Data Periods and Walk-Forward Windows.....	43
Table 3.2 Canonical Data Schema.....	44
Table 3.3 Feature Categories and Counts per Asset	45
Table 3.4 Strategy Taxonomy Used Across H1, H2 and H3.....	50
Table 3.5 Backtest Execution Assumptions and Cost Scenarios	53
Table 3.6 Performance Metrics Definitions.....	55
Table 3.7 Statistical Tests and Robustness Checks Applied	59
Table 4.1 H1 Asset-Level Sharpe Comparison (Pooled by Asset).....	63
Table 4.2 H1 Pooled Statistical Evidence (4 assets, 840 strategy-fold observations).....	65
Table 4.3 H1 Regime Breakdown: Mean Sharpe by Market Regime	66
Table 4.4 H1 Cost-Stress Sensitivity Summary	67
Table 4.5 H2 Asset-Level Drawdown and Statistical Evidence	70
Table 4.6 H2 Pooled Regime, Cost-Stress and Monte Carlo Evidence.....	73
Table 4.7 H3 Ensemble vs Single-Model Sharpe and Drawdown by Asset.....	75
Table 4.8 H3 Strategy Ranking (Pooled Across 4 Assets)	78
Table 4.9 Cross-Hypothesis Verdict Matrix.....	84
Table 4.10 Integrated Findings Across H1, H2 and H3	85
Table 5.1 Practical Deployment Controls Implied by the Evidence.....	102
Table 5.2 Final Hypothesis Decision Table.....	108
Table 6.1 Key Study Limitations and Severity.....	112
Table 6.2 Future Research Agenda by Priority and Hypothesis.....	117

LIST OF FIGURES

Figure 2.1 Conceptual framework of the study	12
Figure 3.1 Feature Engineering Pipeline	47
Figure 3.2 ML_07_regime_lgbm Architecture.....	52
Figure 3.3 Walk-Forward Test Window Timeline	55
Figure 3.4 Walk-Forward Validation Structure.....	56
Figure 4.1 Research framework	61
Figure 4.2 H1: AI/ML versus rule-based Sharpe across the four assets.....	64
Figure 4.3 H1 bootstrap 95% confidence interval for mean Sharpe difference.....	65
Figure 4.4 H1 regime breakdown: mean Sharpe by market regime	67
Figure 4.5 H1 leave-one-asset-out sensitivity	68
Figure 4.6 H2: mean maximum drawdown by asset (regime-aware vs static benchmark).....	71
Figure 4.7 H3: ensemble versus single-model across assets	76
Figure 4.8 H3 strategy Sharpe ranking.....	79
Figure 4.9 H3 per-asset deltas	80
Figure 4.10 H3 ensemble vs single-model by asset.....	81
Figure 4.11 Evidence hierarchy used in the cross-hypothesis synthesis.....	83
Figure 4.12 Cross-hypothesis verdict matrix.....	84
Figure 4.13 Robustness summary: direction of each check by hypothesis.....	86

LIST OF ABBREVIATIONS

Abbreviation	Meaning
AI	Artificial Intelligence
ML	Machine Learning
BTC	Bitcoin (BTCUSDT perpetual futures)
ETH	Ethereum (ETHUSDT perpetual futures)
SOL	Solana (SOLUSDT perpetual futures)
XRP	Ripple (XRPUSDT perpetual futures)
OHLCV	Open, High, Low, Close, Volume
OOS	Out-of-Sample
WFA	Walk-Forward Analysis
CAGR	Compound Annual Growth Rate
Sharpe	Sharpe ratio (CAGR / annualised log-return volatility, in this study)
MDD	Maximum Drawdown
CI	Confidence Interval
PnL	Profit and Loss
RF	Random Forest
LGBM	Light Gradient Boosting Machine
SVM	Support Vector Machine
MLP	Multi-Layer Perceptron
LSTM	Long Short-Term Memory network
GRU	Gated Recurrent Unit network
MC	Monte Carlo
MTC	Multiple Testing Correction
EMH	Efficient Market Hypothesis
HMM	Hidden Markov Model

Abbreviation	Meaning
PSI/CSI	Population / Characteristic Stability Index
B&H	Buy-and-Hold
RB	Rule-Based strategy group
RA	Regime-Aware strategy group

CHAPTER I: INTRODUCTION

1.1 Background to the Study

During the past decade, cryptocurrency markets transformed from an experimental playground of Bitcoin as a peer-to-peer electronic cash system introduced by Nakamoto (2008), with the first block mined in January 2009, to a globally traded asset class worth multi-trillions of dollars with participants actively trading around-the-clock (Baur, Hong and Lee, 2018). Several unique characteristics distinguish cryptocurrency markets from the traditional markets of equity, fixed income and foreign exchange. These include relatively high price volatility, distinct price regimes of price movement, frequent fragility events such as exchange crashes, visible liquidity dislocations and fragmentation of venues among centralized perpetual futures exchanges, spot exchanges and decentralized liquidity pools. All these, together with round-the-clock trading and lack of market maker obligations on many venues, open a potential ground for systematic strategies, but also bring risks.

At the same time, algorithmic trading saw its revolution powered by the evolution of supervised machine learning (ML) techniques and, more recently, practical application of deep learning and large-scale ensemble methods. The common view shared by practitioners, academics, and increasing grey literature claims that it is possible to extract non-linear patterns from financial time series using AI/ML models such as tree-based ensemble, gradient boosting, recurrent neural networks, meta-learning architectures, etc. And the prospect of such models to generate profits by predicting the patterns in time series data appears especially promising in crypto markets, given the retail-driven sentiment, structural volatility and novelty of the asset class (Sebastião and Godinho, 2021).

However, the quantitative finance literature continues to emphasize the gap between backtest performance and live trading performance (Bailey, Borwein, López de Prado and Zhu, 2014). The list of reasons is long and consists of data snooping and multiple comparisons across big space of strategies, look-ahead biases arising from misalignment of features and labels, survivorship bias in asset selection, absence of cost modeling in the evaluation process, inappropriate use of random cross-validation for time series data, insufficient transparency and

details of model selection. These risks are well documented in the broader technical-trading literature (Park and Irwin, 2007) and are multiplied in cryptocurrency markets specifically. Estimation risk increases because of small historical samples, frequent regime events such as collapse of Terra/LUNA tokens in May 2022 and collapse of FTX in November 2022, and small number of liquid assets.

Against this background, the current thesis seeks to provide rigorous, reproducible and audited empirical proof of one single claim: whether AI/ML-based trading strategies, including regime-aware and ensemble architectures, provide quantitatively verifiable benefits over the traditional rule-based and benchmark trading strategies in cryptocurrency perpetual futures markets. The study represents empirical evaluation rather than a demonstration of a particular claim, and the key contribution here is the testing of specific, falsifiable hypotheses related to AI/ML performance under standardized and cost-aware multi-assets, out-of-sample conditions with presentation of findings, whether they are positive or negative.

1.2 Research Problem

The research problem can be stated as follows. While the industry discourse on cryptocurrency trading is heavily biased towards AI/ML techniques, the evidence provided in the academic and grey literature lacks rigor and robustness. It fails to meet all the criteria of a rigorous empirical financial study: evaluation across multiple assets, out-of-sample procedure free of look-ahead leakage, modelling of transaction costs and slippage, walk-forward validation instead of train-test splitting, robustness across various regimes and transaction costs, use of statistical inference rather than storytelling, and finally, audit of the whole pipeline from raw data to the conclusion. Such evidence is required in order to avoid deployment of strategies on the basis of potentially flawed in-sample or weakly validated backtests when the significant amount of capital is at stake.

Further, there are three questions embedded in the research problem. First, whether AI/ML strategies deliver superior performance against rule-based benchmarks in terms of risk-adjusted return when executed identically across several crypto assets? Second, whether regime conditioning, training and deployment of separate models for trending, ranging and high-volatility regimes improve drawdown control? Third, whether ensemble aggregation of multiple

AI/ML models improves cross-asset performance consistency compared to a single model? All these three directly translate into the three formal hypotheses stated in section 2.11.

1.3 Research Aim

The main aim of the research is to empirically test whether AI/ML-based trading strategies including regime-aware and ensemble variations, provide any statistically significant and economically meaningful benefits compared to traditional rule-based and benchmark approaches in cryptocurrency perpetual futures markets, under audited, cost-aware, multi-asset, walk-forward, out-of-sample conditions. A second aim, on the other hand, follows this one: to find the conditions, which market regimes, which architectural choices, which cost regimes, where any positive AI/ML evidence is robust, and where it does not survive under scrutiny.

1.4 Research Questions and Objectives

Three research questions drive the research design, each of them linked to a formal hypothesis and all of them are tested against the audited evidence pipeline generated by the research platform in Chapter III.

1.4.1 Research Questions

1. RQ1. Do AI- and ML-based trading strategies outperform traditional rule-based trading strategies in cryptocurrency markets on a risk-adjusted basis?
2. RQ2. How do different classes of AI and ML algorithms perform across varying market regimes, such as trending, ranging and high-volatility conditions?
3. RQ3. What are the practical limitations and implementation challenges of deploying AI-based trading strategies in real-world cryptocurrency trading?

1.4.2 Research Objectives

Three formal research objectives can be derived from the three research questions stated above in section 1.4.1. They are all empirical, measurable and directly traceable to the parts of the evidence chain generated by the research platform.

RO1, linked to RQ1, examines whether the portfolio of supervised AI and machine learning classifiers achieves a higher risk-adjusted return in comparison with the portfolio of eight rule-based benchmark strategies, under cost-aware, multi-asset, walk-forward, out-of-sample conditions. Risk-adjusted performance measure is the fold-level Sharpe ratio, calculated in line with the convention stated in section 3.8. The evidence is collected for each of four assets (BTCUSDT, ETHUSDT, SOLUSDT, XRPUSDT) and pooled across all the four assets. The objective is fulfilled when bootstrap, Welch and rank-order tests give consistent results and when the leave-one-asset-out checks indicate that no single asset is driving the result.

RO2, linked to RQ2, compares how different classes of AI and ML algorithms perform in different regimes, such as bullish, ranging and high-volatility regimes. Regimes are classified with feature-based classification into bull-trend, ranging, bear-trend groups. The objective is met when the regime-aware strategy shows lower mean maximum drawdown in the pooled multi-asset analysis and at least in three out of four assets being tested.

RO3, linked to RQ3, examines if ensemble AI trading strategies demonstrate better performance compared to single-model AI strategies on the consistency and stability of fold-level Sharpe ratios across the four assets. Ensembles include majority and soft voting, weighted voting, stacking out-of-fold, strategy-signal aggregation and rank-based aggregation. The objective is met if the ensembles demonstrate higher pooled median Sharpe and directional improvement across the four assets. An additional deployment-related objective is whether the ensembles reduce the maximum drawdown for the four assets.

There is one fourth, holistic objective, which is associated with the first three objectives, and it is testing whether the overall claim that AI and ML trading strategies generally improve risk-adjusted performance, robustness and consistency in cryptocurrency markets hold up after all three hypotheses are tested and put together into an audited cross-hypothesis synthesis. None of the four objectives makes any assumptions regarding the result direction.

1.5 Scope of the Study

The empirical scope is intentionally limited in order to allow making statistical claims in tightly controlled conditions. The universe of assets considered in the study consists of four major USDT perpetual futures contracts available on Binance Futures exchange: BTCUSDT,

ETHUSDT, SOLUSDT and XRPUSDT. Data is available at one-hour bar frequency sufficient to generate walk-forward folds. The data is downloaded from 1 January 2021 to 1 March 2026. The common out-of-sample period across all the three hypotheses is from 2 April 2022 to 1 January 2026. Such start date allows for at least twelve months of training and three months of validation in the first walk-forward fold.

The universe of strategies includes eight rule-based benchmark families, experiment-specific AI/ML models and their ensembles. Clarification: since the H1/H2 and H3 experiment runners use partially different namespaces of strategies, the numeric identifiers of the strategies cannot be combined into a single one. The H1/H2 AI/ML set comprises logistic regression, random forest, gradient boosting, support vector machine, GRU recurrent modelling, supplemental ensemble and regime-aware LightGBM implementation. The H3 set consists of six single-model competitors, six primary ensembles, one secondary regime-aware ensemble and two failed candidates. Execution and risk management are the second tier concerns in the thesis and no live trading or order book simulation is conducted. Perpetual futures funding rate cost is not included in the backtest engine and is mentioned in the thesis.

1.6 Significance of the Study

There are several ways how the thesis contributes to the academic and practitioner literature. First, it compares the performance of three different families of AI/ML strategies, namely, single-model, regime-aware and ensemble, against a common set of static rule-based benchmarks under identical conditions of execution and risk management rather than demonstrating one single strategy on case-by-case basis. Second, it employs multi-asset universe of four major cryptocurrencies with common out-of-sample period, thus providing cross-asset generalizability and leave-one asset out sensitivity analysis. Third, it integrates walk-forward analysis, Monte Carlo block bootstrap, six-scenario cost stress testing and regime breakdown into one single robustness framework. Fourth, it develops an end-to-end audit process from raw exchange data to manifest experiments to audit report on the claim level and establishes the reproducibility standards in order to overcome the problem of replication in the quantitative finance literature. Finally, by reporting null findings openly as valid academic results, it addresses the literature that may be biased toward positive AI/ML claims.

1.7 Structure of the DBA thesis

The structure of the thesis proceeds as follows. Chapter II reviews the literature and outlines the characteristics of cryptocurrency markets, market efficiency, algorithmic and rule-based trading, machine learning in finance, regime-aware trading, ensemble learning and backtesting and robustness practices and identifies the research gap. Chapter III describes the methodology, covering the asset universe, data processing, feature engineering, label and signals frameworks, strategy taxonomy, backtest engine, performance metrics, walk-forward validation, statistical test, robustness framework and auditability standards. Chapter IV presents the results, with 4.2 on H1 (AI/ML versus rule-based), 4.3 on H2 (regime-aware versus static), 4.4 on H3 (ensemble versus single-model), and 4.5 on cross-hypothesis synthesis. Chapter V discusses the findings in the context of the literature, examines the reasons for the failure of the AI/ML strategies in the scope of the thesis and covers the implementation issues. Chapter VI summarizes the thesis, answers the research questions posed, states the hypotheses decision, provides the contributions and limitations of the study, and formulates the recommendations and directions for future research.

CHAPTER II: REVIEW OF LITERATURE

2.1 Theoretical Framework

Any empirical analysis of trading effectiveness involves making theoretical assumptions regarding how markets discount information, investor behavior and how predictable future returns can be based on historical data. The theories on which the present research is based are discussed in this section. There are five major theories which are applicable to this particular study, the efficient market hypothesis (EMH), random walk and martingale approach, adaptive markets hypothesis (AMH), behavioral finance and finally the emerging field of machine learning in finance as championed by López de Prado. These theories provide the basis for both the pros and cons of systematic outperformance by AI and machine learning.

2.1.1 Efficient Market Hypothesis

The Efficient Market Hypothesis was formulated by Fama (1970) as a hierarchy of claims about which information prices have already accounted for. Weak-form EMH states that there is no edge from using historical price and volume data once a normal return for risk is subtracted. Semi-strong form EMH extends this claim to all publicly available information, while the strong form adds non-public information to that list as well. The practical consequence for systematic trading is that any strategy generating durable excess returns must be doing at least one of three things: bearing a risk that the market prices but the backtest does not capture, relying on a research process such as overfitting or survivorship bias, or harvesting an actual, but temporary inefficiency in the market. In the case of cryptocurrencies, the empirical picture is truly mixed. The major assets seem to converge to the weak-form EMH due to rising institutional participation, but persistent anomalies emerge when regime shifts happen, during times of stress events, and in small cap tokens. For current purposes, the weak-form version of the theory serves as the null: AI and ML strategies have to outperform it, accounting for costs.

However, according to Urquhart (2016), who examines the weak-form efficiency of Bitcoin returns, there is strong evidence of inefficiency for the entire period, although subsequent sub-periods exhibit increasing levels of efficiency in line with the conditional approach to EMH adopted by this thesis.

The opposite result is found when Nadarajah and Chu (2017) re-analyse the same data set using power-transformed returns. This example underlines just how crucial it is to choose an appropriate test when deciding whether a market is efficient, which is why the thesis assesses AI/ML methods in light of a conditional null.

Finally, Bariviera (2017) estimates the evolution of the Hurst exponent for Bitcoin returns and finds a dramatic drop in long-range dependence after 2014 due to increased market depth.

2.1.2 Random Walk and Martingale Models

The Random Walk Hypothesis makes precise the notion of weak-form efficiency in statistical form: If the price increments are serially uncorrelated, then no pattern-recognition process on historical prices can produce a statistically significant advantage over time. The strict version of the hypothesis rules out autocorrelation at every lag, while the softer versions allow predictability in the short term but still deny predictability once transaction costs are taken into account. Cryptocurrency results are somewhere in between. While the strict random walk at intraday frequencies has been thoroughly rejected, the predictability is small enough that transaction costs will absorb it and so is what we try to test with this technique. The use of block bootstrap in the robustness analysis is also a product of this literature since the hourly returns remain serially correlated at short lag times, so resampling the individual observations would be an underestimation of how different the sampling is.

Samuelson (1965) developed the theory of martingale asset prices. His idea is that competitive markets will price in all available information at each moment, and only return a fair price for risk, and that is the minimum that needs to be demonstrated before any positive AI/ML finding is interpretable.

Fama (1965) operationalises the random walk model for stock prices by using serial correlation and runs tests to show that price changes over time are essentially independent. That null hypothesis is what the walk-forward, cost-aware tests in Chapter III are designed to challenge under realistic conditions.

2.1.3 Adaptive Markets Hypothesis

Lo (2004) formulated the Adaptive Markets Hypothesis, a model that integrates the efficient markets hypothesis with the concepts of evolutionary biology and behavioral science. The AMH treats efficiency not as a fixed property but as a time-varying outcome of an ecosystem of competing strategies. Predictability emerges and vanishes as participants come, learn and leave and depending on how macroeconomic conditions alter which strategy groups are better suited to the market state. This model is particularly appropriate for cryptocurrency markets due to the rapid changes in the participant profile, regulation and product in the last decade. According to the AMH, a good performance of an algorithmic or machine learning strategy under one regime does not mean that it will be equally successful in the other, the point in the market cycle at which a strategy is deployed matters. The focus on regime break-downs, walk-forward training and conditional sub-findings in this thesis is consistent with the AMH approach, where every positive finding should be interpreted within its particular regime and period.

In Lo (2005), the biological analogy underlying the AMH is discussed, where competition for limited resources among different strategies in finance is similar to competition between different species in an ecological system. The profitability of such strategies rises and falls depending on changes in circumstances, and survival depends on adaptation rather than any permanent edge. That framing directly informs the regime-breakdown analysis in section 4.2.5.

Urquhart and Hudson (2013) test the AMH across five international equity markets using rolling-window predictability tests and find that predictability fluctuates over time in ways consistent with the AMH but not with a fixed-efficiency view, providing empirical support for the time-varying conditional framing adopted here.

2.1.4 Behavioural Finance Perspectives

While efficient-markets models assume that investors are rational and process all available information, behavioral finance rests on the observation that it is often not the case. Each of the loss aversion (Kahneman and Tversky, 1979), overconfidence, and herding mechanisms leaves its mark on the distributions of returns. Crypto asset classes, due to their predominance of retail investors, show it more frequently than mature equity asset classes do. Price events driven by social media and sentiment cascades, the kind of narrative contagion Shiller (2017) documents in his account of how stories move markets, are structural features of the crypto landscape and not just a form of noise. While none of the behavioral mechanisms is directly modeled in the features

set, the conditional bear trend superiority of AI/ML models found in section 4.2.5 is likely to be an outcome of all of them. Distressed retail selling in down-trending markets may create the short-horizon directional signal that classifiers pick up but that simple trend-following rules miss. The robustness approach also stems from this logic. Results based on one specific behavior episode would not pass either leave-one-asset-out or leave-one-fold-out test.

De Long et al. (1990) formalize the noise trader model, proving that irrational traders can exist in equilibrium and drive asset prices away from their fundamentals, hence giving rise to the directional patterns over short horizons that AI/ML methods try to find, and which the bear-trend finding of section 4.2.5 is consistent with.

Barberis and Thaler (2003) provide a survey of the whole field of behavioural finance, pointing to overconfidence, conservatism and loss aversion as the reasons for return anomalies. This theoretical account gives a basis for viewing the conditional AI/ML bear trend as resulting from the behaviour of distressed retail investors rather than a general regime-independent informational edge.

2.1.5 Financial Machine Learning

Financial Machine Learning, defined by López de Prado (2018), is the methodological theory that comes nearest to the empirical approach used in this study. It starts from the assumption that backtesting performance suffers from selection bias, look-ahead bias, multiple hypothesis testing and non-stationarity and provides solutions for each case: strict walk-forward validation with embargo periods between training and test windows; purged k-fold cross-validation to block label leakage; explicit modelling of transaction costs, slippage, and turnover; deflated Sharpe ratios that adjust for how many strategies were tested; combinatorial purged cross-validation to estimate the distribution of backtest performance; and probabilistic Sharpe ratio and probability of backtest overfitting (PBO) statistics. The present research implements several of these controls, walk-forward validation with embargo, train-only feature scaling, manifest-tracked configurations, audit reports, and bootstrap inference. The rest, deflated Sharpe and PBO especially, are identified as future-research priorities in Chapter VI.

2.1.6 Integration into the Present Study

Putting the pieces together, the position adopted here is the following. Cryptocurrency markets are not perfectly efficient, but at the hourly frequency tested here they are efficient enough that any AI or ML claim of outperformance bears the burden of evidence. Predictability, where there is any, is conditional – linked to regimes or structural events rather than being present in all market states. Behavioural and adaptive forces keep that predictability moving, which is why single-period, single-asset backtest claims are insufficient. Methodology of financial machine learning provides the framework of credible evidence. This defines the design of the empirical exercise in Chapter III, and it is the standard against which the hypothesis verdicts in Chapter IV and V get measured.

2.2 Conceptual Framework

Figure 2.1 visualises the conceptual framework that links the theoretical foundations reviewed in section 2.1 to the three research questions, the three hypotheses, the empirical design, and the audited verdicts that constitute the contribution of the study. The framework is organised as a five-layer hierarchy from theory to outcome.

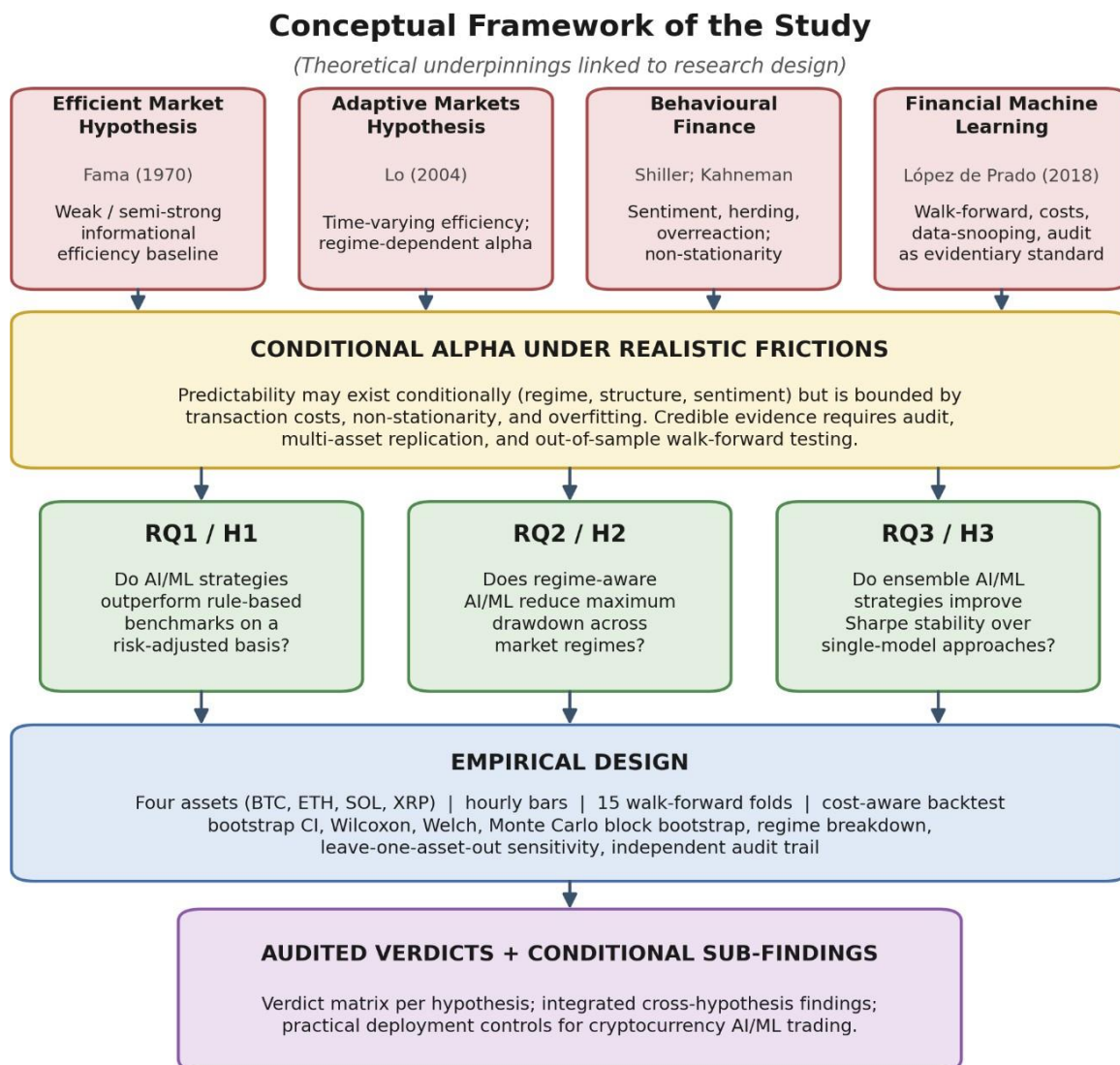


Figure 2.1 Conceptual framework of the study, linking the four theoretical foundations to the three research questions, the empirical design, and the audited outcomes.

Layer one comprises four theoretical foundations: the Efficient Market Hypothesis, the Adaptive Markets Hypothesis, behavioural finance, and financial machine learning. Each of them offers a distinctive view of how AI and ML strategies may or may not work in the face of the existence of easier alternatives and what evidence will have to be shown to justify this position. EMH is the null hypothesis. AMH is the conditional and time-varying view. Behavioural finance is the

means by which short-horizon predictability can persist. And financial machine learning is the methodological paradigm.

All of this converges at layer two, resulting in the assumption that any tradeable alpha will likely be conditional, time-sensitive and frictional. Therefrom follow the answers to three distinct empirical questions (RQ1, RQ2, and RQ3) of layer three. Next, each of these questions can be formulated into an empirical hypothesis (H1, H2, and H3) and tested at layer four.

Layer four is the empirical design. Its ingredients are identical for all three hypotheses: the four Binance perpetual futures contracts, the hourly frequency, fifteen walk-forward folds with embargo, and the backtest engine. This layer introduces statistical inference tools, robustness checks and auditing. Layer five is the empirical output: verdict matrices by hypotheses, cross-hypotheses findings, and a set of practical deployment controls. Moreover, the theory model is circular in nature, and once the data has been gathered, it should be used to re-evaluate the conditional beliefs in the second layer and finally the theoretical assumptions in the first layer.

2.3 Cryptocurrency Market Characteristics

Cryptocurrency markets trade continuously and are globally accessible, and they have drawn growing empirical study since Bitcoin appeared in 2008 (Nakamoto, 2008). Five characteristics appear consistently. First, realised return volatility is materially higher than in mature markets, historically three to five times the realised annualised volatility of equivalent equity benchmarks for the majors. Second, trading is continuous, so there are no overnight gaps and an intraday regime shift can land at any local hour. Third, the venue landscape is fragmented across centralised exchanges, perpetual futures venues, and decentralised liquidity pools, each differing materially in funding rates, fee structures, and liquidity. Fourth, liquidity concentrates in certain instruments such as BTC and ETH, whereas the order book liquidity and bid-ask spread become extremely thin in other assets. Fifth, the large, active retail base now coexists with the increasing institutional presence in the market, creating a situation where some of the traditional arbitrage inefficiencies disappeared while new ones appeared, with exchange-level credit risk being one example, following the FTX failure in November 2022.

The other noteworthy characteristic is that of non-stationarity of return distributions. Crypto price processes show time-varying volatility clustering, regime shifts driven by macroeconomic

events and changes in protocols, and tail events poorly captured by the parametric distributions. The implication for strategy evaluation is straightforward: any methodology based on the stationary distribution assumption or calibration of trading rule parameters based on the entire sample will be flawed due to overestimation of potential performance.

Baur, Hong and Lee (2018) examine Bitcoin's role as the store of value, hedge or speculative asset and find it serves primarily as a speculative investment with very low correlation with traditional asset classes. This trait reduces the scope of generalizability of conclusions from the research in equity ML domain.

Gandal, Hamrick, Moore and Oberman (2018) detect evidence of coordinated price manipulation in the Mt Gox exchange during the 2013 Bitcoin price spike. Their findings illustrate the structural weakness of the early crypto markets and justify the choice of an evaluation period spanning multiple regimes, as opposed to only one advantageous one.

Corbet, Meegan, Larkin, Lucey and Yarovaya (2018) detect that cryptocurrency markets are informationally separated from traditional equity, bond and currency markets. This finding supports the treatment of cryptocurrency as a standalone asset class and warns about the dangers of applying the assumptions from equity ML research to the analysis of the four perpetual futures contracts in the experiment.

2.4 Efficient Markets and Cryptocurrency Inefficiencies

The Efficient Market Hypothesis, weak, semi-strong, and strong, is the canonical baseline that has to be taken as a benchmark for any claim of systematic outperformance. According to the weak form of EMH, past prices and volume do not carry any exploitable information except fair return for risk. The empirical literature on cryptocurrency efficiency is mixed: some researchers detected momentum and short-term reversal anomalies in the major assets, BTC and ETH, while others noted that those assets moved towards weak-form efficiency as the institutions joined the market. Three studies are especially relevant for this discussion. Urquhart (2016) detects strong evidence of Bitcoin inefficiency while noting the improvement in some of the later samples. Nadarajah and Chu (2017) perform power transformation of the same data and come up with the opposite conclusion. The Bariviera (2017) provides the time-series approach to this question,

detecting that Bitcoin long-range dependence significantly declined after 2014 as the institutional participation grew.

The crypto literature discovered several concrete anomalies: momentum at weekly-to-monthly horizons, mean-reversion at very short ones, calendar effects weaker than in equities (due to 24/7 trading), and event-driven effects around protocol upgrades, exchange listings, and regulatory events. Such anomalies provide the theoretical basis for systematic trading research. The same literature mentions the caveats that have to be kept in mind: they tend to be relatively small compared to the transaction costs, change over time, and decay as the exploitation of the anomaly grows. Thus, any AI/ML strategy harvesting such anomalies has to be evaluated with realistic transaction and turnover costs assumptions, not idealized ones.

There are five additional features of the crypto markets worth mentioning, since each of them influences the design and interpretation of the research results. First, venue landscape is highly fragmented: the same asset, BTCUSDT for instance, is traded on dozens of centralized exchanges, each with different fees, funding regimes, and order-book liquidity, and on decentralized exchanges operating according to automated market makers principle rather than centralized limit-order books. Cross-venue arbitrage is possible, but the latency in withdrawal, custody risks, and the cost of cross-venue trade execution limit its scale. For the sake of consistency, the current study evaluates the strategies in one venue, Binance Futures, while extending the analysis to cross-venues is left for Chapter VI.

Second, the instrument class used in the study, perpetual futures, does not expire. Instead, it is aligned with the spot asset price through funding rate paid by longs to shorts or vice versa depending on whether the regime is bullish or bearish. Over the four-year window the level of funding rates changed significantly depending on the regime reaching 20% and more annually at some points. The backtest engine does not take funding rates into account, this material limitation is highlighted in section 3.13 and summarized in Chapter VI. The impact of this oversight depends on the turnover and regime: a trend follower with long positions in a bull regime will incur higher costs than shown in the backtest, while a strategy with predominantly short positions in a bear regime will incur lower ones.

Third, the composition of institutional and retail participants changed over the window. Retail sentiment-driven, 2021 alt-coin bull cycle, Terra-LUNA collapse in May 2022, FTX failure in

November 2022, 2024 Bitcoin spot ETF approval, Solana memecoin episodes in 2024-2025, and many other similar periods show how rapidly the structure changes. From the empirical finance perspective these are non-stationarities that violate the assumption implicitly used in most backtests that data-generating process does not change from the training to testing window. The walk-forward validation mitigates this problem but does not solve it.

Fourth, the regulatory environment matters: it is uneven across jurisdictions and has changed significantly over the evaluation window. Spot ETF approval of several tokens in the United States in January 2024, European Union's Markets in Crypto-Assets (MiCA) regime becoming effective in 2024, and ongoing SEC enforcement actions against certain tokens affect the cross-asset performance in the results. The clearest example is XRP: the regulatory resolution episode in late 2024 drives the outlier fold reported in section 4.2.5. No attempt is made to adjust for the regulatory events: their influence is absorbed into return distribution.

Fifth, on-chain settlement and the transparency of the underlying network introduce a special category of features and risks that are not present in traditional financial markets. On-chain metrics like active addresses count, transaction throughput, exchange inflow/outflow, behavior of miners or validators, stablecoins flows can carry predictive signal for short-term returns. The on-chain features are deliberately excluded from the main experiments: maintaining the internal consistency across four assets is not easy if three of them (ETH, SOL and XRP) operate on different consensus and settlement architectures that do not have comparable on-chain metrics. The inclusion of the on-chain features into the feature set is one of the high-priority future directions.

2.5 Algorithmic and Rule-Based Trading Strategies

Rule-based strategies are included in this research for a specific purpose: they serve as the practical baseline against which any claim of AI/ML superiority has to be measured. A more complicated model is only worth its engineering cost if it beats something that can be implemented by the trading desk with a spreadsheet and a brokerage API. These eight families cover the major categories of systematic discretion (passive holding, trend-following, mean-reversion, breakout and momentum), and they are transparent enough for the study of their regime-dependent weaknesses. Together they provide a challenging benchmark.

Brock, Lakonishok and LeBaron (1992) produce the pioneering work providing the empirical evidence that moving average and trading range breakout rules generate positive returns in the DJIA over nearly a century of data before transaction costs erode much of the advantage. This work establishes the rule-based families used in the current research as historically grounded benchmarks.

Park and Irwin (2007) survey the technical trading literature and find that although most papers claim positive results, data snooping and ignoring costs inflate many of those results. The walk-forward and cost-conscious design of the study is the direct response to the limitations they identify.

Hurst, Ooi and Pedersen (2017) document consistent trend-following returns across asset classes over a century and find that time-series momentum strategies deliver steady risk-adjusted performance in diverse market conditions, providing long-run context for interpreting the buy-and-hold and moving-average crossover results observed across the four crypto assets.

The eight families divide naturally into three groups by their underlying logic. The first is passive exposure: RB_01 buy-and-hold, which makes no directional bets and instead just collects whatever the market delivers. The second group is trend-following: RB_02 moving-average crossover, RB_06 MACD and RB_08 EMA trend filter all attempt to identify and ride directional price moves, differing mainly in how they confirm that a trend is in place before committing. The third group is reversion and breakout: RB_03 RSI and RB_04 Bollinger Band bet on a return toward the mean after extreme price moves, while RB_05 Donchian and RB_07 volatility breakout bet on continuation following a breakout from a range. This means that the presence of all three of these groups in the benchmark pool ensures that the analysis is not biased towards any of the comparators.

The eight families have their own preference for a market environment, but there is no family that always does well under any market environment of its choice. This conditional behavior pattern is the reason why these families can be used as benchmarks, making the issue of whether an AI/ML strategy performs better than the best possible strategy much clearer. Trend-based families struggle when market swings sideways; mean-reversion families struggle if momentum stays up for longer than the signal window expects; breakout families are susceptible to failed

range expansions. The AI/ML group is in the same market state and has to work with the high turnover and estimation risk of learning from a finite training window.

What follows is a description of each of the eight rule-based families, and the families' behaviour in the hourly perpetual futures setting rather than their generic structure. Each family is tested on a grid of parameter configurations, and at each walk-forward fold, the configuration with the best Sharpe in the validation window is taken forward to the out-of-sample test window. That separation is critical, since any parameter that sees the test window even indirectly inflates reported performance, so the grid search is limited to the validation part of the grid.

Buy-and-hold (RB_01) opens a long at the start of the test window and keeps it open until the end of the test window, accumulating whatever returns the market gives, in that time. There are no free parameters, so there is no risk from parameter selection, but also no protection from losses, since every drawdown is absorbed in full. In the four crypto assets tested here that unconditional long bias made RB_01 an unusually good comparison during bull-trend folds, as it merely rode large directional moves that an active strategy could only beat through timing rather than exposure.

Moving-average crossover (RB_02) enters long when a short-window average is higher than a longer-window average, suggesting that the price action is gaining momentum relative to the trend, and short on the reverse signal. We test five parameter sets, fast windows of 10, 20 and 50 bars and slow windows of 50, 100 and 200 bars. At hourly resolutions of crypto, the faster configurations respond fast but are whipsaw-like in choppy times; the slower ones are slow to entry and exit but stay long enough to absorb the transaction cost. That trade-off between reactivity and noise tolerance is why the best configuration is the one that wins the fifteen review folds.

The RSI mean-reversion (RB_03) will bet that extreme momentum readings are unsustainable and will unwind. Eighteen configurations are tested in the sample, with overbought/oversold threshold pairs of 25/75, 30/70 and 35/65. Tighter thresholds fire signals less often but on more extreme markets; looser thresholds trade more frequently on weaker reversals. In the crypto context, with trend persistence common in bull and bear markets, mean-reversion signals can stay on the wrong side of the market for long periods of time and that structural weakness is revealed in the walk-forward results.

Bollinger Band mean-reversion (RB_04) is based on a similar logic to RB_03 but the signal is seen in the volatility-relative terms, not momentum oscillator terms. The signal fires when price has moved out of the rolling mean, measured in standard deviations (not percentages). We test four different configurations, with band periods of 20 and 50 bars and multipliers of 2.0 and 2.5 standard deviations. The multiplier determines how far price must drift before the strategy considers a snap-back likely; narrower bands fire more often in normal conditions, while wider bands only in real extremes. The bands adjust automatically as the realised volatility changes and so they respond to changes in market turbulence and do not respond to fixed thresholds of momentum oscillators.

Donchian breakout (RB_05) treats a break above the N-bar high as evidence that buying pressure has overwhelmed supply and goes long; a break below the N-bar low triggers the mirror-image short. We test three different lookback windows of 20, 50 and 100 bars. The shorter window responds to recent range breaks, which is ideal for fast-moving crypto markets but leads to more false entries if price is simply oscillating. The longer window requires a breakout to be more decisive and we miss some real moves but avoid many traps. At hourly resolution, a 100-bar Donchian covers roughly four days of price history, so it is a medium-frequency trend entry mechanism rather than a high-frequency signal.

MACD trend (RB_06) trades on the spread between a 12-period fast EMA and 26-period slow EMA, using a 9-period signal line to smooth the crossover. These are standard parameters that have been in use for decades in retail and professional charting tools, and they are used here without modification because they are widely used and they are a good representational benchmark: any trader who uses a MACD-based system is likely doing something similar in this case. The strategy is directional in nature and can struggle in flat or choppy markets where the MACD line crosses the signal line all the time with no follow-up.

Volatility breakout (RB_07) treats a large rise in realised range or true-range volatility as the sign that a direction is moving and goes in the direction of the first significant bar after the growth. We test six configurations, with lookback periods of 20 and 50 bars and a threshold of 1.0 to 2.5 standard deviations. Low multipliers fire on mild volatility spikes and generate high turnover rates; high multipliers wait until larger expansions are achieved and do not observe as much movement. In crypto, volatility spikes are often high enough that a high-multiplier

configuration can still produce signal frequency, but low multipliers trade so much that the cost drag dominates the gross return.

The exponential-moving-average trend filter (RB_08) layers a directional filter over a momentum signal: the fast EMA indicates that the price is accelerating, and the slow EMA indicates that the trend is coming from the broader market during trade. Two configuration pairs are tested: 20/100 and 50/200. The dual confirmation scheme reduces the number of signals when crossing a single EMA crossover, which leads to fewer turnover and costs. The trade-off is that, because both conditions must hold simultaneously, RB_08 enters into a trend later than RB_02 does, which takes some early-move gain out of the picture but is at the cost of fewer false entries when trading in non-standard flows.

The frequency of trade is quite different in the eight families. At one end, RB_01 makes a single trade per fold; at the other end, RB_07 makes about 200 trades per fold. The rule-based group makes around 85 trades per fold compared to about 272 for the AI/ML strategies. This three-to-one turnover ratio is important to the H1 comparison: since commission and slippage come into play on every round trip, a strategy with three times the number of trades needs a gross edge to get the same net return. It is one of the structural reasons why rule-based benchmarks are more difficult to beat than they might appear on a gross-return basis.

2.6 Machine Learning in Financial Prediction

Machine learning is now increasingly prominent in financial prediction research, especially since the mid-2010s. The appeal is simple: It can handle non-linear relationships, large feature sets and pick up patterns that traditional statistical models miss. In finance it has been used to model equity returns, foreign exchange, commodities, derivatives, and, more recently, cryptocurrency assets.

Breiman (2001) developed an algorithm for random forests and showed that unifying many decorrelated decision trees greatly reduces the variance of the predictions without bias. This theoretical guarantee is why ML_02 random forest is still one of the best single-model results in H1 and H3 and is the basis for ensemble variance reduction in H3.

Chen and Guestrin (2016) developed a scalable gradient-boosted tree system that is used in structured financial data sets. Its presence in the experiments as both an active H1/H2 strategy

and a failed H3 implementation of them illustrates a very relevant point: good general-purpose performance on financial tabular data does not mean that production behaviour is guaranteed in every configuration and hyperparameter setting.

Hochreiter and Schmidhuber (1997) propose the Long Short-Term Memory network to overcome the vanishing gradient problem in recurrent architectures. The LSTM is included in the H3 model pool as ML_09 but fails to generate any trades on all folds and assets. It is a direct reminder that architectures with good theoretical properties for sequential data are not expected to generate any signals on noisy hourly crypto price series.

Pedregosa et al. (2011) present scikit-learn, a Python-based machine learning library for all the supervised classifiers in their software. Writing it here is part of the reproducibility discipline: a future researcher can verify the hyperparameter grids, scaling procedures and cross-validation logic for the same software environment.

The literature on financial ML has four main modelling approaches that deal with various types of models and bet on what kind of output will be most useful for trading. Directional classifiers sort each bar into a few outcome categories and fire a signal based on which category wins, and this is what all ML strategies here produce. Return-magnitude regressors try to estimate how big the next move will be, which in principle allows more nuanced position sizing but in practice adds estimation noise in low-signal environments. Asset-ranking approaches compare securities against each other and trade the spread, which works well in large cross-sections but is less suited to a small universe of four correlated perpetual futures. And reinforcement learners treat the trading account as an environment and the strategy as an agent that discovers rules through reward feedback, a promising direction but one that introduces non-stationarity into the learning process itself. The choice of directional classification is deliberate and it leaves the comparison between these and the rule-based benchmarks clean: the two groups give the same kind of output, a long, short or no-trade signal over each bar.

Within that classification framework, the label is a three-class directional outcome over a four-bar forward horizon, with a 15-basis point activation threshold for the minimum move that is a tradeable signal. We calibrate the threshold to the round-trip transaction cost of the base case so that a signal only fires when the expected gross edge at least covers the cost of acting on it.

A forward return above the positive threshold is classified as positive, and one below the negative threshold as negative. The rationale behind this approach is very straightforward: a trading model must not trigger a trade unless the expected move is significant enough to cover the transaction costs. It is important to note that no-trade is not a residual category but rather a deliberate trading decision in itself.

There are four groups of trading models used in the experiment. Logistic regression is the low-complexity anchor; it is the floor that no more sophisticated approach should fall under and, if it outperforms more flexible models, that ultimately holds true about signal quality. The tree gradient boosting and random forests have the middle ground, dealing with the types of non-linear feature interactions that are characteristic of momentum and volatility signals without the bulk of training data that neural networks need to generalise to. The SVM provides a different geometry to the decision boundary, which looks for the greatest margin separation in the feature space and not growing recursive partitions. And at the flexible end, the MLP trades interpretability for capacity, which in noisy hourly crypto data has been found to have higher variance in fold-level performance than actual better expected returns.

Recently, attention-based deep learning architectures have developed across a variety of financial forecasting domains. Transformer models are designed to capture long-range temporal dependencies more accurately than recurrent neural networks. They have shown competitive performance for a number of multivariate time-series forecasting tasks (Lim et al., 2021; Wen et al., 2023). However, it is increasingly clear that architecture alone cannot guarantee better trading performance. The improvements in predictive accuracy are usually diluted when realistic transaction costs, walk-forward validation, and strict control of data leakages are taken into account. With this in mind, recent financial machine learning research has been increasingly focused on testing and developing the models to make them more accurate, and this is the approach we take when designing the empirical approach.

Three failure modes recur across the ML trading literature, and all three are specifically guarded against in the methodology here. The first one is that the training set is contaminated by information that would not be present at the bar that is predicted, for example, a specific feature computed on future prices or a label whose horizon is shorter than the next training period. The second is that the validation procedure is borrowed from i.i.d. settings where folds are randomly

shuffled are acceptable, but in a time series they generate training observations from the future of what is nominally a test observation. The third is the classic overfitting story: a 35-to-39 feature matrix trained on 12 months of hourly data contains far more degrees of freedom than the true signal, so any classifier that is not regularised or that over-searches hyperparameters will find in-sample patterns that dissolve as time goes on.

A fourth failure mode runs at the level of the research process rather than the individual model: if the strategy variants are tested on the same historical data, at least some will appear profitable by chance alone, and the one that gets picked is likely to be the one that looked best. Walk-forward validation with pre-specified hypotheses is the key defense to this. The three hypotheses were fixed before the experiments ran, and the verdict criteria were specified in the experiment configuration files before a single fold was evaluated. However, such discipline does not make one totally free from the risk of data snooping, given that the feature set and model families had to be chosen at some point, but it restricts the freedom of choice after the data has been looked at.

Chapter III discusses the solution to this problem through the walk-forward validation method. Each model is trained on historical data and tuned on a different validation window, then evaluated on a different test window. Feature scaling is fitted for the training data and then applied to the validation and test data without refitting. The test period is never used for parameter selection. An embargo around fold boundaries keeps label leakage in check. These controls are what make the AI/ML results in the thesis credible for the study.

Machine learning's role in the study therefore is not assumed in advance. The thesis is not based on the idea that AI/ML must beat the less sophisticated strategies. It treats that as a claim, something that must be tested. The following sections cover the crypto-specific evidence of AI/ML trading, regime-aware models, ensemble approaches, and robustness of the strategy evaluation process.

2.7 AI/ML Trading in Cryptocurrency Markets

The usage of AI and machine learning is becoming more widespread within the field of crypto trading in recent years in two ways. Firstly, high-frequency prices, volume, and derivative data of crypto assets provide good raw material for analytics, which can be analyzed, and the amount

of data that can be used as an input is growing constantly. Secondly, the availability of ML techniques for many researchers and practitioners allows them to analyze non-linear dependencies that cannot be detected by the means of statistical models.

The published evidence is mixed. While some research has found that ML models exhibit useful directional information regarding the pricing of cryptocurrencies against basic statistical benchmarks, others have been much more reserved in their conclusions, pointing out that the gains observed are reduced significantly when transaction costs and true out-of-sample tests are applied. This difference is critical in relation to the thesis, since improved prediction precision does not necessarily mean profitability or even risk adjustment. The problem of trading is not merely predicting direction, but doing so in such a way that translates into profits.

Recent studies further support this conflicting trend. State-of-the-art approaches such as Transformers, Temporal Fusion Transformers, as well as graph neural networks, tend to perform better than statistical approaches in prediction tasks. Still, the gains in prediction accuracy do not necessarily result in trading performance that is meaningful from an economic perspective when realistic assumptions about the trading process are taken into account (Wen et al., 2023; Bouteska et al., 2024). The academic debate is becoming more aware that there is a difference between predictive and realized trading performance.

A number of limitations can be found in the literature repeatedly. First of all, many studies concentrate exclusively on Bitcoin, making the results hardly applicable to the market in its entirety. Moreover, some out-of-sample periods used in the experiments are too short and never encompass one full cycle. Others report classification accuracy and do not go further, never translating predictions into trades, drawdowns, Sharpe ratios, or cost-adjusted returns. Transaction costs and slippage get left out or understated. These omissions are material, because crypto strategies, intraday ML ones especially, can generate high turnover. Even a model with modest predictive skill can turn commercially unviable if the signal demands frequent trading and the costs absorb the edge.

Another recurring concern is reproducibility. Many published studies rely on proprietary datasets or selective reporting of successful model configurations, making verification difficult. Recent methodological work in quantitative finance has therefore encouraged transparent experiment governance, version-controlled research pipelines and reproducible evaluation frameworks that

allow empirical findings to be replicated and audited (Bailey et al., 2014; López de Prado, 2018). These recommendations closely align with the audit and traceability framework adopted throughout this thesis.

For instance, Chen (2023) uses ML techniques on Bitcoin price prediction and concludes that random forest, support vector machine, and LSTM models beat basic out-of-sample directional benchmarks. The result is useful because it supports the idea that crypto prices hold some exploitable structure. But directional accuracy does not answer the trading question. A strategy still has to answer for return, volatility, drawdown, transaction costs, and robustness across conditions. This thesis extends that line of work by moving from prediction-level evaluation to trading-level evaluation across four assets and a common walk-forward framework.

Fischer and Krauss (2018) is another point of reference, though it works equity markets rather than cryptocurrency. They show LSTM networks beating other model families in a daily long-short portfolio. The relevance is clear, it demonstrates what deep learning can do in financial prediction. But the findings cannot just be assumed to carry over to cryptocurrency, which differs in volatility, market structure, liquidity, trading hours, and participant behaviour. And the way the LSTM attempt in the H3 experiments failed to produce any trades is a useful reminder: success in one setting guarantees nothing in another.

According to Krauss, Do and Huck (2017), deep neural networks, gradient boosted tree and random forest algorithms compete in a statistical arbitrage scenario with transaction costs and it is demonstrated that tree-based algorithms do relatively good in the presence of transaction costs. This is important in the current thesis because random forest and gradient boosting turn out to be one of the best single models AI/ML techniques in the presented results. Moreover, their findings confirm a general idea in finance regarding machine learning – complexity does not necessarily imply good performance.

Earlier, Atsalakis and Valavanis (2009) survey neuro-fuzzy and AI methods in financial forecasting. The work predates the deep learning era, but it is included here for two reasons. It shows that bold claims about AI in financial prediction are not new, and it documents the recurring problems, overfitting, in-sample evaluation, not accounting for trading costs. Those problems have not gone away in crypto research, where abundant data and market excitement

can push strong claims to be made before the evidence has been tested under sufficiently demanding conditions.

More recent work has focused on hybrid and deep learning architectures, attention-based BiLSTM, Temporal Fusion Transformers, decomposition-based LSTM models. These can improve forecasting accuracy in some situations. What is still thin is evidence that they consistently improve risk-adjusted returns after accounting for costs. This thesis is aimed at a more tractable set of supervised learning models before moving any further, and the logic behind that is straightforward: if simpler supervised models cannot show robust value under audited, cost-aware, multi-asset testing, then more advanced models require an even stronger justification.

Alessandretti, ElBahrawy, Aiello and Baronchelli (2018) employ ML methods for next day return prediction for hundreds of tokens and find that gradient-boosted trees and random forests significantly outperform simple baselines while predictability occurs only at shorter horizons. Directional predictability obtained in the study is consistent with the label approach and relative performance of tree-based models revealed in the results.

Sebastião and Godinho (2021) examine various ML models for their application to Bitcoin, Ethereum, and Litecoin, and demonstrate that out-of-sample predictive power of those models cannot be converted to positive trading outcomes once transaction costs are included. The discrepancy between predictability and trading potential that is demonstrated in the paper is precisely the issue that motivates the trading-level evaluation employed in this thesis.

Jaquart et al. (2021) study short-horizon predictability of Bitcoin returns and find that models outperform the benchmarks to some degree at intraday frequencies but the advantage depends on particular choice of features and costs, which is exactly the feature of model sensitivity to certain parameters described in section 5.3.1.

McNally et al. (2018) test LSTM and RNN models for forecasting Bitcoin prices and find that they have higher directionality accuracy than a random walk but emphasize that this cannot directly lead to profitable trading because of a specific set of circumstances, which is an observation made throughout the literature and addressed by the evaluation method of this thesis.

Supervised models chosen for this analysis are a good representation of various ML models applied to finance problems. Logistic regression (ML_01) is one of the most basic classifiers and low variance baseline that can be easily interpreted but is hard to overfit due to limited capacity to capture non-linearity. RandomForest (ML_02) grows multiple decision trees on different samples and sets of features.

Gradient-boosted decision trees are represented by two LightGBM implementations, ML_03 and ML_05. These build trees one after another, each new tree correcting the errors of the last. LightGBM is a widely used model for tabular data, efficient and often strong on structured feature sets. We have kept the second implementation, as a test that we know it will hold in a different hyperparameter configuration and not to rely on one model specification.

Support vector machine (ML_07_svm) is another non-linear classifier that finds a maximum-margin decision boundary and will work well in high-dimensional data, but it gets more demanding as the size of the training sample grows. At the hourly frequency we use in the analysis, training is still possible. Multi-layer perceptron (ML_08) stands in for the neural-network family among the single models. It is more flexible than the linear models, and that flexibility is accompanied by more variance and less stable out-of-sample behaviour on noisy financial data.

Two of the intended H3 models never produced usable trading signals. However, the ML04_xgboost and ML09_lstm models did not make any trades for all folds and assets. Rather than removing these failed models from the research, both are included in the data pool because the picture will be completely different if they are excluded. One clarification is that this caveat belongs to the H3 strategy namespace and should not be confused with H1/H2 where ML_04_svm and ML_05_gru are active implementations. In the field of AI/ML trading research, not every planned model works and it is essential to report failures as transparently as possible.

Every AI/ML strategy in this thesis has the same kind of label. It is a three-class directional outcome over a four-bar forward horizon with a 15-basis point activation threshold. The model learns to distinguish positive, negative and no-trade conditions and not simply whether the next return goes to zero. The threshold represents a practical constraint: a trade is only worth making if the expected move is large enough to cover transaction costs. It is the labelling framework of all models that helps them to be compared on equal footing.

The literature reviewed here is what motivates H1. Prior work suggests AI/ML models may find predictive patterns in crypto and other financial markets. But it also shows those patterns often fade out of sample, across assets, and after costs. This thesis puts H1 under stricter conditions: four assets, fifteen walk-forward folds each, common execution assumptions, transaction-cost modelling, and formal statistical evaluation.

2.8 Regime-Aware Trading

Regime-aware trading is based on a simple but important idea, markets do not behave the same way in every condition. A strategy that thrives in a bull market may perform poorly in a bear one; a mean-reversion strategy that operates in a range can fail by a strong trend. In crypto this is especially relevant, since clear transitions between bull, bear, and consolidation phases are routine.

Hamilton (1989) introduces the Markov regime-switching model, in which a time series alternates between latent states with distinct statistical properties. That framework is the theoretical ancestor of the regime-classification literature and motivates the deterministic regime features used here as a tractable alternative that avoids the calibration risks of latent-state models.

Ang and Bekaert (2002) apply regime-switching models to international equity return predictability and show that predictability varies substantially across regimes, a finding that directly motivates H2's hypothesis that conditioning on the detected regime should improve drawdown control relative to a static strategy.

Nystrup, Hansen, Madsen and Lindström (2017) compare hidden Markov models with threshold-based regime classifiers for financial time series and find that simpler rule-based regime definitions perform comparably to HMMs in trading applications, providing empirical support for the deterministic slope-based regime features used in `ML_07_regime_lgbm`. Chen and Tsang (2020) review the broader problem of classifying normal and abnormal regimes in financial markets and note that the choice of classification scheme matters less than the discipline of testing whether the chosen scheme actually improves downstream decisions, exactly the question the H2 experiment is designed to answer.

The theoretical case for regime-aware trading is closely linked to the Adaptive Markets Hypothesis. If efficiency is not fixed but shifts as participants, liquidity, and macro conditions change, then a signal's usefulness can shift by regime too. On that view a single static strategy is too rigid to keep up. A regime-aware strategy attempts to adapt, adjusting its model, its parameters, or its decision to be in the market at all, to match the current state.

The regime classifier used throughout this thesis is deliberately simple: the slope of the 200-bar simple moving average, thresholded against rolling price volatility, sorting bars into bull-trend, ranging, or bear-trend states. That choice was not the only option, and it is worth being explicit about why the alternatives were set aside. The most theoretically grounded alternative is a latent-state model such as a Hidden Markov Model or Gaussian Mixture Model, which infers regime membership probabilistically from the statistical behaviour of returns and volatility rather than reading it off an observable rule. The appeal is that regime assignment becomes a soft, data-driven judgement rather than a hard threshold. The cost is that the number of regimes and the distributional assumptions have to be fixed in advance, the model has to be refit at every retraining boundary alongside everything else in the pipeline, and the resulting state labels often resist a clean economic interpretation, a real liability when the regime feature also has to double as a conditioning variable that a reader can audit by eye.

Online changepoint detection methods (CUSUM, Page-Hinkley, Bayesian online changepoint detection) sit at the other extreme: rather than classifying the current state, they flag the moment a structural break occurs. These are well suited to markets prone to sudden statistical breaks, but they introduce a calibration problem of their own, a threshold set too sensitively manufactures false regime changes, set too conservatively it reacts only after the transition has already done its damage. The slope-based rule used here trades some of that statistical sophistication for transparency: every regime assignment can be recomputed by hand from the published threshold, which matters given that the same feature feeds both the ML models and the H2 regime-routing architecture. Its weakness is the mirror image of the HMM's strength, a deterministic rule reacts only after a moving average has actually turned, so it can lag a genuine transition by several bars.

Regime information can then be used in several ways. The simplest is to include the regime label as one more feature in a broader model. Here the model is not split into regime-specific versions; it just learns how the regime feature interacts with the other predictors. This is what the static

ML strategies do in the H2 comparison group, not fully regime-routed, but not blind to regime either.

A second approach is regime-conditional deployment, where a strategy only runs in selected regimes. For example, if AI/ML signals perform better in bear or ranging markets, the strategy may be activated in those regimes and replaced by passive exposure or no trade during bull-trend conditions. This treats regime information as a portfolio-level control rather than as part of the model's architecture.

A third, more aggressive way is regime-routed modelling. Train separate sub-models for different regimes, then route the live prediction to whichever sub-model matches the regime detected now. This is the architecture H2 tests through ML_07_regime_lgbm. The rationale is that the feature-to-return relationship differs across bull, bear, and ranging markets, and that dedicated sub-models can capture those differences better than one global model can.

The strength of regime-routed modelling is also what makes it risky. It assumes regimes can be identified accurately, and early enough to help. It also assumes each regime carries enough training observations to support its own model. Both assumptions can break. If the classifier reacts late, the strategy continues using the wrong sub-model right through a transition. If a regime's sample is thin, its sub-model overfits to a narrow slice of history. Both risks are practical in crypto, where transitions are often abrupt and event-driven.

The literature is mixed on regime-aware trading. Some studies report that regime conditioning helps, especially where markets show persistent, state-dependent behaviour. Others find it adds complexity, and can make things worse if the classifier is unstable or the model reacts too slowly. That unsettled debate is exactly the issue that H2 tests, testing whether a regime-aware AI/ML strategy actually cuts maximum drawdown relative to static benchmarks.

The study draws a clear line between regime information and regime routing. Regime information can still be useful, as a feature, or as a portfolio overlay. But the H2 analysis tests a more demanding claim: that splitting the model into separate per-regime routes improves drawdown control. The distinction matters, because a negative result for routing does not imply that regime awareness has no value. It only means this particular regime-aware architecture must demonstrate value against simpler alternatives.

This section provides the basis for H2. If crypto markets are non-stationary and regime-dependent, then a regime-aware model should cut drawdown and improve robustness, in principle. Whether that expectation is supported is what Chapter IV tests, evaluating the strategy across four assets, fifteen walk-forward folds, and realistic cost assumptions.

Additional empirical studies have indicated that information about the market regimes could add more value when applied to portfolio optimization than as a selection tool for predictive models themselves. Instead of changing the predictive algorithm itself depending on the regime, regime indicators could be utilized to change portfolio exposure or position size based on current market regimes (Gu et al., 2020). The approach offers another viewpoint on regime-based investments and fits into conditional implementation guidelines presented in Chapter V below.

2.9 Ensemble Learning for Trading

Ensemble learning combines several models or signals to get a steadier prediction than any one model provides independently. The intuition is simple: different models make different mistakes. As long as those mistakes are not perfectly correlated, combining the models cuts variance and steadies out-of-sample behaviour. It is a well-established idea in machine learning, and it has seen wide use in financial prediction.

Dietterich (2000) provides the theoretical case for why ensembles work, demonstrating through statistical, computational and representational arguments that combining classifiers reduces prediction variance. That reduction in variance is the core guarantee underlying H3's hypothesis that ensemble AI strategies should deliver more stable Sharpe ratios than any single model.

Wolpert (1992) introduces stacked generalisation (commonly called stacking) as a method of combining base learners through a meta-learner trained on their out-of-fold predictions, so the meta-learner never simply memorises in-sample errors. That design directly informs the ML_06D stacking out-of-fold architecture evaluated in H3.

Ke, Meng, Finley, Wang, Chen, Ma, Ye and Liu (2017) present LightGBM, a gradient-boosted framework using histogram-based leaf-wise tree growth that is substantially faster and more memory-efficient than earlier implementations. It serves as the base architecture for ML_03, ML_05 and the regime-aware ML_07, and its computational efficiency is what makes the multi-fold, multi-asset experiment tractable.

In trading, the appeal of ensembles is that financial data are noisy and unstable. A single model can perform well in one period and poorly in the next. Pool several learners and the strategy is less dependent upon one or the other. That is especially important in cryptocurrency where the performance of the model can vary in different regimes, assets, and volatility conditions.

The theoretical rationale for ensembles comes from the well-known tension in supervised learning: rigid models miss patterns in data and highly flexible models capture patterns that are really just noise and then perform poorly out of sample. Ensemble methods are primarily a tool for the second problem. When multiple separately trained learners disagree on a prediction, averaging their outputs can mute the effects of any one model's mistake and will shift the result towards the signal rather than the noise. In the case of hourly crypto returns, where the signal-to-noise ratio is low and the fold-level Sharpe ratios vary widely over retraining windows, this variance-reduction property is the main theoretical explanation why it is expected that ensembles will outperform individual models in consistency metrics.

Recent studies similarly indicate that the value of an ensemble method lies in improving robustness and not in being able to increase returns. By combining multiple independent models, the model improves prediction variance and stability over a wide range of market environments, and the improvement is found more clearly in downside risk measures than in average return measures (Fischer and Krauss, 2018; Gu et al., 2020). This distinction is especially important in financial environments in which the ability to store capital when a bad market is trading may be as important as improving returns.

However, in financial markets, the benefits are not automatic. The strongest theoretical payoff comes when base model errors are just weakly correlated and trading data are consistently out of line. Models trained on the same price, same features, same labels tend to make the same mistakes. If they are all reacting to the same market noise, then the ensemble is just averaging correlated errors, not diversifying them. And worse, it can dilute the one strong model's useful signal in the process.

And this is the main challenge in moving from stable predictions to actual trading performance. Smoother predictions do not necessarily mean higher Sharpe ratios. If averaging leads to larger losses and bigger gains, then drawdown can improve and return could decrease. That tension

runs through the H3 results: ensembles reduce maximum drawdown, but they do not improve the primary Sharpe-stability criterion.

There are three voting-based aggregation methods in H3. The simplest is the one that counts directional predictions and acts on whichever class achieves a plurality and drops all confidence information. A second variant combines the raw probability output of each learner, which is more meaningful if some of the models are much more certain than others. A third variant weights the probability outputs of each model according to the most recent validation performance of each model. The goal is to make the ensemble more likely to trust the base learners when the previous training set was more accurate. All three are open enough to be audited and cheap enough to be relearned from scratch at each walk forward step.

The stacking approach is yet another step down this road. Rather than choosing some form of aggregation algorithm a priori, it chooses a rule of aggregation: a very simple second stage classifier is trained to predict the correct class using as input the prediction made by the base learners. The problem with such approach is that it may just be memorizing the actual events seen in training; for this reason, the approach we have taken involves predicting the base learners prediction on the held-out data before even seeing them in the second stage.

A fourth architecture combines realised trade decisions with predicted probabilities. Where probability-averaging keeps gradient information about each model's confidence in learning, this approach reduces every base learner to a binary vote before aggregating, making the logic easy to understand and explain but discarding the degree of conviction information that distinguishes a borderline signal from a strong one. That loss is most costly in low-signal environments where the trade quality of calibration is highly variable for the base learners.

The final architecture replaces raw probability values with ordinal rankings before aggregating, which avoids the need for well-calibrated probability estimates across base learners. This is appealing, as calibration quality in short walk-forward windows is difficult to verify, and rank aggregation is robust to models that are directionally accurate while the probability outputs are often different in scale or distribution. The downside is that fine-grained probability information leads to signal strength, so converting to ranks fires more trades and amplifies the cost impact that the other ensemble variants are more sensitive to.

A regime-aware ensemble is also presented as secondary evidence in the thesis. It lets the ensemble weights or decisions shift with the detected market regime. It is outside the primary H3 test because it fuses two ideas at once, regime conditioning and ensemble aggregation. Even so, it points to a worthwhile future direction, particularly because conditioning at the ensemble level may prove less brittle than routing the whole prediction process through separate regime-specific models.

On balance the literature supports the claim that ensembles steady predictions. On trading-level improvement it is less settled. In markets the objective is not just to predict a bit better; it is to trade in a way that survives costs, drawdowns, and shifting regimes. An ensemble that improves accuracy but increases turnover may not improve realised performance at all. By the same token, one that trims drawdown can still fail to improve Sharpe if it clips the upside too.

This section is the basis for H3. If ensemble learning really does cut variance and steady performance, then ensemble AI/ML strategies should trade more consistently than single-model ones. Chapter IV tests this directly, comparing ensemble strategies with single-model AI/ML strategies across the four crypto assets.

2.10 Backtesting and Robustness in Financial Research

Every number in Chapters IV and V is a backtest output and a backtest is exactly the kind of evidence that can be misleading without anyone actually meaning it to. A strategy might have a good historical Sharpe ratio for reasons that have nothing to do with predictive skill: a parameter grid searched until something worked, a feature that leaked tomorrow's price into today's prediction, a cost assumption that was left out, or a single market episode; the SOL and XRP outlier folds in section 4.2.3 are a case in point here and unlikely to happen again. But none of these failure modes shows up in the headline number. What follows is an account of how the thesis tries to rule each one out, not as an abstract methodological survey but as a description of the guardrails that we built into the H1, H2 and H3 experiments.

Data-snooping is the risk that has received the most attention in the methodology literature and it is an important one: we have sixteen rule-based and AI/ML strategy families with four assets and seven cost scenarios in our sample and the search area here is large enough that something would be good by chance if the evaluation protocol did not guard against it. White (2000) and Hansen

(2005) developed formal reality-check and superior-predictive-ability tests for this problem, asking whether the best performer in a large search remains better when the search itself is taken into account. Romano and Wolf (2005) have developed this logic further by implementing stepwise control of inference over many simultaneous comparisons.

Harvey, Liu and Zhu (2016) argue that the standard five per cent significance level is too lenient in empirical finance once data snooping and publication bias are taken into account, and recommend a minimum t-statistic of around 3.0 before a new factor or strategy is treated as credible. The results of this thesis do not reach that level, which reflects the weakness of the underlying effects rather than any shortcoming in the methodology.

Aronson (2007) shows through a practitioner-oriented approach to evidence-based technical analysis that most published technical trading results become irrelevant once proper data-snooping corrections are made. The hypotheses tested and the walk-forward methodology used in this thesis are consistent with such an adjustment even in the absence of a White's reality check.

Combinatorial Purged Cross Validation, suggested by López de Prado (2020), would allow us to build a distribution of backtest outcomes in order to reduce the problem of selection bias beyond a simple chronological train-test approach. The latter is applied in the current thesis for practical reasons, while the former is the preferable method.

It does not use White's reality check or Hansen's SPA test directly. It reduces selection bias by pre-specified hypotheses, walk-forward validation, hyperparameter selection on the validation window only, bootstrap confidence intervals, and a formal audit of the evidence chain. But it provides a more powerful basis for inference than a single backward-looking backtest ever could.

A second concern is whether the Sharpe ratio can be trusted. Lo (2002) showed that Sharpe estimates are sensitive to distributional assumptions and to serial dependence in returns. Bailey and López de Prado (2014) followed with the deflated Sharpe ratio to correct for non-normality and multiple testing. Bailey, Borwein, López de Prado and Zhu (2017) also introduced the probability of backtest overfitting as an estimate of how likely a strategy chosen for strong in-sample numbers is to fail out of sample. And neither method is implemented here, so we take both as future research priorities.

A third relevant method is combinatorial purged cross-validation, which scores performance across many combinations of training and test folds and removes the observations that could leak, yielding a distribution of out-of-sample outcomes rather than a single chronological test path. CPCV itself is not implemented here, but its underlying logic is reflected in the walk-forward design: training, validation and testing are separate in time and embargo controls prevent leakage around the fold boundaries.

In Chapters IV and V, all results are based on the same walk-forward design, a time-series cross-validation discipline Bergmeir and Benítez (2012) point out to be a key component once observations can no longer be regarded as independently and identically distributed: 12 months of training data, 3 months of validation and 3 months of out-of-sample testing are run over fifteen non-overlapping cycles for BTCUSDT, ETHUSDT, SOLUSDT, and XRPUSDT for 60 fold-asset evaluations. The model never sees a fold's test data until its weights are frozen, and the entire fifteen-fold sequence is repeated for every strategy in each hypothesis, again and again. That repetition is the point: a single train-test split would ensure that one lucky time-frame in the coin history to the end of the crypto history would decide which one to choose, while 15 cycles of retraining for each of the assets would force a strategy to prove itself again and again through the bear market in 2022, the bull market in 2024, and the mixed conditions of 2025.

Cost realism is treated as a first-order design constraint rather than an afterthought applied to an already-computed gross return. The base scenario is 6 basis points commission and 2 basis points slippage per side (16 basis points round-trip on every trade) and 6 other scenarios stress the H1 verdict up to forty-four basis points round-trip and the worst case would be done by each scenario. The constraint is strongest on AI/ML strategy groups, which trade about three times more than rule-based comparators over the evaluation window. A classifier that crosses its probability threshold more often always earns more round trips, therefore the total gross signal would need to be passed over a higher bar in terms of costs than a comparator that trades relatively little. If we report only gross returns and apply costs as a single, uniform haircut after the fact, we would have hidden this asymmetry that is one of the structural reasons the H1 mean-Sharpe verdict is in favor of those with lower turnover rule-based comparators.

The Monte Carlo simulation and block bootstrapping techniques are used to check the robustness through the alternative paths other than the historical one that we have observed. The former

technique would also destroy the short dependence that is present in the crypto data by considering each bar in the dataset as independent while the truth is that volatility and movements occur in clusters of several bars. The purpose is not forecasting but checking if what we have observed is not sufficient for an alternative shuffle of the return series.

Multiple testing is performed for each hypothesis using the Holm-Sidak stepdown method. We do not perform a single correction for H1, H2, and H3 because the three hypotheses are pre-specified to be tested for genuinely distinct claims. That is justified, but it is also a decision with consequences. A more conservative family-wise correction for all three might be useful in future work, especially if there are more claims to test.

Robustness is further assessed by regime breakdown and leave-one-asset-out analysis. The regime breakdown shows if a result is widespread or narrowly focused on one market condition. The leave-one-asset-out test is used to determine if the pooling method is strongly dependent on one particular asset. This is extremely important for both SOL and XRP, whose outlier folds impact mean-based performance comparison considerably. Pooled performance would have been considerably more stable without such testing techniques.

It is the backtesting technique described in chapter III that is applied here consistently. The research involves a rolling training window of twelve months, a validation window of three months, a test window of three months and an embargo of 24 bars. Feature scaling is fitted on training data only and then applied to validation and test data without refitting. Hyperparameters are chosen only on the validation window. And the backtest engine applies a single-bar execution lag, transaction costs, and one signal-to-fill process across all strategies.

The objective of this design is to bridge the gap between a backtest result and credible empirical evidence. It cannot close it completely. The study is still backtest only; it does not model funding costs, does not assume live execution, does not make every statistical correction in the literature. What it does do is hold to a higher standard than much of the published AI/ML trading work and it is clear where its limits are.

This section sets the evidentiary bar for the thesis. Hypotheses are not accepted or rejected on the basis of one model's performance. They go through walk-forward testing, cost-aware backtesting, statistical inference, robustness checks and audit traceability. Chapter III shows how those safeguards are built. Chapter IV presents the results.

Recent developments in quantitative finance have also made reproducibility a cornerstone of empirical credibility. In addition to statistical validation, experiment versioning, deterministic execution pipelines, configuration management and audit trails are proposed in recent years and evidence from published results are now independently checked and verified (Bailey et al., 2014; López de Prado, 2018). These practices decrease levels of freedom of researchers, increase confidence in findings and make the empirical trading research more transparent.

2.11. Hypothesis Development

Three hypotheses formalise the research questions from section 1.4. Each is set out below with its theoretical motivation, its operational form, and the criterion it will be judged against. Together they form the falsifiable structure the empirical study is built to test.

2.11.1 Hypothesis H1: Risk-Adjusted Outperformance

H1 addresses RQ1 and risk-adjusted outperformance. Its motivation arises from a tension in the literature. The predictive-accuracy work in section 2.6 reports several cases where AI and ML models pull some directional signal out of crypto price data; the efficient-markets and random-walk work in sections 2.1 and 2.4 implies that signal should be too faint to beat simple benchmarks after costs. H1 is framed as a positive claim about risk-adjusted superiority, so that even a failure to reject the null is informative. Operationally: over the four assets and fifteen walk-forward folds, the mean fold-level Sharpe ratio of the AI/ML group is higher than that of the rule-based group at the five per cent level of significance in the bootstrap test (see section 3.10). In order for H1 to be true, we need the above result to hold for three out of the four assets. H1 holds when both the pooled and asset-level tests succeed, and not otherwise.

2.11.2 Hypothesis H2: Regime-Aware Drawdown Control

H2 addresses RQ2 and regime performance, and the goal is to reduce drawdown. This motivation is based on the regime-aware trading literature in section 2.8 and the Adaptive Markets Hypothesis. If markets are non-stationary, and regimes have different return distributions, then it is likely that a model that is based on the current regime should be more powerful than one that ignores it, at least in terms of drawdown. Operationally: across the four assets and fifteen folds, the mean maximum drawdown of a regime-aware AI strategy (a

LightGBM with per-regime signal routing) is lower than that of a static benchmark pool combining eight rule-based strategies with six regime-informed but non-routed AI strategies. Asset-level support requires the same direction in at least three of the four assets. Choosing static-regime-informed AI comparators deliberately raises the bar, it isolates what per-regime routing adds architecturally from what simply including the regime label as a feature adds informationally.

2.11.3 Hypothesis H3: Ensemble Stability

H3 addresses RQ3 and the question of ensemble versus single-model stability. The motivation is the bias-variance decomposition of supervised learning. Ensemble averaging cuts the variance of base-learner predictions, so it should steady out-of-sample performance, especially in low signal-to-noise settings like financial time series. Operationally: across the four assets, the pooled median fold-level Sharpe ratio of the ensemble group exceeds that of the single-model AI group, and the direction holds across all four assets. The primary metric is median Sharpe, not mean, because the goal of ensembling is to stabilise the typical fold rather than to stretch the tail. A secondary, deployment-relevant criterion asks whether ensembles reduce mean maximum drawdown across the four assets.

2.11.4 Integrative Claim

An integrative claim links the three hypotheses together. If all three are true (H1, H2 and H3), the broad thesis claim that AI and ML strategies generally improve risk-adjusted performance, robustness and consistency in cryptocurrency is still valid: if any one fails, the claim is only partial and only survives in the dimensions where its supported hypotheses cover. If none does, then the overall claim is rejected in the tested area, although conditional sub-findings might still be reported. The synthesis of the three hypotheses in Chapter IV is where that integration is done.

2.11.5 Formal Hypothesis Statements and Mapping Table

The formal statements of H1, H2 and H3, with their primary metrics and evidence sources, are reproduced below. Table 2.1 plots the three hypotheses onto their research questions, primary metrics, and audited evidence pools.

1. H1: AI- and ML-based trading strategies generate statistically significant superior risk-adjusted returns, as measured by Sharpe ratio, compared to traditional rule-based benchmark strategies in selected cryptocurrency markets.
2. H2: Regime-aware AI and ML trading strategies exhibit lower maximum drawdowns and higher robustness across different market conditions compared to static benchmark strategies.
3. H3: Ensemble-based AI trading strategies outperform single-model approaches in terms of stability and consistency of returns across multiple cryptocurrencies.

Table 2.1 Research Questions, Hypotheses and Primary Metrics

Research Question	Hypothesis	Primary Metric	Evidence Source
RQ1, Risk-adjusted outperformance	H1: AI/ML vs rule-based	Mean and median Sharpe ratio (fold-level)	H1 multi-asset audited pool, 840 observations
RQ2, Regime performance	H2: Regime-aware vs static	Mean maximum drawdown; regime breakdown	H2 multi-asset audited pool, 60 asset-fold pairs
RQ2 / RQ3, Stability across architectures	H3: Ensemble vs single-model	Fold-level median Sharpe; cross-asset rank stability	H3 multi-asset audited pool, 4 assets × 15 folds
RQ3, Implementation limits	Cross-cutting	Cost stress, robustness, audit caveats	Cross-hypothesis synthesis (SYN_H1H2H3)

2.12 Research Gap

Pulling the literature together, the gap this thesis fills can be stated as follows. Many studies report promising AI/ML trading results; far fewer bring multi-asset, out-of-sample, cost-aware, regime-aware, and ensemble-comparative evidence across major cryptocurrencies with reproducible experiment governance and an independent audit trail. Specifically, the literature lacks a single coherent study that does all of the following at once: (a) tests AI/ML against rule-based benchmarks under identical execution assumptions in a multi-asset universe; (b) tests regime-aware AI/ML against static-benchmark approaches under those same conditions; (c) tests ensemble AI/ML against single-model AI/ML under those same conditions; and (d) integrates all three into an audited cross-hypothesis synthesis with structured claim governance. The present

thesis fills that gap by running three independently designed and audited hypothesis experiments, H1, H2 and H3, within one unified empirical framework.

2.13 Summary

Chapter II reviewed the theory, the conceptual framework, the empirical literature, and the systemic considerations that frame the study. On theory, five strands were identified as most relevant: efficient markets, random walk, adaptive markets, behavioural finance, and financial machine learning. The conceptual framework in Figure 2.1 connects those theories to the three research questions, the three hypotheses, the empirical design, and the audited outcomes. And the empirical review, across cryptocurrency markets, algorithmic and rule-based trading, ML in finance, regime-aware trading, ensemble methods, and backtesting practice, identified a clear gap: no multi-asset, audited, cost-aware, walk-forward evidence spanning risk-adjusted return, regime conditioning, and ensemble stability, all integrated into one cross-hypothesis synthesis. Chapter III operationalises those requirements through a methodology.

Collectively, the above literature shows that despite the promise of AI and ML in finance, there is still work to be done on multi-asset cryptocurrency analysis, backtesting with realistic transaction costs, statistically valid evaluation, and independently verifiable methods. These unsolved problems are the direct motivation for the method presented in Chapter III.

CHAPTER III: METHODOLOGY

3.1 Research Philosophy and Design

Every claim in this thesis traces back to a number computed from historical market data through a documented, version-controlled pipeline, rather than to a qualitative judgement or an interview transcript. The unit of analysis is the strategy-fold observation, one strategy, one asset, one walk-forward fold, and the out-of-sample performance metrics that are produced. Hypotheses are cast as falsifiable claims about distributions of those strategy-fold observations and tested with bootstrap, parametric, and non-parametric procedures. The structure is deductive in its hypothesis-testing structure: H1, H2 and H3 are pre-specified and effectively pre-registered through the experiment configuration files. Exploratory elements are limited to the sensitivity analyses, the regime breakdown, and the search for conditional sub-findings.

There are three design principles that guide our approach. First, every strategy in a given comparison must be executed under the same assumptions (the same bar code, the same cost model, the same signal-to-fill convention) in a single execution code path in the backtest engine. Second, no test period information can affect any training or hyperparameter choice (processed in walk-forward folds with embargo bars, train-only feature scaling, and validation-only hyperparameter selection). Third, all the results should be traceable back to a concrete configuration, a concrete code commit, and a concrete random seed.

This commitment to traceable quantitative evidence is most apparent in the pre-registration discipline built into the experiment configuration files. Each hypothesis experiment is set up with a configuration, before the experiment runs, that describes the strategies, the evaluation period, the performance metrics, the statistical tests, and the verdict criteria. It is frozen, committed to version control, and run by the platform. Any change to that configuration will generate a new experiment identifier with a new manifest, but the results will still be auditable. It is essentially the empirical counterpart of pre-registration practices in clinical trials and behavioural science in the last decade which Nosek et al. (2018) refer to as pre-registration of hypotheses before data analysis.

The deductive central part is balanced by some inductive work at the level of sub-findings. After those primary findings have been obtained, the cross-hypotheses synthesis reexamines the evidence behind them, looking for any patterns that might have implications going beyond the preregistered hypothesis. Such findings, the conditional sub-findings IF-01 to IF-05, found post hoc, are described in the thesis along with the supporting statistics but are not elevated to the status of primary hypotheses. This is an intermediate position. Pre-registration means excluding any post hoc pattern-finding, while pure inductive reporting would mean leaving the door open to snooping for data. The study reports both, clearly labelled, and allows the reader to evaluate them separately.

3.2 Asset Universe and Data Sources

The asset universe consists of four USDT-margined perpetual futures contracts on Binance Futures: BTCUSDT, ETHUSDT, SOLUSDT, and XRPUSDT. The four assets were chosen to meet three criteria at the same time: to represent the major-cap tier, to have sufficient historical liquidity to support 1-hour bar testing, and to be exposed to genuinely different price behaviour for 2021-2026 (BTC and ETH being macro-sensitive majors; SOL and XRP being event-based mid-caps with big bull run episodes). All data is collected from the Binance Futures API and no spot data or other exchange data are used in our primary experiments. We use only one exchange and one instrument type, USDT-margined perpetuals, to keep OHLCV prices, funding rates, open interest, mark prices and index prices the same.

Table 3.1 Asset Universe, Data Periods and Walk-Forward Windows

Asset	Symbol	Instrument	Data Range	OOS Window
Bitcoin	BTCUSDT	Binance perpetual	2021-01-01 → 2026-03-01	2022-04-02 → 2026-01-01
Ethereum	ETHUSDT	Binance perpetual	2021-01-01 → 2026-03-01	2022-04-02 → 2026-01-01
Solana	SOLUSDT	Binance perpetual	2021-01-01 → 2026-03-01	2022-04-02 → 2026-01-01
Ripple	XRPUSDT	Binance perpetual	2021-01-01 → 2026-03-01	2022-04-02 → 2026-01-01

All four assets share a common out-of-sample window. BTC has an extended dataset in canonical storage going back to 2020-01-01, but the H1, H2 and H3 experiments load it from 2021-01-01 to align the cross-asset window. Wherever BTC training data is from before 2021-01-01 in any experiment, the out-of-sample window is the same as all the other three assets.

3.3 Data Preparation and Canonical Schema

The raw exchange data go through a canonicalisation pipeline that has a fixed column schema, monotonic UTC timestamps, gap detection, synthetic candles imputation, and asset manifest generation. If the feed contains bar gaps, synthetic candles are generated (labelled with a Boolean) and are ignored in ML training. The canonical schema contains timestamp, asset, exchange, open, high, low, close, volume, funding rate, open interest, mark price, index price, and synthetic flag. Open interest is only available to BTC; for ETH, SOL, and XRP it is dropped, and downstream features that depend on it are dropped (35 feature vector, not 39).

Table 3.2 Canonical Data Schema (Fields per Bar)

Field	Type	Description
timestamp	DatetimeIndex (UTC)	Bar open time, hourly
asset	string	Asset symbol e.g. BTCUSDT
exchange	string	binance
open / high / low / close	float64	OHLC prices in quote currency
volume	float64	Quote volume in USDT
funding_rate	float64	Hourly funding rate, forward-filled
open_interest	float64	USD-denominated open interest (BTC only; NaN otherwise)
mark_price	float64	Mark price
index_price	float64	Index price
_synthetic	bool	True if bar was gap-filled (synthetic flat candle)

3.4 Feature Engineering

Constructing a feature matrix for a trading strategy backtest that is going to work in real-time means treating each bar as if the future does not yet exist. There are three ironclad rules that must be followed when developing the algorithm in code. First, no feature may depend on a future bar. Second, no feature may encode the outcome being predicted, directly or through a transformation that contains the forward return. Third, every feature value is checked to confirm it could have been computed with data available at the bar's timestamp and no later. Violations of any of these rules produce a backtest that overstates what live performance would have been, often dramatically, which is why they are enforced as code-level assertions rather than documentation conventions. After applying these constraints, the pipeline produces thirty-nine features for BTC and thirty-five for ETH, SOL and XRP, with the four-feature gap reflecting BTC's availability of open-interest data from the Binance Futures API at the time of collection, while ETH, SOL and XRP had gaps that made inclusion inadvisable. The resulting feature set is summarised in Table 3.3.

Table 3.3 Feature Categories and Counts per Asset

Category	Examples	Count (BTC / others)
Log returns	log_return_{1, 3, 6, 12, 24}	5 / 5
SMA levels	sma_{20, 50, 100}, price_vs_sma_{20, 50}, ma_slope_{20, 50}	8 / 8
EMA levels	ema_{20, 50, 100}	3 / 3
RSI	rsi_{14, 21}	2 / 2
MACD	macd, macd_signal, macd_hist	3 / 3
Rate of change	roc_{12, 24}	2 / 2
ATR	atr_14	1 / 1
Bollinger width	bollinger_width_20	1 / 1
Rolling volatility	rolling_vol_{24, 72}	2 / 2
Candle structure	high_low_range, candle_body_size, upper_wick, lower_wick	4 / 4
Volume	volume_zscore_{24, 72}, volume_change_1, volume_ma_ratio_24	4 / 4
Derivatives (BTC only)	funding_rate, funding_rate_lag_1, open_interest_change, open_interest_zscore	4 / 0

Category	Examples	Count (BTC / others)
Regime	trend_regime, volatility_regime, drawdown_regime	3 / 3
Total features		39 / 35

Three of the thirty-five to thirty-nine features carry extra weight because they do double duty: they enter the ML models as predictive inputs and they also serve as the conditioning variable in the H2 regime-routing experiment. Each is computed as a deterministic rule rather than a learned output, which keeps interpretation clean and avoids adding a second layer of model risk. The trend feature reads the slope of the 200-bar simple moving average for the last 24 bars, normalized against the price volatility. A slope above half a standard deviation of rolling price returns is bull-trend; a slope below that level is bear-trend; anything in between is ranging. We chose the 200-bar SMA because it lasts about 8 days at hourly resolution and provides a sense of the direction that the price is going and does not have to be affected by each intraday fluctuation. The volatility feature compares the 24-bar annualised realised volatility against its 72-bar moving average, moving between high-volatility and low-volatility at that point. It is a regime of volatilities (as compared to the recent history of the asset) and is the same for BTC and SOL, even though they have very different baseline volatility profiles. The drawdown feature is used to show how far price has fallen back from its trailing 72-bar high: below 5 per cent down is near peak, above 5 per cent below is a drawdown state, and the border between them is a transitional zone. The 72-bar window is approximately three days long and we have selected it to capture the short-duration drawdown episodes common in crypto without getting triggered by regular intraday swings.

The feature engineering pipeline operates in three stages. First, the canonical OHLCV-plus-derivatives data is converted into the indicator vectors, and under the point-in-time rule, no forward shift and no use of the label as a feature is allowed. The second normalises those vectors with a StandardScaler fitted on training data only, then applies the fitted transformer to the validation and test partitions without refitting. The third applies fold-specific anti-leakage guards: the last horizon-plus-embargo bars of the training window are masked as not-applicable, so the label cannot leak across the validation boundary.

Two choices in the feature set warrant comment. First, the 35-to-39 feature dimensionality is a deliberate trade-off between expressiveness and estimation noise. Adding more features increases overfitting risk in a low signal-to-noise setting; removing too many risks omitting genuinely informative inputs. The chosen count is consistent with what López de Prado (2018) recommends for daily and intraday classification. Second, the regime features that drive the H2 treatment strategy are computed as deterministic rules over the underlying price and volatility series, not as output of a separate ML model. Keeping them deterministic means the regime classifier itself does not add any additional overfitting risk and it attributes our H2 result to the routing architecture rather than an upstream classifier.

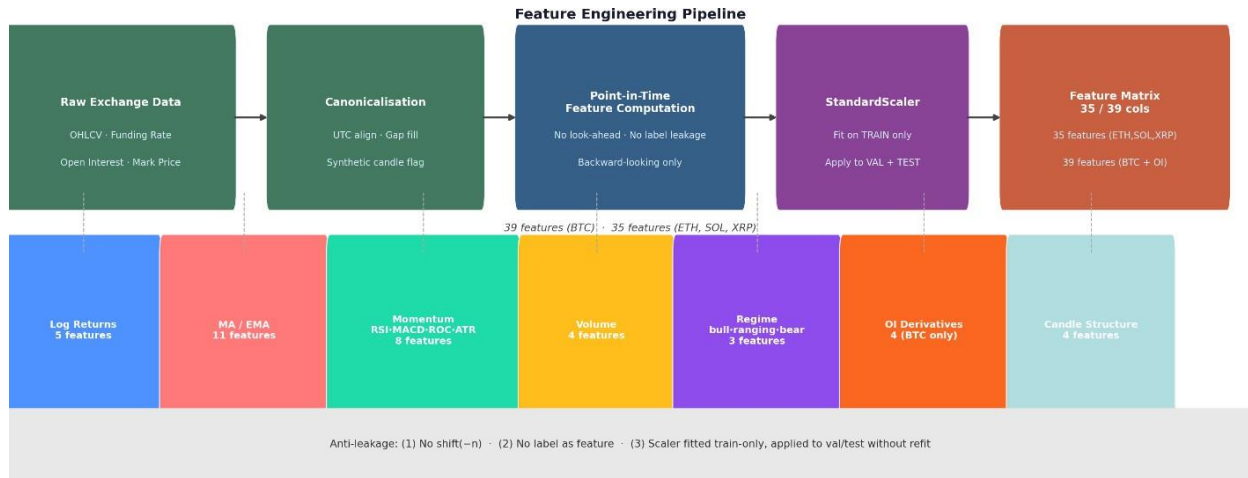


Figure 3.1 Feature Engineering Pipeline

Three pipeline stages have strict anti-leakage invariants. BTC has 39 features, including open-interest derivatives; ETH, SOL and XRP have 35.

The pipeline in the figure above translates the raw Binance Futures data into the feature matrix shown in Table 3.3 in three steps of the pipeline, each of which has its own set of point-in-time invariants. There are three rules at the code level that apply throughout: no feature column can use a negative shift or forward look-up, no feature column can contain the forward label or forward return, and all transformations can be performed from data at bar t without reference to bar $t+1$ or later. The StandardScaler fitted at stage three is asset-specific and fold-specific. A new scaler is fitted at each fold boundary on the training data of a fold and applied without refit to validate and test the results of that fold. This design means the scaler cannot absorb any test-

period distributional shift, which would constitute a subtle but material form of data leakage on a non-stationary return series.

The four open-interest features available only for BTC reflect data availability rather than a modelling choice: the Binance Futures API provided reliable open-interest history for BTC across the full evaluation window, but ETH, SOL and XRP had gaps or inconsistencies that made inclusion inadvisable. The mild four-feature asymmetry is noted in section 3.13 Methodological Limitations and carried through to the cross-asset sensitivity analysis in section 3.11.

3.5 Label and Signal Framework

Every ML strategy in H1, H2 and H3 is trained on one directional label specification: the label is +1 if the 4-bar forward log return exceeds +0.0015 (15 bps), -1 if it falls below -0.0015, and 0 otherwise. The label is computed across the entire historical series in a single pass and then masked as not-applicable in the final (horizon + embargo) bars of each training window, to prevent leakage into validation or test. Predictions become trading signals through an asymmetric probability-threshold rule: a long signal requires the predicted positive-class probability to reach at least 0.55, a short signal requires the predicted negative-class probability to reach at least 0.45, and otherwise returns a zero signal.

Every strategy, rule-based or AI/ML, produces a `SignalFrame`, a pandas `DataFrame` indexed by UTC timestamp with a single integer signal column taking values in $\{-1, 0, +1\}$. The backtest engine then imposes a single-bar execution lag: a signal at bar t fills at the open of bar $t+1$. That lag is the primary anti-leakage control at the execution layer.

The directional label needs further detail as it is in financial machine learning, one of the most important methodological decisions. There are three key features of this specification that need to be discussed.

First of all, the four-bar horizon (4 hours at hourly resolution) is a good choice that balances two things. If we want to have the label with a lower signal-to-noise ratio (the noise in the horizon is less likely to accumulate), then having a shorter horizon would shorten the holding period for the model and increase turnover-related costs. A longer horizon would minimize turnover but add intra-horizon noise and make the label more susceptible to regime changes along the way. Four

bars would put about twenty per cent of the bars in each directional class and roughly 60 per cent in no trade (a good mix for classifier training).

Second, the fifteen-basis-point activation threshold was adjusted to have the cost of trade in the base scenario at about sixteen basis points (6 bps commission plus 2 bps slippage per side). So fifteen basis points is essentially the minimum forward return required to break even on a single round-trip trade. Lowering the threshold would have more signals but the bulk of them would be absorbed by costs; raising it would reduce signal frequency but increase the expected value of each trade signal.

Wilder (1978) introduces the Relative Strength Index as a bounded momentum oscillator to detect overbought and oversold conditions. RSI at 14-bar and 21-bar periods is included in the 35-to-39 feature set and also forms the basis for the RSI mean-reversion rule-based family RB_03, making it both a predictive feature and a benchmark strategy component.

Pring (2002) provides a comprehensive treatment of technical indicators including MACD, Bollinger Bands and Donchian channels (the indicator families underlying RB_04, RB_05 and RB_06) and documents the market conditions under which each is expected to perform reliably. That account informs the parameter ranges evaluated during the validation step of each walk-forward fold.

Third, the asymmetric probability threshold (0.55 for longs, 0.45 for shorts, no-trade otherwise) is chosen to correspond to the base rates of the three classes. The asymmetry is intentional. Positive class observations appear slightly less frequently than negative class ones in cryptocurrency because the median hourly return is slightly negative once transaction costs come out of the realised return. The threshold values were fixed in advance and held constant across every strategy; testing alternative combinations is left for future sensitivity work.

There is a labelling alternative that has received attention in recent financial ML work, the triple-barrier method, which combines a horizontal take-profit barrier, a horizontal stop-loss barrier, and a vertical time barrier, then labels each observation by whichever barrier is hit first. The purpose is to line label generation up with trade execution by modelling the path-dependent exit decision directly. This thesis does not use triple-barrier labels in its primary experiments, and for a specific reason: bringing them in would change the comparison interface between AI/ML strategies (which can generate triple-barrier-style exit logic directly) and rule-based strategies

(which exit on opposite signals). That asymmetry would complicate the cross-family comparison. Triple-barrier labelling is identified as a future direction.

3.6 Strategy Taxonomy

Strategies are classified along three axes: primary family (rule-based, ML-based, hybrid, mutually exclusive), regime-awareness (Boolean), and ensemble type (none, voting, weighted, stacked, hierarchical). This taxonomy enables structured cross-family comparisons, rather than ad-hoc naming-based ones. Table 3.4 summarises the strategy universe across the three hypothesis experiments. Read the strategy identifiers with their full experiment-specific suffix wherever it is given: the same numeric prefix can mean different implementations in H1/H2 and H3.

Table 3.4 Strategy Taxonomy Used Across H1, H2 and H3

Strategy ID(s)	Description	Family	Regime-Aware	Ensemble	Used in
RB_01	Buy-and-hold	Rule-based	No	None	H1, H2 static control
RB_02	Moving-average crossover variants	Rule-based	No	None	H1, H2 static control
RB_03	RSI mean-reversion variants	Rule-based	No	None	H1, H2 static control
RB_04	Bollinger Band mean-reversion variants	Rule-based	No	None	H1, H2 static control
RB_05	Donchian breakout variants	Rule-based	No	None	H1, H2 static control
RB_06	MACD trend	Rule-based	No	None	H1, H2 static control
RB_07	Volatility breakout variants	Rule-based	No	None	H1, H2 static control
RB_08	EMA trend filter variants	Rule-based	No	None	H1, H2 static control
ML_01-ML_03	Logistic regression, random forest, gradient boosting	ML	No	None	H1 primary, H2 static control, H3 single-model
ML_04_svm; ML_05_gru	Support vector machine and GRU recurrent model	ML	No	None	H1 primary, H2 static control
ML_06_ensemble	Supplemental/static ensemble implementation	ML	No	Ensemble	H1 supplemental; H2 static control
ML_07_regime_lgbm	Regime-aware LightGBM with per-regime routing	ML	Yes	None	H1 primary; H2 treatment
H3 active singles	ML_01, ML_02, ML_03, ML_05_lightgbm, ML_07_svm, ML_08_mlp	ML	No	None	H3 single-model group

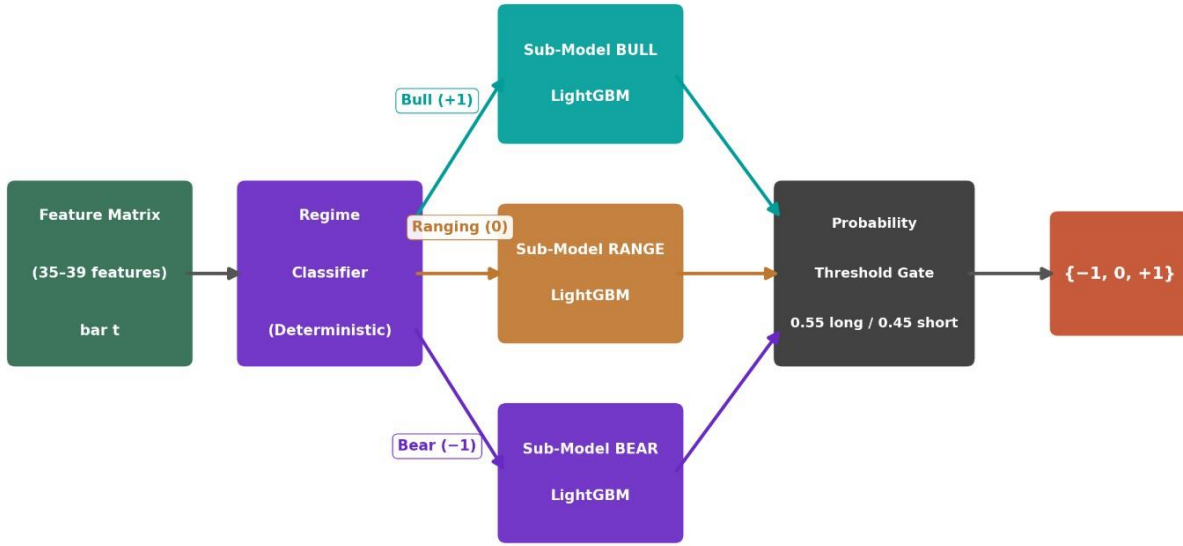
Strategy ID(s)	Description	Family	Regime-Aware	Ensemble	Used in
ML_06A-ML_06F	Majority, weighted, soft, stacking OOF, strategy-signal and rank-based ensembles	ML	No	Multiple	H3 primary ensemble group
ML_06G	Regime-aware ensemble	ML	Yes	Hierarchical	H3 secondary only
ML_04_xgboost; ML_09_lstm	Failed H3 candidate models; zero trades	ML	No	None	H3 failed/excluded

The strategy implementations differ across the three hypothesis groups in their internal mechanics, but they all converge on one output interface, a `SignalFrame` indexed by UTC timestamp with a single integer signal column. That shared interface allows the backtest engine to treat every strategy the same. Rule-based strategies are based on indicator thresholds and comparators; single model AI strategies make probabilistic predictions and convert them into signals with the asymmetric threshold rule in section 3.5; ensemble strategies combine base-learner predictions by voting, averaging, stacking, or rank-based aggregation before applying that rule; and the regime-aware `ML_07_regime_lgbm` strategy makes predictions with one of three sub-models, according to the regime as a rule. The full hypothesis-asset-strategy mapping is given in Appendix M, so the methodology section does not need to include the entire 180-row evidence map.

There are some implementation details that are related to the audit findings in Chapter IV. First, the H2 regime-aware strategy `ML_07_regime_lgbm` has a two-bar execution lag that is the result of a loss of signals due to double shift. The audit described it, and it is considered a conservative bias in favour of the strategy in the sense that the extra lag lowers the turnover and the costs associated with it. The H2 negative is large enough that this bias does not change the verdict. Second, in the H3 namespace, `ML_04_xgboost` and `ML_09_lstm` failed to trade for every fold and all four assets. The reason for this was an interaction between the probability threshold rule and the calibration of these particular implementations, which kept the predicted probabilities under the entry threshold. Both are classified as `FAILED` rather than `rebuilt`, in order to report the raw evidence from the pre-specified design; failure thins the H3 ensemble's base-learner diversity pool and appears in the H3 audit caveats. All of this has no impact on the active H1/H2 strategies `ML_04_svm` and `ML_05_gru`.

Regime-Aware ML_07 Architecture (H2 Treatment Strategy)

Static comparators (RB_01-RB_08, ML_01-ML_06) receive regime as a feature but do NOT route through per-regime sub-models



Sub-models trained on regime-labelled fold segments · Scaler fitted on training window only

Figure 3.2 ML_07_regime_lgbm Architecture

The deterministic regime classifier routes each bar to one of three per-regime LightGBM sub-models (Bull, Range, Bear). Each sub-model is trained on regime-labelled fold segments only. A probability threshold gate (0.55 long / 0.45 short) converts the selected sub-model's output to a $\{-1, 0, +1\}$ signal. Static comparators receive regime as a feature but do not route through separate sub-models.

The architecture in the figure above makes concrete the distinction between regime-informed and regime-routing strategies that runs through the H2 experiment. Static comparators in the H2 benchmark pool, namely RB_01 through RB_08 and ML_01 through ML_06, receive the trend_regime feature as one of 35 or 39 inputs, allowing each model to learn how regime interacts with other predictors. However, they do not split their modelling architecture by regime. ML_07_regime_lgbm takes the stronger position: it trains separate LightGBM sub-models on regime-labelled training window segments and routes each live prediction to whichever sub-model matches the contemporaneous regime. The routing decision is made by the same deterministic SMA-200 slope rule which generates the trend_regime feature in stage two of the pipeline, so it carries no additional learnable parameters and adds no overfitting risk.

This architectural difference is important for interpreting the H2 result in section 4.3. Since static ML comparators already condition on regime as a feature, a positive H2 finding would have to show that each sub-model routing has something architecturally to build on instead of simply validating regime information. The H2 negative finding thus has a clear meaning: the routing architecture, implemented in the walk-forward protocol with realistic transaction costs, does not improve the drawdown control compared to a static model that only conditions on the same regime signal as a feature. Other routing architectures like HMM-based regime classifiers and regime-adaptive ensemble weighting were not tested here and may behave differently; it is commented on as item 2 in the future research agenda at section 6.6.

3.7 Backtesting Framework

Every strategy is based on a strategy-agnostic backtest engine that applies the same execution assumptions to rule-based and AI/ML strategies alike. The engine takes in a `SignalFrame`, applies a single-bar execution lag, fills in per trade at the $t+1$ open, applies commission and slippage per side and returns a continuous equity curve, a trade log and summary metrics. Position sizing is fixed-notional at \$10,000 per trade, both long and short positions are allowed, same-bar reversals on an opposite signal are allowed and no stop-loss, take-profit or time-exit rules apply. The exit is always an opposite signal. This uniform engine is one of the core ingredients of our study to protect the comparability.

Table 3.5 Backtest Execution Assumptions and Cost Scenarios

Parameter	Value	Notes
Initial capital	\$10,000	Reset per strategy per fold
Position sizing	Fixed notional \$10,000 per trade	Creates implicit leverage after drawdowns; disclosed limitation
Base commission	6 bps per side	12 bps round-trip
Base slippage	2 bps per side	Applied uniformly to all fills
Funding rate	Not modelled	Perpetual futures funding cost omitted; disclosed limitation
Stop-loss / take-profit	Disabled	Exit only on opposite signal
Cost stress scenarios	6 scenarios from (6 bps fee + 5 bps slip) up to (12 bps fee + 10 bps slip)	Applied as robustness check; see Table 3.7

Parameter	Value	Notes
Execution lag	Single-bar ($t \rightarrow t+1$ open)	Anti-leakage at the execution layer

One caveat on implicit leverage: with fixed-notional sizing, the \$10,000 notional can be more than $1\times$ as much as equity after a drawdown. This is a known limitation in this thesis. Fixed-notional sizing was kept because the alternative of sizing as a fraction of equity brings with it a path-dependent feedback loop that makes cross-strategy comparison harder.

The only backtest engine employed by all of the strategy implementations makes the same execution assumptions at the code level. That uniformity is a methodological strength: any difference in reported performance can be owed to the strategy signals as opposed to some difference in the execution model. What follows describes the engine a little more than Table 3.5 does.

Signal-to-fill conversion is uniform: a signal at the open of bar t fills at the open of bar $t+1$. That one-bar execution lag is the primary anti-leakage control at the execution layer, it guarantees no strategy can profit from information that would not have existed at the moment the signal fired. Position sizing is fixed-notional at ten thousand dollars per trade, applied the same way to rule-based and ML strategies. The fixed-notional choice is a known limitation that magnifies the influence of outlier folds, a position with a thousand dollars of remaining equity trading at ten thousand notional carries implicit leverage of ten times, but it has a compensating advantage: performance metrics that compare directly across strategies, with no path-dependent feedback from prior performance affecting the comparison.

The cost application itself is uniform as well: six basis points of commission and two basis points of slippage per side charged upon fill. The slippage value is relatively conservative compared to the published cryptocurrency microstructure literature on large-cap assets and bar size aggregated orders but not taking into account position in the queue and the corresponding price impact which will kick in for larger trades. Cost stress testing takes place through six additional cost scenarios described in Table 3.5.

The position management policy remains straightforward: an opposite signal will close the current position and open a new one in the same bar; reverse trading in the same bar is allowed;

there are no stop loss, take profit, time exit or trailing stops rules. Ignoring stop losses is an intentional decision. A stop loss introduces a non-trivial parameter – stop loss level – that requires validation and complicates the analysis due to interaction with the signal logic. This way, by leaving stop losses aside, we separate the effect of the signal logic from the effect of a risk management layer. Stop loss-based variants should be easy to implement later.

The funding rates of perpetual futures are not accounted for in the backtesting engine which is a limitation. Perpetuals pay or receive funding at periodic intervals, set by the premium of the perpetual price over the index. Across the four-year window the cumulative funding cost on a permanent long in BTCUSDT has been substantial, so leaving it out biases the reported equity curves of long-bias strategies, buy-and-hold (RB_01) especially, upward. The H1 evidence synthesis names this bias as one of the factors inflating the rule-based mean Sharpe advantage. Adding funding rate modelling is identified as a high-priority future direction in Chapter VI.

3.8 Performance Metrics

The Sharpe ratio originates from Sharpe’s performance measure, although the present study uses a modified CAGR-based implementation for internal consistency across strategy comparisons (Sharpe, 1966). For H1 and H3 the primary metric is the Sharpe ratio, defined here as the compound annual growth rate over the annualised standard deviation of log returns, with annualization using a factor of $\sqrt{(24 \times 365)} \approx 93.58$ for 1-hour bars. The formula remains internally consistent across all the strategy comparisons, but it departs from the academic convention of $\text{mean}(r) / \text{std}(r)$, so it is not directly comparable with Sharpe ratios in external literature that use the standard form, a disclosure repeated wherever Sharpe appears. For H2 the primary metric is the maximum drawdown of the per-fold equity curve, expressed as a percentage of the rolling peak.

Table 3.6 Performance Metrics Definitions

Metric	Definition / Formula
Sharpe (this study)	CAGR / annualised standard deviation of log returns
Sortino	CAGR / downside deviation of log returns
Calmar	CAGR / max drawdown

Metric	Definition / Formula
Maximum drawdown	Peak-to-trough percentage decline of the equity curve
Total return	Equity curve final value / starting value – 1
Win rate	Profitable trades / total trades
Profit factor	Gross profit / gross loss
Trade count	Total executed trades in the fold
Time in market	Fraction of bars with an open position
CAGR	Compound annual growth rate of the equity curve

3.9 Walk-Forward Validation

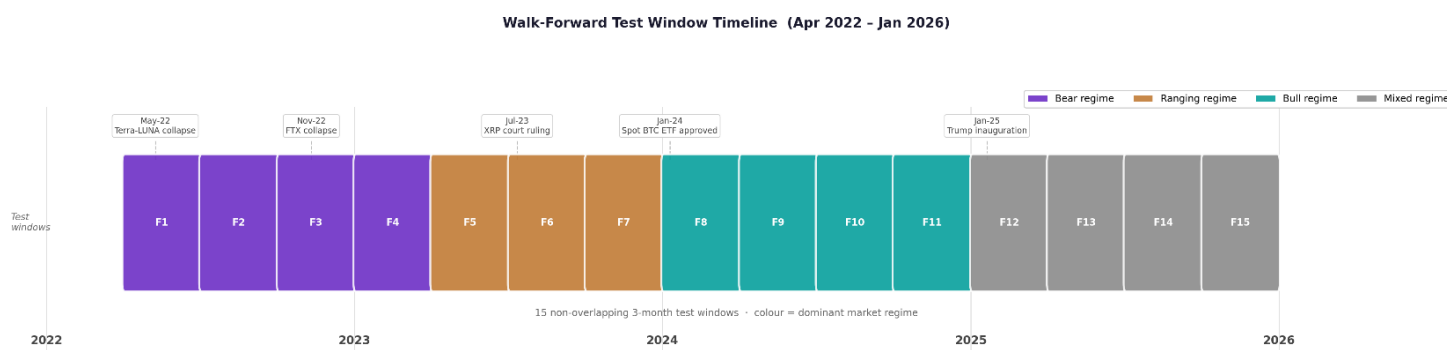


Figure 3.3 Walk-Forward Test Window Timeline

Walk-forward test window timeline (Apr 2022 – Jan 2026). Fifteen non-overlapping 3-month test windows colour-coded by dominant market regime: Bear (F1–F4), Ranging (F5–F7), Bull (F8–F11), Mixed (F12–F15). Key market events marked.

The fifteen test windows in the figure above span four qualitatively distinct market regimes within the evaluation period. The 2022 bear-market folds (F1–F4) are bounded by the Terra-LUNA collapse in May 2022 and the FTX failure in November 2022. Ranging conditions (F5–F7) cover the 2023 consolidation period through the XRP court ruling in July 2023. Bull-trend folds (F8–F11) begin with the spot Bitcoin ETF approval in January 2024 and extend through the SOL and XRP bull-run episodes that produce the extreme outlier Sharpe values discussed in section 4.2.5. The final four folds (F12–F15) cover the 2025 mixed regime following the Trump administration inauguration. This distribution matters for interpretation: bear-trend folds are over-represented relative to bull-trend folds, which bears directly on the H1 sub-finding about AI/ML bear-trend performance discussed in section 4.2.5.

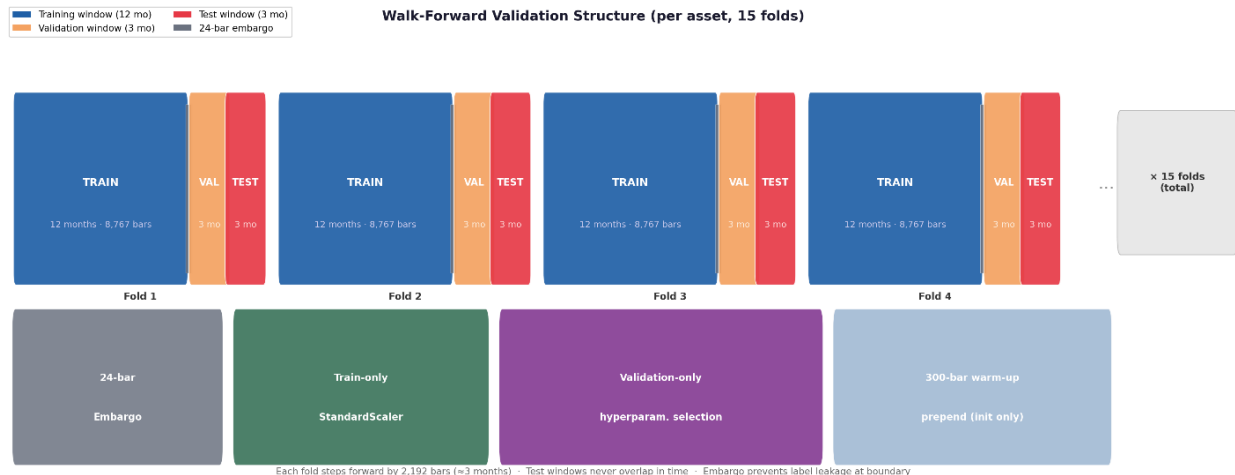


Figure 3.4 Walk-Forward Validation Structure

Top: single-fold anatomy showing training window (12 months, 8,767 bars), 24-bar embargo, validation window (3 months) and test window (3 months), repeated across 15 folds. Bottom: four anti-leakage controls applied at every fold boundary.

The fold anatomy shown above enforces four anti-leakage controls at every fold boundary. The 24-bar embargo prevents labels computed at the end of the training window from bridging into the validation window. The StandardScaler is fitted on training data only and frozen before any contact with validation or test partitions, ensuring no test-period distributional information reaches the model. Hyperparameter selection is based on validation Sharpe only; the selected configuration is then fixed before the test evaluation begins. The 300-bar warm-up prepend gives indicator calculations a full initialisation period without contributing to the performance measurement window. A signal at bar t fills at the open of bar $t+1$ and there is a one-bar lag at the fill level. These controls combine to eliminate the main leakage paths that inflate naive ML backtests like in section 2.10.

Each asset is split into fifteen non-overlapping walk-forward folds over the common window. A fold runs a training window of 8,767 bars (about 12 months), a validation window of 2,192 bars (about three months), and a test window of 2,192 bars (about three months). The step between folds is 2,192 bars, so the test windows never overlap in time. A 24-bar embargo separates training labels from the validation/test boundary, and a 300-bar warm-up window is prepended to the validation and test partitions for indicator initialisation but kept out of performance evaluation. The StandardScaler in the ML preprocessing pipeline is fitted on training data only, then applied to validation and test data without refitting. As a result, these controls block the

major channels through which naively conducted machine learning back-tests of trading rules are quietly tainted by corruption, ensuring the proper chronological order required by Bergmeir and Benítez (2012) for valid cross validation in time-series settings.

3.10 Statistical Testing

Hypothesis testing combines parametric, non-parametric, and bootstrap methods. For each strategy-group comparison and each metric, the statistical engine runs a normality test (Shapiro-Wilk for $n \leq 50$ and D'Agostino-Pearson for $n > 50$) to guide test selection and a Mann-Whitney U test for rank-order comparison, a Welch t-test for mean differences in cases of unequal variances and a bootstrap confidence interval with 1,000-5,000 samples and a random seed of 42. The Holm-Šidák stepdown procedure is used for multiple-comparison control. The effect size is Cohen's d (for parametric tests), Cliff's delta (for non-parametric tests) and rank-biserial correlation r (for non-parametric tests). For paired analysis for $n = 4$ assets, such as the H3 pooled Wilcoxon test, the smallest two-sided p-value we can obtain is 0.0625, and this test is seen to be indicative of directional consistency rather than significance.

Bootstrap inference is the primary statistical procedure for H1 (mean Sharpe difference) and H3 (median Sharpe difference). It runs 1,000 resamples for asset-level analyses and 5,000 for pooled ones, and has a random seed of 42 to ensure reproducibility. In both cases, the resampling is used at strategy-fold level (not on the bar level), so the within-fold dependence in returns is preserved. H1 also reports a clustered bootstrap variant where the resampling unit is the asset rather than strategy-fold; this variant results in a wider confidence interval, which highlights how little diversification a four-asset universe provides.

We use block bootstrap to test the Monte Carlo robustness. The block size is 5 days (120 hourly bars), large enough to ensure that the crypto returns retain serial correlation at the autocorrelation horizons reported in the literature but small enough to make the resampling meaningful in terms of path diversity. We generate one thousand block bootstrapped paths for each strategy per fold and summarise the resulting Sharpe distribution with its 5th, 25th, 50th, 75th and 95th percentiles. The Monte Carlo p95 Sharpe statistic in Table 4.6 is the 95th percentile of that block bootstrapped Sharpe distribution (the null of no signal is not recorded).

Multiple-testing correction is applied selectively. Within a single hypothesis, the per-asset bootstrap tests are corrected with the Holm (1979) stepdown procedure (Holm-Šidák version), which controls the family-wise error rate without Bonferroni's heavy hand. Across hypotheses, no correction is applied, since H1, H2 and H3 are pre-specified, distinct claims, not members of a single data-snooping exercise. This is consistent with López de Prado's (2018) suggestion that post-hoc multiple-testing corrections should be avoided for pre-registered, distinct hypotheses.

3.11 Robustness Testing

Our robustness testing is based on a five-stage methodology: walk forward analysis as the main out-of-sample protocol; a Monte Carlo block bootstrap with 5-day blocks and 1,000 paths; six-scenario cost stress testing as described in Table 3.5; regime breakdown performance reported separately for bear, ranging, and bull regimes and leave-one-asset-out sensitivity analysis for any cross-asset claim. In H1, H2 and H3, we treat only the primary claim as thesis-level until it is able to survive both cost stress and leave-one-asset-out perturbation and report any claim that does not, with the caveats stated. This is consistent with Bailey et al. (2014) standard for distinguishing real trading performance from backtest artefacts.

Table 3.7 Statistical Tests and Robustness Checks Applied

Component	Method	Purpose
Primary test	Bootstrap CI (n = 1,000–5,000, seed 42)	Estimate the distribution of the mean / median metric delta
Secondary test	Welch t-test (unequal variances)	Mean-difference test, parametric
Secondary test	Mann–Whitney U / Wilcoxon signed-rank	Rank-order test, non-parametric
Effect size	Cohen's d, Cliff's delta, rank-biserial r	Magnitude of difference, distribution-agnostic
Walk-forward analysis	15 rolling folds per asset	Out-of-sample stability
Monte Carlo	Block bootstrap (5-day blocks, 1,000	Distributional uncertainty of the OOS

Component	Method	Purpose
	paths)	Sharpe estimate
Cost stress	6 alternative cost scenarios	Sensitivity to fee + slippage assumptions
Regime breakdown	Bear, ranging, bull regimes	Conditional performance evidence
Leave-one-asset-out	Drop each asset in turn from the pool	Sensitivity to single-asset influence
Multiple-testing control	Holm–Šidák stepdown	Family-wise error rate control

3.12 Auditability and Reproducibility

Every experiment run generates a manifest: the experiment configuration hash, the git commit hash, software versions, strategy identifiers, evaluation period, random seeds and all the reported metric values. Each hypothesis series then goes through an independent audit, data quality, leakage controls, statistical methodology, claim traceability, disclosure completeness, and so on, before it can be used in the thesis. The audit hierarchy is crystal clear: per-asset reports are the evidence base, pooled multi-asset reports are the primary evidence base and the audited cross-hypothesis synthesis is the definitive proof. Audit findings bind. Any issue found by audit has to be part of the matching research claim. At the point of finalisation, H1 had passed audit with all twelve issues closed, H2 had been formally accepted for thesis use under the H2 pooled strict audit report, and H3 had passed after remediation under audit identifier AUD_POOL_20260515_001. The cross-hypothesis synthesis (SYN_H1H2H3_20260515_001) had passed audit with all five issues closed.

3.13 Methodological Limitations

Several methodological limitations are documented up front, not added afterward as caveats. First, the Sharpe ratio uses a non-standard formula ($CAGR / \text{annualised log-return volatility}$) rather than the academic $\text{mean}(r) / \text{std}(r)$ convention, internally consistent, but not directly comparable to external literature. Second, perpetual futures funding costs are not modelled in the backtest engine, a material omission for high-drawdown strategies. Third, fixed-notional position sizing creates implicit leverage after drawdowns and amplifies the influence of extreme outliers in the pooled analyses. Fourth, BTC has four open-interest-derived features that ETH, SOL, and

XRP lack, which introduces a mild model-capacity asymmetry. Fifth, two intended H3 ML strategies (ML_04_xgboost and ML_09_lstm) produced zero trades across all folds and assets and are classified as failed; that reduces the H3 ensemble base-learner pool from five intended models to three active ones, and it does not affect the active H1/H2 implementations ML_04_svm and ML_05_gru. Sixth, the ML_07 regime-aware LightGBM strategy in H2 carries a documented two-bar execution lag from a double-shift in its signal pipeline, which biases its reported performance conservatively downward. Each limitation is referenced in the relevant results section in Chapter IV and consolidated in the limitations discussion in Chapter VI.

3.14 Ethical Considerations

The research uses only historical market data obtained from a public exchange interface, no human subjects, no personally identifying information. It is not investment advice, and it makes no recommendation to any investor, or class of investors, to deploy any of the tested strategies. Reproducibility and transparent reporting of negative results are treated as ethical obligations here: null findings are reported with the same rigour as positive ones, and every audit caveat is disclosed in the chapter where the matching claim is made. The thesis does not attempt to influence cryptocurrency prices or to enable market manipulation, and the publication date is chosen so that every reported strategy detail refers only to historical evaluation.

CHAPTER IV: RESULTS

4.1 Introduction

This chapter reports the empirical results of the three hypothesis experiments and integrates them into a cross-hypothesis synthesis. Section 4.2 reports the H1 experiment, AI/ML strategies against rule-based benchmarks on a Sharpe-ratio basis. Section 4.3 reports H2, a regime-aware AI/ML strategy against a static benchmark pool on maximum drawdown. Section 4.4 reports H3, ensemble AI/ML strategies against single-model AI/ML strategies on Sharpe stability and consistency. Section 4.5 reports the audited cross-hypothesis synthesis and the verdict on the overall thesis claim. Throughout, asset-level evidence counts as supporting, pooled multi-asset evidence as primary, and the audited cross-hypothesis synthesis as the authoritative source of truth. Figure 4.1 below summarises the research framework behind all of these results.

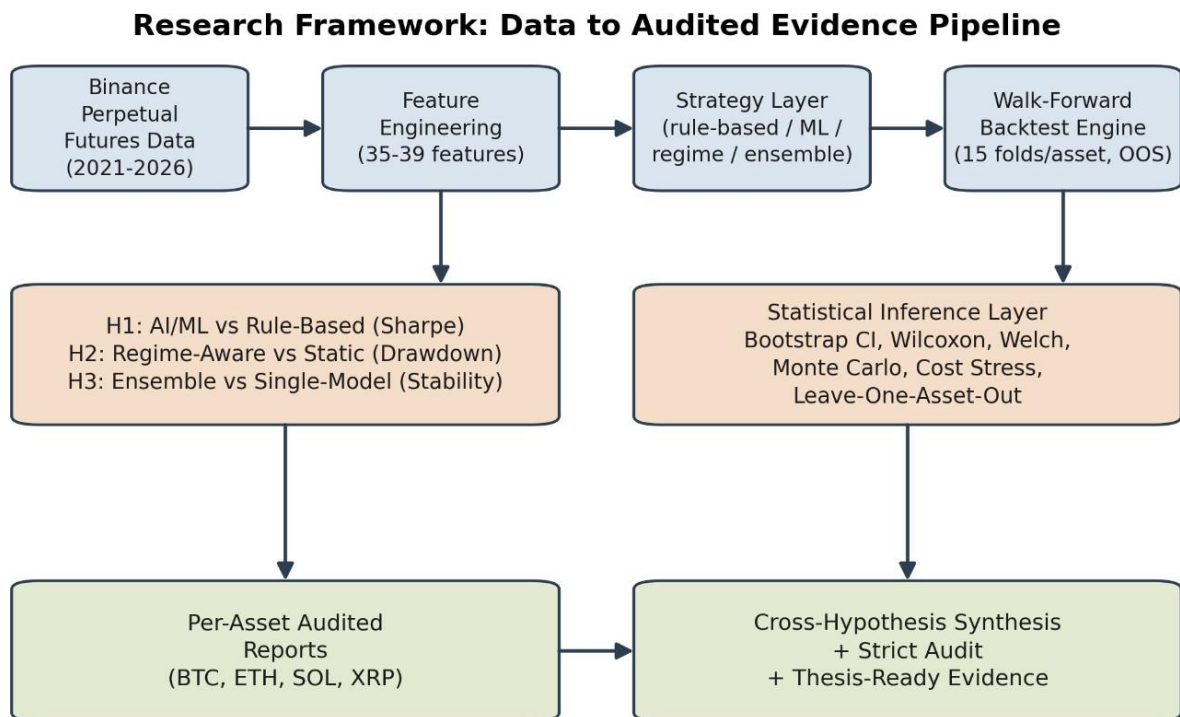


Figure 4.1 Research framework: data to audited evidence pipeline.

4.2 H1: AI/ML versus Rule-Based Strategies

4.2.1 H1 Statement and Objective

H1 states that AI- and ML-based trading strategies generate statistically significant superior risk-adjusted returns, measured by the Sharpe ratio, against traditional rule-based benchmark strategies in selected cryptocurrency markets. The objective is to test whether the ML strategies, drawn from a portfolio of supervised classifiers trained on engineered features, beat a benchmark pool of eight rule-based families on fold-level Sharpe, both at the asset level (BTC, ETH, SOL, XRP) and in a pooled multi-asset analysis spanning 840 strategy-fold observations.

4.2.2 H1 Experimental Design

Six primary AI/ML strategies are compared against the eight-family rule-based pool. The AI/ML primary set is ML_01 logistic regression, ML_02 random forest, ML_03 gradient boosting, ML_04 support vector machine, ML_05 GRU recurrent model, and ML_07 regime-aware LightGBM. ML_06_ensemble is kept as supplemental evidence but left out of the primary H1 statistical tests, to avoid double-counting ensemble evidence that H3 evaluates separately. The rule-based comparators carry hyperparameter configurations chosen on each fold's validation window by best validation Sharpe, with that best-validated configuration then tested out-of-sample. The pooled primary analysis brings together 480 rule-based observations and 360 AI/ML primary observations, 840 strategy-fold observations in all. When ML_06_ensemble is added as supplemental evidence, the H1 evidence file holds 900 fold-level observations.

4.2.3 Asset-Level Results

Asset-level results appear in Table 4.1 and Figure 4.2. At the strategy-mean level, rule-based strategies perform better in three of the four assets, BTC being the near-parity exception. At the median level, the pattern reverses, AI/ML strategies lead in all four assets. And all four asset-level H1 experiments independently produce a verdict of "not supported," because the primary bootstrap criterion is not met.

Table 4.1 H1 Asset-Level Sharpe Comparison (Pooled by Asset)

Asset	Mean Sharpe (RB)	Mean Sharpe (ML)	Median Sharpe (RB)	Median Sharpe (ML)	Bootstrap p	Verdict
BTC	-1.146	-1.120	-1.560	-1.177	0.438	Not supported
ETH	-0.603	-0.660	-1.073	-0.840	0.546	Not supported
SOL	+3.202	+0.412	-0.672	-0.366	0.740	Not supported (outlier-driven)
XRP	+3.491	-0.212	-0.716	-0.463	0.912	Not supported (outlier-driven)

Two exceptional outlier folds drive the SOL and XRP mean Sharpe figures. SOL Fold 6 (October 2023 - January 2024) exhibits a buy-and-hold out-of-sample Sharpe Ratio of 364.41 for a bullish movement in the price of SOL. Likewise, XRP Fold 10 (October 2024 - January 2025) has an MA-Crossover Sharpe Ratio of 331.86 during the rally in XRP due to the regulatory resolution. In between, those two folds are approximately 0.24% of the 840 total primary H1 observations. Excluding those, the SOL and XRP Sharpe ratios from rule-based mean turn negative at about -0.40 and -0.26.

H1: AI/ML vs Rule-Based Sharpe Comparison Across Assets

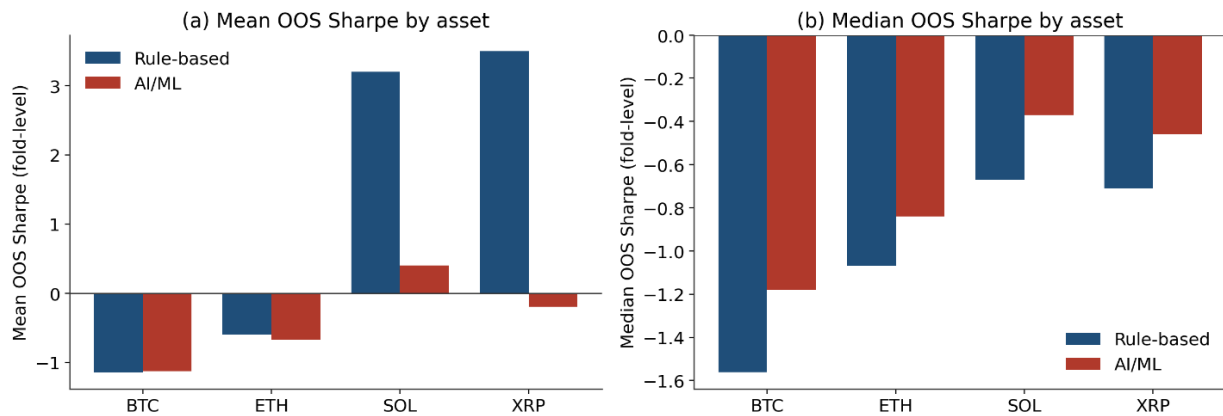


Figure 4.2 H1: AI/ML versus rule-based Sharpe across the four assets, mean panel (a) and median panel (b).

Per-Asset Notes

BTC: The ML and rule-based mean Sharpe figures are within 0.03 of each other, both negative. The Mann-Whitney U test is significant at $p = 0.007$ (Holm-corrected $p = 0.021$) and says that the ML distribution does not yield the worst fold-level results, while the bootstrap and Welch test are not significant. So BTC returns a mixed result, rank-statistically significant, with a small Cohen's d (0.022) and is not supported by the primary mean-Sharpe criterion. Buy-and-hold is the best rule-based performer (+2.818) and well above the best ML mean (−0.677).

ETH: ML and rule-based mean Sharpe are essentially level (−0.660 vs −0.603), with the median favouring ML by 0.232. The maximum drawdown is significantly smaller for rule-based strategies (23 per cent vs 47 per cent) which reflects the higher turnover and exposure variance of the ML strategies themselves. Bootstrap $p = 0.546$.

SOL: The SOL fold 6 outlier dominates everything; take it out and ML will actually have a higher mean Sharpe than rule-based. ML wins the bear-trend regime significantly (+8.36 vs −0.16) and the 2022 bear-market sub-period (+2.838 vs +0.039), but loses the bull-trend regime, where buy-and-hold just rides the SOL rally. ML's mean maximum drawdown is more than double rule-based's (69 vs 33 per cent).

XRP: A similar trend as SOL: rule-based mean Sharpe affected by the XRP fold 10 outlier, ML median Sharpe higher. But unlike BTC, ETH and SOL, XRP does show no clear ML bear-trend advantage. The bear-fold sample is too small to trust here, so the XRP regime result is unreliable (+0.570 vs RB +1.381 just reflects the small bear-fold sample). The best individual rule-based strategy is moving average crossover that catches the XRP rally; the best ML strategy is ML_07_regime_lgbm at +0.098 mean Sharpe.

4.2.4 Pooled H1 Results

The pooled multi-asset analysis stacks the fold-level observations across all four assets and applies a set of statistical tests to them, to see whether AI/ML strategies generate significantly superior Sharpe ratios in aggregate. Table 4.2 summarises the main test statistics.

Table 4.2 H1 Pooled Statistical Evidence (4 assets, 840 strategy-fold observations)

Test	Statistic	p-value	95% Confidence Interval	Significant?
Bootstrap (n = 5,000)	diff = -1.594	0.9548	[-3.946, +0.130]	No
Mann–Whitney U	U = 99,258	0.0001	N/A	Yes (rank-order only)
Welch t-test (unequal var)	t = -1.482	0.9305	N/A	No
Clustered bootstrap	diff = -1.594	0.8852	[-4.807, +0.258]	No
Fixed-effect meta-analysis	est = -0.047	0.592	[-0.444, +0.350]	No
Stouffer combined-p (4 assets)	N/A	0.836	N/A	No

The pooled mean Sharpe is +1.236 for the rule-based group and -0.358 for the AI/ML primary group, a mean delta of -1.594. The bootstrap 95 per cent confidence interval, [-3.946, +0.130], straddles zero while sitting almost entirely in negative territory, which gives no statistical basis for rejecting the null of no AI/ML advantage. Cohen's d for the mean comparison is -0.090 (a small negative effect); rank-biserial r is +0.149 (a small positive rank advantage for ML). The Mann–Whitney U result ($p = 0.0001$) confirms ML strategies rank higher on average, but H1 is specified in terms of mean Sharpe superiority, and there the evidence is not present.

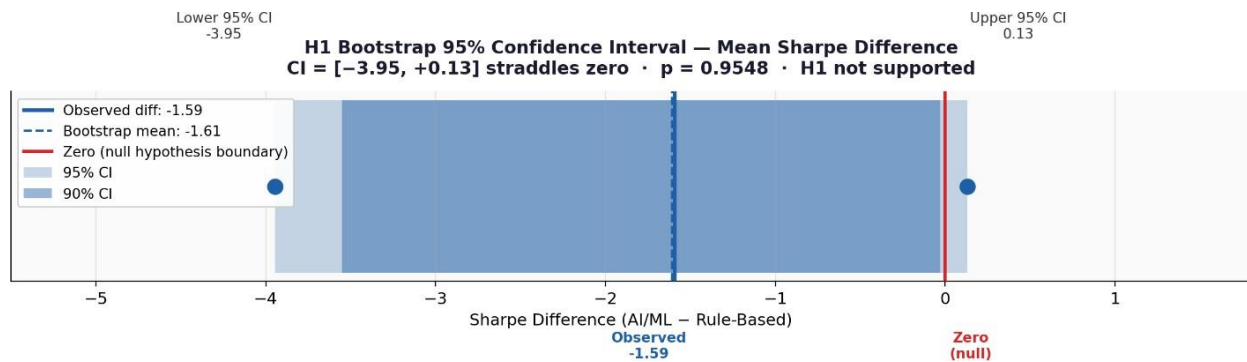


Figure 4.3 H1 bootstrap 95% confidence interval for mean Sharpe difference (AI/ML – Rule Based)

Difference = -1.59. CI = [-3.95, +0.13]. Upper bound slightly positive, due to extreme variance in the SOL fold 6 and XRP fold 10 outlier folds. Bootstrap one-sided p = 0.9548.

The confidence interval in the figure above places the numerical data in Table 4.2 in a visual representation. The mean Sharpe difference of -1.59 lies close to the centre of the bootstrap distribution, and the 95% confidence interval [-3.95, +0.13] is nearly entirely negative. The zero line is crossed only slightly inside the upper bound. That the upper bound is barely positive is not indicative of just a small advantage of AI/ML but because the rule-based mean is very close to zero with the two outlier folds, it is impossible to exclude the null from the bootstrap CI. The bootstrap difference of -1.61 is an excellent match to the observed value; it proves that the bootstrap is centred. The width of the interval of nearly four Sharpe units is similar to the volatility concentration problem that the leave-one-asset out analysis in Figure 4.5 explores in more detail.

4.2.5 Regime and Cost-Stress Breakdown

Table 4.3 reports mean Sharpe by market regime. The most striking finding is the bear-trend advantage of AI/ML: in bear-trend folds, AI/ML strategies achieve mean Sharpe of +2.25 against -0.18 for rule-based strategies, a finding that is consistent across BTC, ETH and SOL. In bull-trend folds, the rule-based mean Sharpe of +24.40 is inflated by the SOL and XRP outliers; without them, the bull-trend rule-based advantage moderates substantially. In ranging conditions, both groups produce small negative Sharpe means, with rule-based marginally ahead.

Table 4.3 H1 Regime Breakdown: Mean Sharpe by Market Regime

Regime	Mean Sharpe (RB)	Mean Sharpe (ML)	Direction	Notes
Bear-trend	-0.182	+2.246	ML much greater than RB	Consistent across BTC, ETH, SOL; small bear-fold sample in XRP
Ranging	+0.483	-0.565	RB marginal	Most common regime
Bull-trend	+24.404	-1.099	RB much greater than ML	Dominated by SOL fold 6 and XRP fold 10 outliers

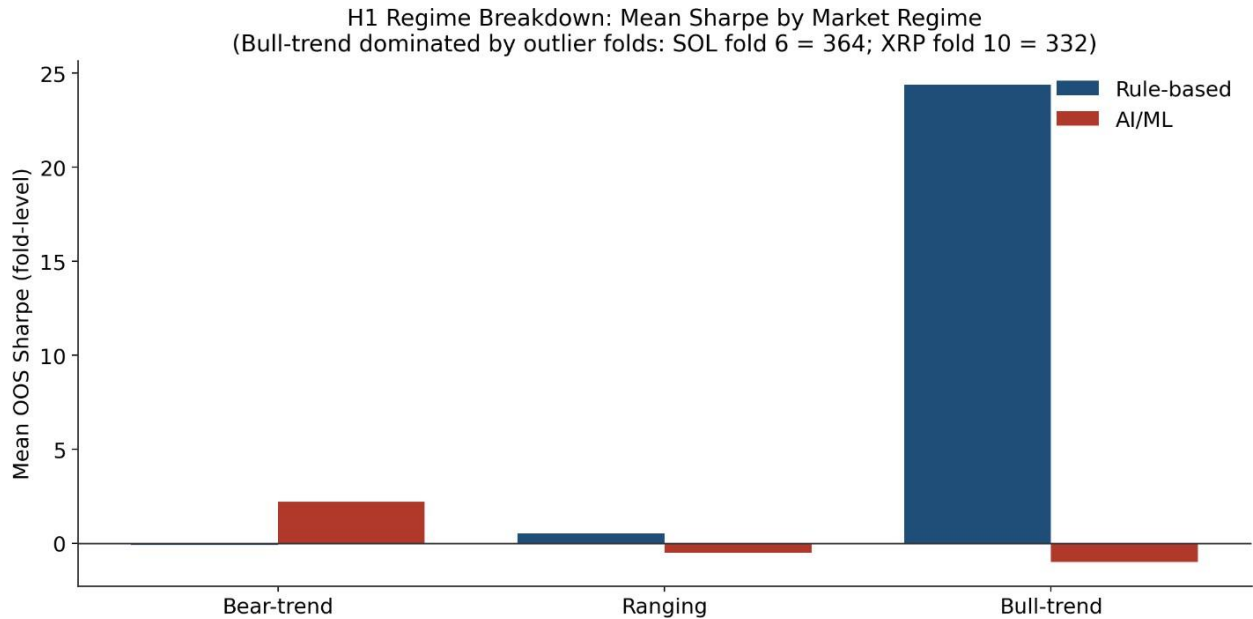


Figure 4.4 H1 regime breakdown: mean Sharpe by market regime

Cost-stress sensitivity is summarised in Table 4.4. At the pooled level, rule-based dominates in all seven cost scenarios, mostly because of the SOL and XRP outlier folds. At the ETH-only level the result inverts: AI/ML beats rule-based in the six non-base stress scenarios, reflecting ETH's lighter outlier exposure. Because AI/ML strategies trade about three times as much as rule-based ones (272 trades per fold against 85), they are systematically more cost-sensitive.

Table 4.4 H1 Cost-Stress Sensitivity Summary

Cost Scenario	Pooled Direction	ETH-Only Direction	Comment
Base (6 bps fee + 2 bps slip)	RB wins	Near-parity	Standard scenario; SOL/XRP outliers dominate pool
6 bps + 5 bps slippage	RB wins	ML wins	ML more cost-sensitive globally; ETH local exception
6 bps + 10 bps slippage	RB wins	ML wins	Same pattern
6 bps + 20 bps slippage	RB wins	ML wins	Extreme slippage stress
9 bps + 2 bps slippage	RB wins	ML wins	Higher fee tier
12 bps + 2 bps slippage	RB wins	ML wins	Top of fee schedule
12 bps + 10 bps slippage	RB wins	ML wins	Compound stress

The Monte Carlo block bootstrap, with 5-day blocks and 1,000 paths, moves both groups to deeply negative mean Sharpe, neither strategy class survives trade-return resampling. Only rule-based buy-and-hold clears a positive Monte Carlo $P(\text{Sharpe} > 0)$, at roughly 80 per cent, which is just the long-run drift in cryptocurrency prices over the window being reflected. Leave-one-asset-out analysis then exposes how fragile the rule-based advantage is: dropping SOL cuts the rule-based mean advantage by 62 per cent, dropping XRP by 44 per cent. The "not supported" verdict for H1 holds in every leave-one-asset-out scenario, but the size of the rule-based advantage is not stable.

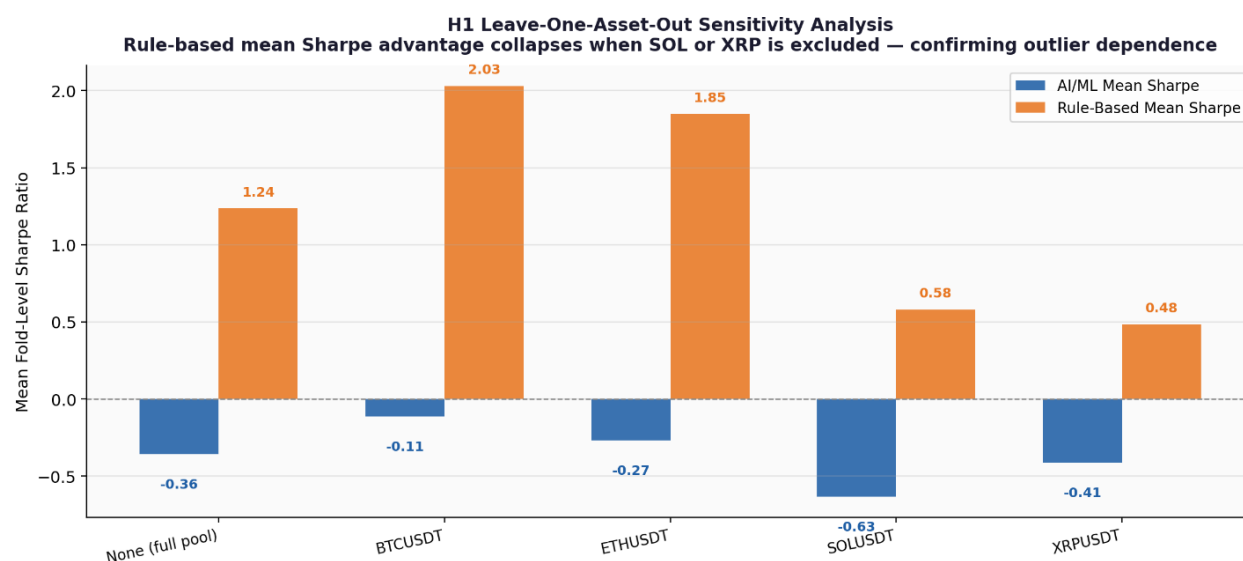


Figure 4.5 H1 leave-one-asset-out sensitivity

Pooled AI/ML (blue) and rule-based (orange) mean Sharpe when each asset is excluded in turn. The rule-based mean Sharpe advantage falls from +1.24 (full pool) to approximately +0.24 when SOL is excluded and +0.32 when XRP is excluded, confirming the pooled rule-based advantage is concentrated in two bull-run outlier folds rather than distributed across the asset universe.

The leave one asset analysis makes the rule-based advantage very clear. In the total set, the difference between the rule-based Sharpe ratio (+1.24) and AI/ML Sharpe ratio (-0.36) is huge. Leaving out BTC or ETH does not change much since none of these assets represents the outlier folds. Leaving out SOL narrows down the difference quite significantly; XRP narrows the difference by about the same amount. The implication is that the pooled H1 result is not distributional noise: it is structurally dependent on two specific market episodes in the bull-trend period. This does not change the H1 verdict, which is not supported by the evidence regardless.

But it does establish that the rule-based mean Sharpe advantage, taken at face value, would mislead a reader into thinking rule-based strategies consistently outperform AI/ML strategies across diverse market conditions and assets.

4.2.6 H1 Audit Observations

The pooled H1 audit (POOL_20260513_162408_63acb78) returned PASS, with all twelve identified issues closed by 13 May 2026. The main observations: (i) the study's Sharpe formula is non-standard and must be disclosed wherever Sharpe appears; (ii) perpetual futures funding cost is not modelled, which understates the true holding cost for high-drawdown strategies; (iii) RB_01 buy-and-hold sits inside the rule-based group rather than standing alone as a benchmark, which inflates the rule-based mean Sharpe in bull-trend folds; (iv) the SOL fold 6 and XRP fold 10 outliers have to be reported alongside any pooled mean-Sharpe comparison; and (v) the BTC experiment predates the strict audit protocol and is classified as eligible with caveats rather than fully strict-audited.

4.2.7 H1 Hypothesis Decision

H1 is NOT SUPPORTED across the selected cryptocurrency markets. Three of the ten criteria are met for primary comparison. The null hypothesis is not rejected in both bootstrapping test and Welch t-test. The rule-based mean Sharpe advantage appears to be very weak since it relies upon two extreme outliers. However, there are two secondary findings that hold true: AI/ML strategies have a higher median Sharpe for all four assets; and, AI/ML strategies exhibit a bear trend advantage.

4.3 H2: Regime-Aware versus Static Strategies

4.3.1 H2 Statement and Objective

H2 states that regime-aware AI and ML trading strategies have lower maximum drawdowns and are more robust in different market conditions than static benchmark strategies. We will test whether the regime-aware ML strategy ML_07_regime_lgbm, a LightGBM with per-regime signal routing, has lower mean maximum drawdown than the static benchmark pool with 8 rule-based families and 6 static ML strategies (ML_01-ML_06) running in the same execution assumption for the four assets.

4.3.2 H2 Experimental Design

The H2 treatment is a single regime-aware strategy, ML_07_regime_lgbm. In the control group, there are fourteen static strategies, the eight rule-based families RB_01–RB_08 and six static ML strategies ML_01–ML_06. The static ML strategies have the trend_regime feature in their input matrix, but do not route predictions through per-regime sub-models. The design is to make H2 more difficult, because the benchmark itself is already partly regime-informed. The resulting analysis covers four assets and fifteen folds, meaning that 60 asset-fold pairs in total.

4.3.3 Asset-Level Results

The results at the asset level are presented in Table 4.5 and Figure 4.6. For each asset, the regime-aware approach generates substantially greater mean maximum drawdown compared to the static approach and performs worse in all fifteen folds in BTC and ETH and obviously most of the folds in SOL and XRP. H2 experiment for all four assets at the asset level results in "not supported" with large to large-magnitude effect sizes.

Table 4.5 H2 Asset-Level Drawdown and Statistical Evidence

Asset	Static Mean DD %	Regime-Aware Mean DD %	Delta (Static – RA) pp	Bootstrap p	Wilcoxon p	Cliff's delta	Verdict
BTC	29.49	67.16	–37.67	1.0000	1.0000	–0.764 (large)	Not supported
ETH	33.03	62.84	–29.81	1.0000	1.0000	–0.668 (large)	Not supported
SOL	48.51	68.85	–20.34	0.9944	0.9949	–0.438 (medium)	Not supported
XRP	39.30	70.13	–30.83	0.9998	0.9990	–0.535 (large)	Not supported

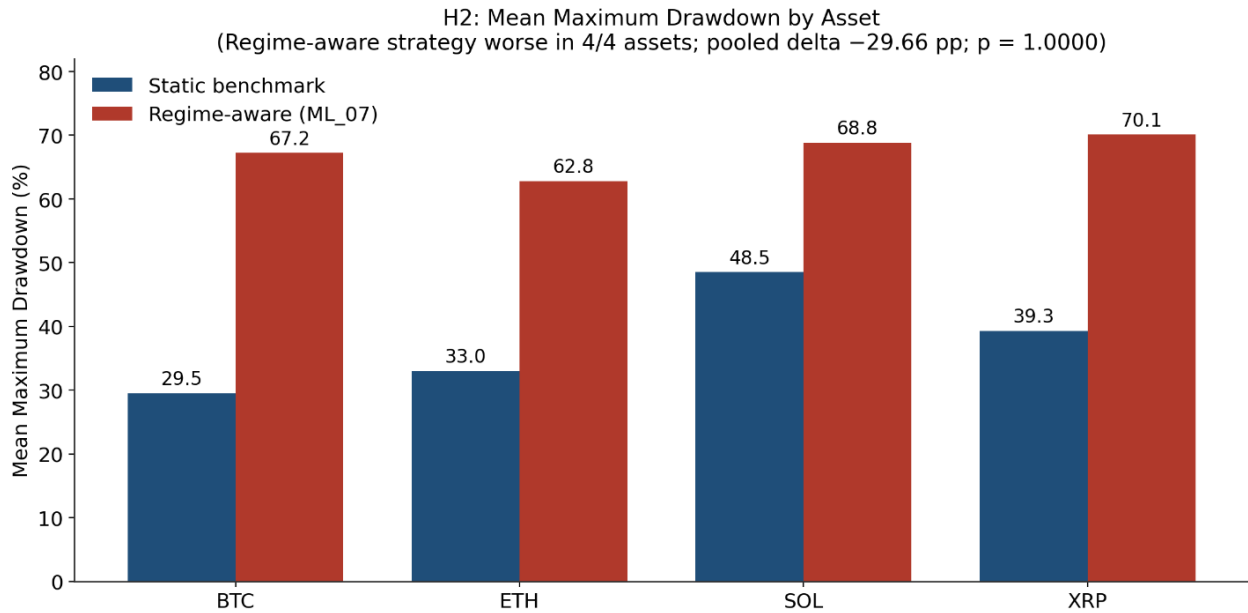


Figure 4.6 H2: mean maximum drawdown by asset (regime-aware vs static benchmark)

In BTCUSDT, the regime-aware strategy achieved higher drawdown in all fifteen folds of the walk-forward direction which shows complete directional consistency. ETHUSDT showed the same fold-level trend. SOLUSDT was slightly less extreme on effect size, however, and that is because the static benchmark has higher baseline drawdown in the volatile environment of SOL. XRPUSDT had the highest absolute regime-aware drawdown of all four, at 70.13 per cent. The ranging-regime per-regime proxy gave the regime-aware strategy a slight edge in SOL and XRP, but that regime covers less than 1.1 per cent of the out-of-sample bars and thereby has limited statistical significance.

Per-Asset H2 Interpretation

In BTCUSDT, regime-aware ML07 regime lgbm achieved larger maximum drawdown than the static benchmark pool in all fifteen folds. The mean drawdown delta of minus 37.67 percentage points is the largest in the four-asset set, and is comparable to a Cliff's delta of minus 0.764, which is a substantial effect. Fold by fold unanimity is important: the H2 negative finding for BTC is not a bad regime episode but the result of a systematic failure of the regime-routing architecture for every time period. From the per-fold equity curves, the weakness is apparent: the routing strategy is most exposed when regime boundary markers arrive at the fold boundaries. When the classifier sees a transition too late, the strategy keeps throwing out predictions from the

prior regime sub-model and is losing the same amount of time as a static benchmark would have avoided.

ETHUSDT repeats the BTC story: in all fifteen folds regime-routing had a higher drawdown, mean delta minus 29.81% and Cliff's delta minus 0.668. ETH's smaller outlier exposure (and relative low drawdown in comparison to SOL and XRP) leads to lower absolute drawdowns in both groups, though relative underperformance of regime-routing is much the same. ETH has the lowest cost stress amplification of the four, so the drawdown gap is more of a signal quality problem and not a cost sensitive one.

In SOLUSDT the regime-aware strategy lost in eleven of those fifteen folds to the static benchmark, mean delta minus 20.34% and Cliff's delta minus 0.438 (medium-large). This is not a significant difference but comes due to two forces in opposite directions. First, the static benchmark itself has a higher baseline drawdown in SOL than in the other three assets because of its volatility and outlier returns, and hence the relative difference. Second, the ranging-regime sub-model in the routing architecture performed better in SOL than in the other three assets and mitigated some of the bear and bull-regime losses.

In XRPUSDT the regime-aware strategy had the highest absolute drawdown of the four assets, 70.13 per cent on average, with a mean drawdown delta of minus 30.83 percentage points and a Cliff's delta of minus 0.535. The reason is that XRP has a very strong event-driven profile: the regime classifier missed the late-2024 regulatory resolution rally, and the strategy stayed short straight through it, producing the worst fold-level loss in the whole H2 study.

4.3.4 Pooled H2 Results

In pooled multi-asset analysis, we collect 60 asset-fold pairs and ask if the regime-aware strategy reduces mean maximum drawdown compared to the static benchmark. It is the opposite. The pooled mean drawdown is 67.25 per cent for the regime-aware strategy and 37.58 per cent for the static benchmark, a mean delta of -29.66 percentage points. In other words, the regime-aware strategy is 1.79 times worse on the first criterion. The bootstrap 95 per cent confidence interval for the drawdown delta is [-35.83, -23.48]. The bootstrap p-value is 1.0000 and the Wilcoxon signed-rank p-value is 1.0000 in the one-sided case that the regime-aware strategy reduces

drawdown. H2 is supported in only 5 of 60 asset-fold pairs (8.3 per cent), zero of four assets and zero of three market regimes. In this case, zero of the eight key verdict criteria are met.

Table 4.6 H2 Pooled Regime, Cost-Stress and Monte Carlo Evidence

Dimension	Static Benchmark	Regime-Aware (ML_07)	Comment
Pooled mean max DD %	37.58	67.25	Delta -29.66 pp; CI [-35.83, -23.48]
Pooled median max DD %	35.05	69.99	Same direction as mean
Cost-stress mean DD %	55.90	118.13	RA fails further under stress
Monte Carlo p95 Sharpe	+0.07	-6.67	Static survives in tail; RA does not
Calmar delta	reference	+5.42 in favour of static	Calmar penalises high RA drawdown
Sortino delta	reference	+1.94 in favour of static	Downside-adjusted favours static
Regime support count	N/A	0 of 3 regimes	Bear, ranging, bull regimes all against H2

4.3.5 H2 Audit Observations

The pooled H2 audit (h2_pooled_strict_audit_report.md) came back ACCEPTED FOR THESIS USE, with three principal observations. First, the static benchmark deliberately includes static ML strategies that are regime-informed (they receive trend_regime as a feature) but do not route through per-regime sub-models, a design that raises the bar, so that any positive H2 finding would have been especially defensible against it. The decisive negative result, then, is not due to a weak comparator pool. Second, ML_07_regime_lgbm carries a documented two-bar execution lag from a double-shift in its signal pipeline, which biases its reported drawdown performance conservatively downward; but the negative finding is so decisive (delta = -29.66 pp; p = 1.0000) that the defect does not affect the verdict. Third, the per-regime drawdown figures are proxy values from regime-segmented equity-curve segments, not directly comparable with full out-of-sample drawdown percentages, and are reported only as supporting evidence.

4.3.6 H2 Hypothesis Decision

H2 is NOT SUPPORTED across the selected cryptocurrency markets, with strong evidence against it. Zero of eight key verdict criteria are met. The pooled regime-aware mean maximum drawdown is 67.25 per cent compared to 37.58 per cent for the static benchmark and the bootstrap confidence interval is entirely below zero. The negative result holds across all four assets, all three market regimes, cost-stress and Monte Carlo checks and at the asset-fold level (55 of 60 folds run against H2). One important limitation on our claim is that this significant negative is for ML_07_regime_lgbm as an architecture. It does not rule out the value of regime information as model features, and it does not generalise to other regime-conditioning architectures, Hidden Markov Model classification, online changepoint detection, regime-adaptive ensemble weighting. These remain open research directions.

4.4 H3: Ensemble versus Single-Model Strategies

4.4.1 H3 Statement and Objective

H3 states that ensemble-based AI trading strategies beat single-model approaches on the stability and consistency of returns across multiple cryptocurrencies. The objective is to test whether ensemble strategies post a higher fold-level median Sharpe ratio than single-model strategies in the same asset, and whether that edge holds across all four assets. The design also reports a secondary criterion on maximum drawdown reduction, treated as a risk-management sub-finding rather than a primary support criterion.

4.4.2 Ensemble Construction and Single-Model Comparators

The H3 ensemble group is six primary ensemble strategies: ML_06A majority vote, ML_06B weighted vote, ML_06C soft vote, ML_06D stacking out-of-fold, ML_06E strategy-signal ensemble, and ML_06F rank-based ensemble. A seventh, ML_06G regime-aware ensemble, is reported as secondary evidence only, since it combines H2 regime conditioning with H3 ensemble aggregation. The single-model comparator group is six active strategies: ML_01 logistic regression, ML_02 random forest, ML_03 gradient boosting, ML_05 LightGBM, ML_07 SVM, and ML_08 MLP. ML_04_xgboost and ML_09_lstm are classified as FAILED in the H3 namespace, having produced zero trades across all folds and all four assets. That caveat is

specific to H3 and should not be confused with the active H1/H2 implementations ML_04_svm and ML_05_gru.

A naming clarification is necessary. The identifier ML_07 reports two distinct implementations: ML_07_regime_lgbm, the regime-aware LightGBM used as the H2 treatment, and ML_07_svm, a support vector machine in the H3 single-model group. They share a model number but are different strategies, so the H3 and H2 evidential scopes are not confounded by a common strategy.

4.4.3 Asset-Level Results

Asset-level results are in Table 4.7 and Figure 4.7. On the primary Sharpe criterion the pattern is consistent: ensemble median Sharpe sits below single-model median Sharpe in every asset, though the gaps vary a lot in size. On the secondary drawdown criterion the pattern is equally consistent, but in favour of ensembles, ensemble median maximum drawdown is lower than single-model median maximum drawdown in every asset, by somewhere between 4.2 and 13.7 percentage points.

Table 4.7 H3 Ensemble vs Single-Model Sharpe and Drawdown by Asset

Asset	Ensemble Median Sharpe	Single-Model Median Sharpe	Delta	Ensemble Median DD %	Single-Model Median DD %	DD Delta (pp)
BTC	-1.365	-1.266	-0.099	35.04	39.24	-4.20
ETH	-0.871	-0.778	-0.093	39.10	45.22	-6.12
SOL	-0.537	-0.344	-0.193	49.36	63.03	-13.67
XRP	-0.425	-0.425	-0.001	41.91	47.69	-5.78
Pooled	-0.704	-0.602	-0.096	40.51	46.46	-5.95

H3: Ensemble vs Single-Model Across Assets

(Sharpe: ensemble below in 4/4 assets; Drawdown: ensemble lower in 4/4 assets, pooled -5.95 pp)

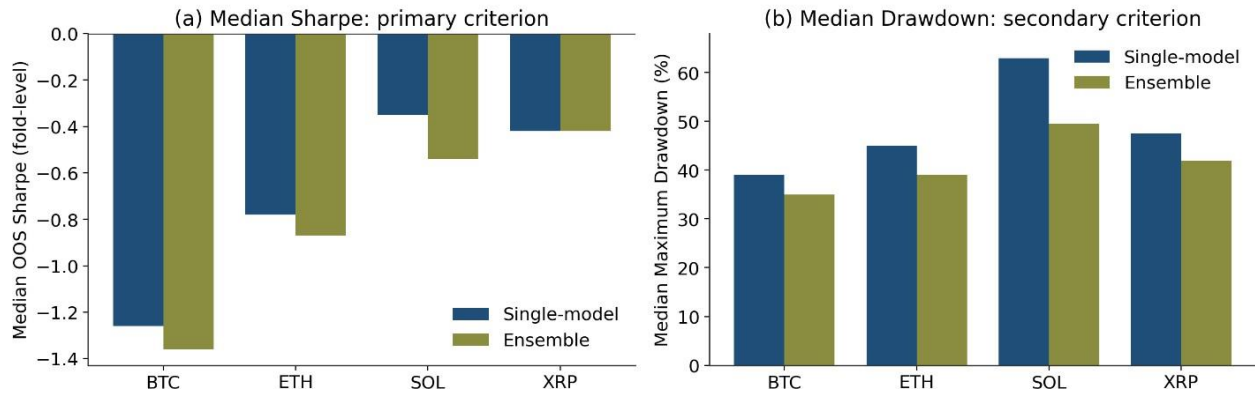


Figure 4.7 H3: ensemble versus single-model across assets: (a) median Sharpe; (b) median maximum drawdown

Per-Asset Notes

BTC: The strongest ensemble (ML_06G regime-aware ensemble, secondary only) posts a mean Sharpe of -0.609 against a single-model best of -1.126 , but the primary ensemble group, with ML_06G excluded, records a lower median Sharpe than the single-model group. Stacking out-of-fold (ML_06D) records the highest mean Sharpe within that primary ensemble group (-0.836), though on the lowest fold count, which leaves it highly uncertain. Random forest (ML_02) is the strongest single-model strategy in BTC.

ETH: Stacking OOF (ML_06D) records a mean Sharpe of $+0.058$ in ETH, slightly above zero. The primary ensemble group as a whole still records a lower median Sharpe than the single-model group. ETH is the weakest evidence against H3, since the median delta is small (-0.093) and stacking OOF comes out positive.

SOL: SOL carries the largest negative ensemble delta on the primary criterion (-0.193). But it also carries the largest ensemble drawdown reduction (-13.67 percentage points), which makes it the asset where the secondary H3 finding is strongest. The strategy-signal ensemble (ML_06E) posts a positive mean Sharpe in SOL, a sign that ensemble construction can pick up cross-strategy diversification benefits in volatile bull-market conditions.

XRP: The median value of the Sharpe delta for XRP is the lowest among all the other assets in terms of absolute value (-0.0007). The value of Cliff's delta is the only one that is positive in the

entire H3 sequence (+0.011, insignificant). Bootstrap $p = 0.500$, statistically indeterminate. XRP does not support H3, but it offers the least evidence against it.

Per-Asset H3 Interpretation

In BTCUSDT, the ensemble strategies have a median Sharpe of minus 1.365 to the single model median of minus 1.266 and a delta of minus 0.099 in favour of single-model strategies. The difference is small in absolute terms, but is consistent across the six primary ensemble variants. In the case of the single models, random forest (ML_02) is the best BTC performer, with a median Sharpe of minus 0.812 among the ensemble models; in the case of stacking out-of-fold (ML_06D) we get minus 0.836 but still very sparsely traded at about twenty-four trades per fold. For the second drawdown criterion, the ensemble strategies have a relatively small advantage, between four and five per cent, among the six primary variants.

For ETHUSDT, the ensemble median is minus 0.871 to the single model median of minus 0.778, with a delta of minus 0.093. ETH is the least convincing evidence for H3, as stacking out-of-fold results in a slightly positive mean Sharpe of plus 0.058, just over breakeven. The ETH ensemble drawdown advantage is around six percentage points, which is consistent with the cross-asset picture.

In the case of SOLUSDT, the median Sharpe delta of minus 0.193 is the largest in absolute terms among the four assets. But SOL also has the largest ensemble drawdown reduction, at minus 13.67 percentage points, and therefore is the asset where the secondary H3 finding is strongest. The strategy-signal ensemble (ML_06E) shows a positive mean Sharpe in SOL, which shows that ensemble construction can pick up cross-strategy diversification in bull-market-driven markets. That comparison, a large negative primary delta and a large positive secondary delta, is the most important single-asset finding in the whole H3 study.

In XRPUSDT, the median Sharpe delta is the smallest (at minus 0.001) and statistically indeterminate at 95 per cent. XRP does not support H3, but it also gives the least evidence against it. The drawdown reduced for the XRP ensemble of around 6 percentage points is consistent with the cross-asset pattern. That combination of a small Sharpe delta and a big drawdown reduction is consistent with what the variance reduction reading of the ensemble tells

us about the return distribution: ensembles trim the tails of the return distribution in both directions, leaving the median untouched, but pulling the maximum drawdown in.

4.4.4 Pooled H3 Results

The pooled primary analysis finds a median Sharpe of -0.704 for the ensemble compared to a median Sharpe of -0.602 for the single-model, which leads to a pooled delta of -0.0956 for the single-model strategies. The bootstrap 95 per cent confidence interval for that delta, [-0.1929, -0.0007], is entirely below 0, but its upper bound is only -0.0007 from zero, so the result should be termed borderline significant rather than strongly significant. The Wilcoxon signed-rank test on the four paired asset-level deltas gives $p = 1.0000$ for the one-sided alternative. The smallest p-value achievable with four paired assets is 0.0625, so this test is treated as evidence of directional consistency rather than a true significance test. For the 13 ranked H3 strategies, there are six primary ensembles, one secondary regime-aware ensemble, and six active single-model strategies. When we only consider the primary H3 strategies, we have six primary ensembles and six active single-model strategies. For the second objective, we use ML_06G as secondary evidence, as it combines regime conditioning and ensemble aggregation. Stacking out-of-fold (ML_06D) has the best individual median rank in the study of 2.0.

The pooled evidence is more coherent for the secondary drawdown criteria. For all four assets, the ensemble median maximum drawdown is significantly less than the single-model median maximum drawdown (pooled delta of -5.95 percentage points; 95% bootstrap CI [-13.67, -4.20]). This is the only unanimous positive cross-asset evidence observed across the AI/ML architectural enhancements tested in this thesis, but it should be interpreted as a secondary risk-management benefit rather than direct support for H3.

Table 4.8 H3 Strategy Ranking (Pooled Across 4 Assets)

Strategy	Group	Median Rank (Across 4 Assets)	Comment
ML_06D_stacking_oof	Ensemble (primary)	2.0	Best overall median rank; sparse trading (10–49 trades/fold); high uncertainty
ML_02_random_forest	Single-model	2.5	Strong cross-asset stability; top-ranked single-model comparator
ML_06G_regime_aware_ensemble	Ensemble (secondary)	4.5	Secondary only; combines regime conditioning with ensemble aggregation

Strategy	Group	Median Rank (Across 4 Assets)	Comment
ML_01_logistic_regression	Single-model	5.0	Stable linear baseline
ML_06B_weighted_vote	Ensemble (primary)	6.5	Mid-pack ensemble
ML_06E_strategy_signal_ensemble	Ensemble (primary)	7.0	Strategy-signal ensemble; cost-sensitive
ML_08_mlp	Single-model	7.0	Neural-network single-model comparator
ML_07_svm	Single-model	8.0	SVM comparator distinct from ML_07_regime_lgbm
ML_03_gradient_boosting	Single-model	8.5	Gradient-boosting comparator
ML_05_lightgbm	Single-model	9.0	LightGBM single-model comparator
ML_06A_majority_vote	Ensemble (primary)	9.0	Simple voting baseline
ML_06C_soft_vote	Ensemble (primary)	11.5	Probability-weighted voting
ML_06F_rank_based_ensemble	Ensemble (primary)	13.0	Consistently last; high-turnover ensemble

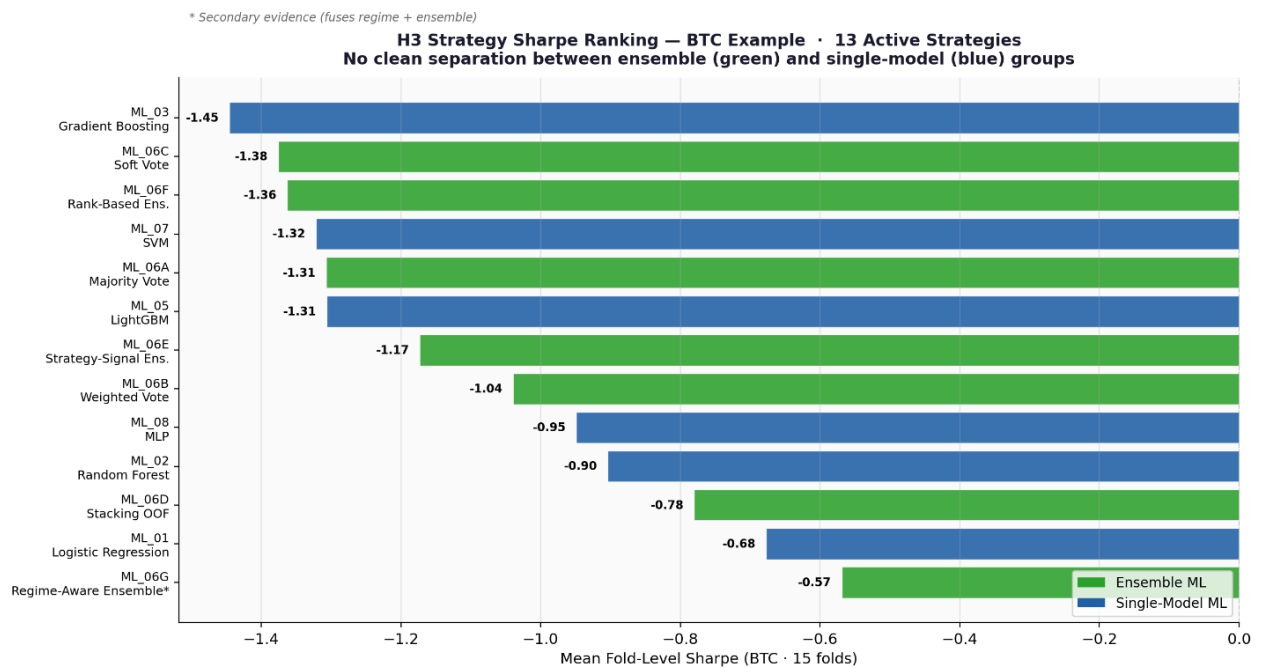


Figure 4.8 H3 strategy Sharpe ranking

The ranking chart shows the BTC-specific ordering; it differs in detail from the pooled median ranks in Table 4.8 but illustrates the same underlying pattern. The top position is held by ML_06G, the regime-aware ensemble that is secondary evidence only (it fuses the H2 regime architecture with H3 ensemble aggregation). The best primary-evidence ensemble, ML_06D stacking out-of-fold, ranks third in BTC with a mean Sharpe of approximately -0.78 , the closest

to zero in the study and the best median rank across assets. The key observation is that neither the top nor the bottom of the ranking separates cleanly along ensemble versus single-model lines: ML_01 logistic regression (single-model) sits second, and ML_06F rank-based ensemble sits last. Ensemble membership is not a predictor of ranking position. This is the pattern the pooled bootstrap formalises: the median Sharpe delta of -0.096 reflects a small, directionally consistent but architecturally diffuse disadvantage to ensembles on the primary criterion rather than a systematic ranking separation.

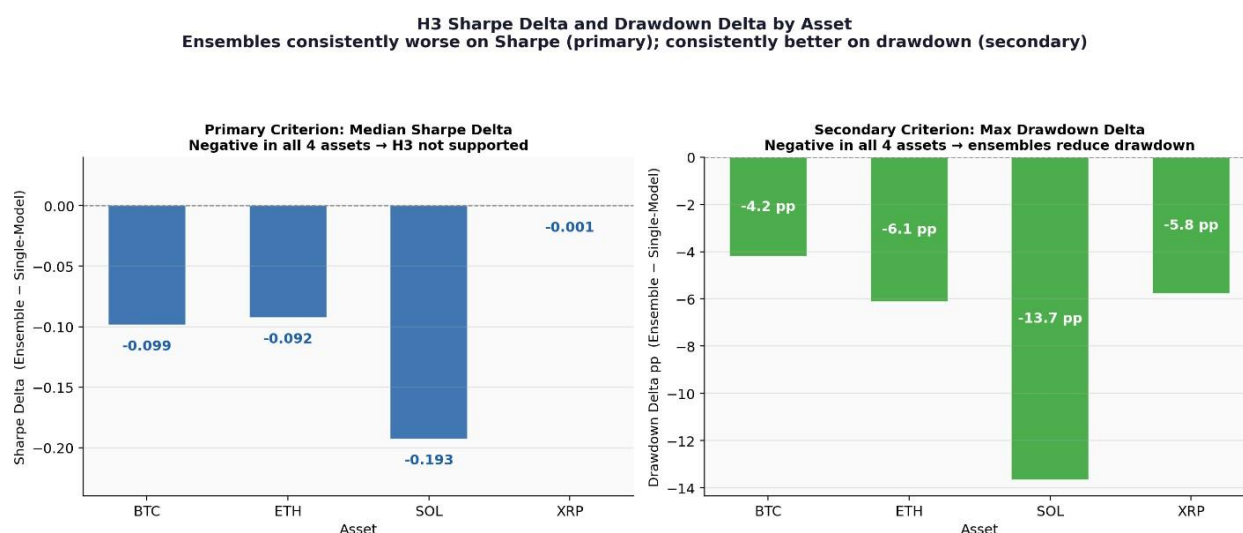


Figure 4.9 H3 per-asset deltas

Left: median Sharpe delta (Ensemble – Single-Model), negative in all 4 assets, confirming H3 not supported on primary criterion.

Right: median maximum drawdown delta (pp), negative in all 4 assets, confirming ensemble drawdown reduction is unanimous.

SOL carries the largest effect in both directions.

The left panel of the chart shows the primary criterion result broken down by asset. The median Sharpe delta is negative in all four assets: ensembles underperform single-model strategies on Sharpe in every part of the asset universe. The magnitudes range from -0.001 in XRP (essentially zero and statistically indeterminate) to -0.193 in SOL (the largest effect). The pooled estimate of -0.096 with a 95% confidence interval of $[-0.193, -0.001]$ is entirely below zero, but only marginally so at the upper end. The direction is consistent and the lower bound meaningful, but the effect size is small.

The right panel shows the secondary criterion: median maximum drawdown delta is negative in all four assets without exception. Ensemble strategies consistently produce lower maximum

drawdown than single-model strategies, with the pooled estimate of -5.95 percentage points and a 95% confidence interval of $[-13.67, -4.20]$ entirely below zero. SOL contributes the largest drawdown reduction (-13.67 pp) while also carrying the largest primary-criterion shortfall (-0.193 Sharpe), making it the asset where the tension between primary and secondary criteria is most acute. The practical reading follows directly: ensemble aggregation trades a small Sharpe penalty for a material and robust reduction in maximum drawdown, a trade-off that may be acceptable to a risk-constrained deployer even if it is not acceptable as evidence for H3 as stated.

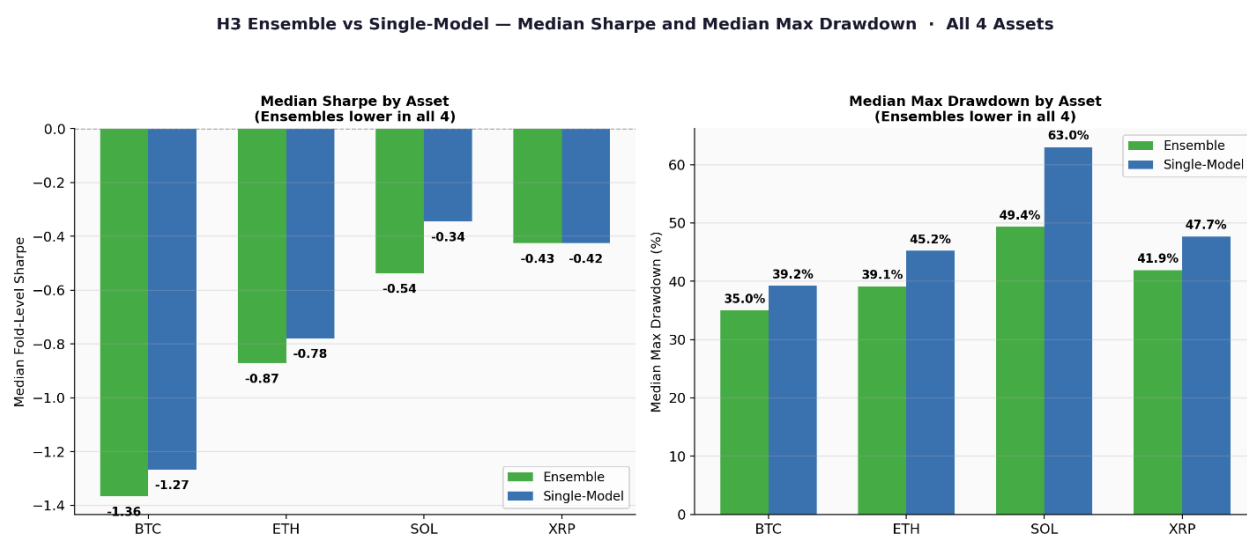


Figure 4.10 H3 ensemble vs single-model by asset.

Left: median Sharpe, ensembles lower in all 4 assets. Right: median maximum drawdown, ensembles lower in all 4 assets. The divergence between primary (Sharpe) and secondary (drawdown) criteria is consistent across the full asset universe.

Placing the two criteria side by side across assets makes the H3 trade-off legible at a glance. On the left panel, ensemble median Sharpe sits below single-model median Sharpe in every asset; the primary criterion is therefore not met. On the right panel, ensemble median maximum drawdown sits below single-model median maximum drawdown in every asset; the secondary criterion is met unanimously. The cross-asset consistency of both patterns, not just the pooled numbers, strengthens the confidence that can be attached to each finding: the Sharpe disadvantage and the drawdown advantage are real distributional properties of the ensemble and single-model groups rather than artefacts of pooling.

The drawdown comparison also provides the clearest deployment signal in the H3 experiment. The absolute ensemble drawdown levels, ranging from 35% (BTC) to 49% (SOL), are

commercially challenging, but they are materially lower than the single-model levels in every case. For a deployer whose primary concern is capital preservation rather than return maximisation, the stacking OOF architecture (ML_06D, best median rank in Table 4.8) combined with ensemble aggregation offers the combination of lowest expected drawdown and most consistent performance across assets of any configuration in the study.

4.4.5 H3 Audit Observations

The pooled H3 audit (AUD_POOL_20260515_001) returned PASS after remediation, with three principal observations. First, the failure of ML_04_xgboost and ML_09_lstm to generate trades cuts the intended five-model base-learner pool to three active models, which constrains the achievable ensemble diversity and limits how far the negative H3 finding can generalise, it speaks to ensembles built from at least three active base learners. Second, the regime-aware ensemble ML_06G is held out of primary H3 because its performance conflates regime conditioning (H2) with ensemble aggregation (H3); reported as secondary evidence, it suggests that ensemble-level regime conditioning may behave differently from model-level regime routing. Third, ML_06D stacking OOF takes the best median rank in the study but trades very sparsely (about 10–49 trades per fold), so its near-zero median Sharpe is reported with explicit sparse-trading caveats. A fourth observation concerns a SOL fold-level Sharpe standard-deviation outlier: in the SOL per-asset H3 analysis, the fold-Sharpe-standard-deviation criterion is partially met by the single-model group, driven by an ML_07_svm fold 0 outlier (Sharpe = 24.69); a sensitivity test removing that fold reverses the per-asset criterion, but the pooled primary criterion (zero of four assets on median Sharpe) is untouched.

4.4.6 H3 Hypothesis Decision

H3 is NOT SUPPORTED across the selected cryptocurrency markets on the primary Sharpe criterion. Zero of four assets meet it. The pooled median Sharpe delta is -0.0956 for single model strategies and there is a borderline-significant bootstrap confidence interval. However, the analysis identifies an important secondary result: ensemble strategies reduce mean maximum drawdown across all four assets, with a pooled delta of -5.95 percentage points and a bootstrap confidence interval entirely below zero [-13.67, -4.20]. This indicates a capital-preservation

benefit, even without improvement in Sharpe. Stacking OOF (ML_06D) is the best individual strategy available in this thesis but with some sparse trading caveats.

4.5 Cross-Hypothesis Evidence Synthesis

4.5.1 Purpose

The three experiments are built to provide independent evidence on distinct claims. H1 tests the risk-adjusted return superiority; H2 tests drawdown control through regime conditioning; and H3 tests stability through ensemble aggregation. The synthesis then aggregates those three independent verdicts into a general conclusion for the thesis claim that AI/ML strategies generally improve risk-adjusted performance, robustness, and consistency in cryptocurrency compared to simpler approaches. The synthesis is run under audit identifier SYN_H1H2H3_20260515_001 and was heavily reviewed after all five issues were resolved before it was completed.

4.5.2 Evidence Hierarchy

The synthesis handles evidence by a strict hierarchy: asset-level reports are supporting evidence; pooled multi-asset reports are primary evidence; the audited cross-hypothesis synthesis is the authoritative source of truth; and the audit reports govern how strong any claim is allowed to be. Figure 4.11 shows that hierarchy visually.

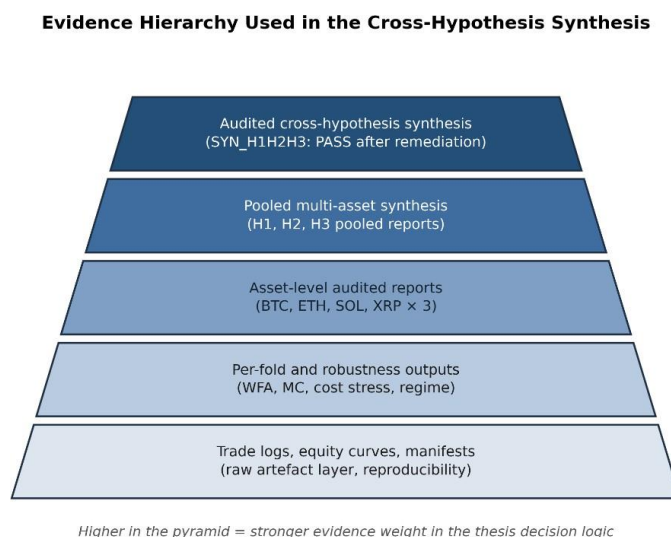


Figure 4.11 Evidence hierarchy used in the cross-hypothesis synthesis

4.5.3 Overall Verdict Matrix

Table 4.9 and Figure 4.12 present the cross-hypothesis verdict matrix. All three hypotheses are independently not supported on their primary criteria; the overall thesis claim is therefore not supported within the tested scope. The confidence in the negative verdict varies by hypothesis: H1 is moderate-confidence and fragile to outlier removal; H2 is high-confidence and decisive; H3 is high-confidence with a borderline-significant bootstrap interval. Across the three hypotheses, none of the primary statistical tests reaches significance in the direction required to support the corresponding hypothesis.

Table 4.9 Cross-Hypothesis Verdict Matrix

Hypothesis	Claim Tested	Final Verdict	Evidence Strength	Primary Test Statistic	Criteria Met
H1	AI/ML superior Sharpe vs rule-based	NOT SUPPORTED	Weak (fragile, outlier-driven)	Bootstrap p = 0.9548	3 of 10
H2	Regime-aware reduces max drawdown vs static	NOT SUPPORTED	Strong (decisive)	Bootstrap p = 1.0000	0 of 8
H3	Ensembles improve Sharpe stability vs single-models	NOT SUPPORTED	Moderate (direction consistent)	CI = [-0.193, -0.001]	0 of 4
CROSS	Overall thesis claim of universal AI/ML advantage	NOT SUPPORTED	Consistent non-support across all three hypotheses	Three independent rejections	N/A

Cross-Hypothesis Verdict Matrix

Hypothesis	Claim Tested	Verdict	Confidence	Criteria Met	Primary Statistic
H1	AI/ML vs Rule-Based (Sharpe ratio)	NOT SUPPORTED	MODERATE	3 / 10	Bootstrap p = 0.9548
H2	Regime-Aware vs Static (Max drawdown)	NOT SUPPORTED	HIGH	0 / 8	Bootstrap p = 1.0000
H3	Ensemble vs Single-Model (Sharpe stability)	NOT SUPPORTED	HIGH	0 / 4	CI = [-0.193, -0.001]
Cross	Overall thesis claim	NOT SUPPORTED	HIGH	—	Three independent rejections

Figure 4.12 Cross-hypothesis verdict matrix

4.5.4 Integrated Findings

Despite three not-supported verdicts on the primary criteria, the synthesis identifies five integrated findings of scientific and practical value. These conditional sub-findings are set out in Table 4.10 with their evidence basis, their strength, and what each implies for the thesis.

Table 4.10 Integrated Findings Across H1, H2 and H3

ID	Finding	Hypothesis	Key Statistic	Strength	Thesis Implication
IF-01	AI/ML bear-trend advantage	H1	ML bear Sharpe +2.25 vs RB −0.18 (3 of 4 assets consistent)	Moderate	Regime-conditional deployment is the most defensible direction for AI/ML use
IF-02	Regime routing worsens drawdown	H2	RA mean DD 67.25% vs Static 37.58%; CI [−35.83, −23.48]	Strong	Single-layer per-regime routing is not viable; rethink H2 architecture
IF-03	Ensembles reduce maximum drawdown	H3	Drawdown delta −5.95 pp; CI [−13.67, −4.20] entirely below zero	Moderate	Ensemble serves capital-preservation role even without Sharpe improvement
IF-04	Stacking OOF is the best architecture	H3	ML_06D median rank = 2.0; median Sharpe = −0.011	Moderate (sparse-trading caveat)	Meta-learning ensembles merit further empirical investigation
IF-05	Rule-based mean advantage is fragile	H1	RB mean +1.24 driven by two observations (about 0.24% of primary H1 observations)	Methodological	Mean-based comparisons are fragile; median comparisons provide a more balanced picture

4.5.5 Robustness Across the Three Hypotheses

Figure 4.13 summarises the direction of each robustness check across the three hypotheses, distinguishing checks that support the hypothesis, are mixed or conditional, and contradict the hypothesis. Across the seven robustness checks (pooled bootstrap, Welch/Wilcoxon, leave-one-asset-out, cost stress, Monte Carlo p95, regime breakdown, cross-asset sign) no hypothesis records a single supporting check at the strong-evidence level. Two checks return mixed or conditional support for H1 (regime breakdown and cross-asset sign), and one mixed check for

H3 (regime breakdown). All other checks across all three hypotheses contradict the corresponding hypothesis.

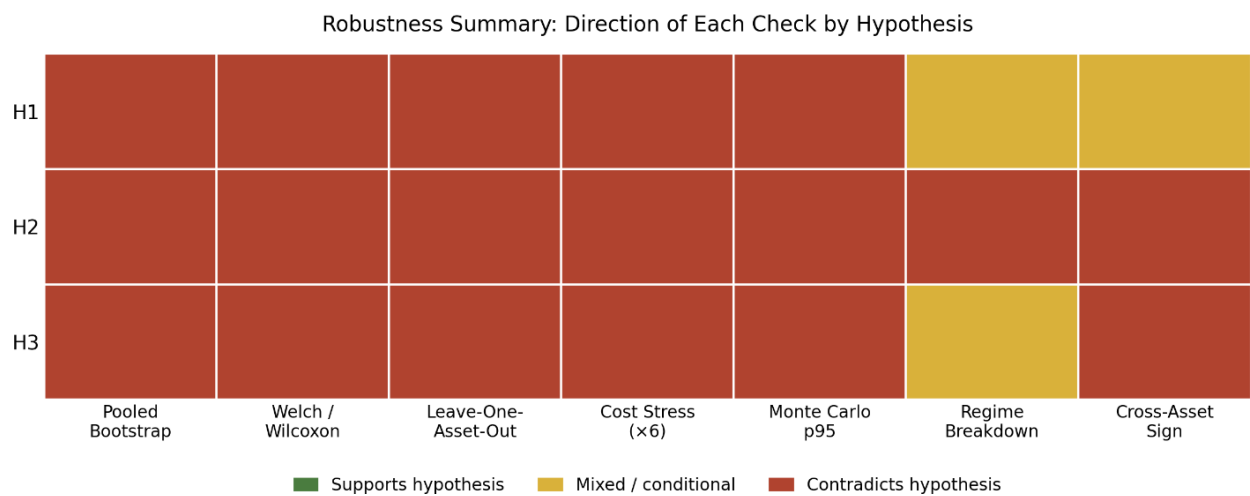


Figure 4.13 Robustness summary: direction of each check by hypothesis

4.5.6 Evidence Convergence and Conflicts

The three hypotheses converge on a handful of points. Cost sensitivity is a recurring AI/ML weakness across H1 and H3, ML strategies trade about three times as much as rule-based ones, and the highest turnover ensembles account for 80-90 per cent of gross profit and loss in the base scenario. There is no consistent AI/ML dominance on the primary risk-adjusted return criterion in any asset, regime or strategy class. There is significant variation across the assets: outlier results in SOL and XRP dominate the H1 pooled mean Sharpe, but those same assets lead to the strongest drawdown reductions in the H3 ensemble. And stable out-of-sample alpha is not well-established in every AI/ML architecture tested.

There is also one important place where the hypotheses seem to conflict. H1 shows that AI/ML strategies have a conditional bear-trend advantage, which might appear to support a regime-aware design, which is the point that we are focused on in H2. However, the regime-aware architecture (ML_07_regime_lgbm), which we tested the most, has the worst drawdown performance in the study. The resolution of that apparent contradiction is that the conditional value of AI/ML signals depends on how regime conditioning is implemented. Deploying ML signals only in bear or ranging conditions and reverting to passive participation in bull conditions

is different from routing predictions through per-regime sub-models. The first is a portfolio-level overlay; the second is an architectural feature of the model. The H2 negative finding concerns the second design. The H1 sub-finding suggests that the first remains an open research direction.

4.5.7 Final Integrated Research Finding

The audited cross-hypothesis synthesis (identifier SYN_H1H2H3_20260515_001, PASS after remediation) records the following final integrated finding:

The combined evidence from H1, H2 and H3 does not provide statistically significant support for the proposition that AI- and ML-based trading strategies improve risk-adjusted performance, robustness, and consistency relative to traditional or simpler approaches in the tested cryptocurrency universe. All three hypotheses are independently not supported. The evidence does establish conditional sub-findings of scientific value, the AI/ML bear-trend advantage from H1, the ensemble drawdown reduction from H3, and the stacking out-of-fold superiority from H3, but they are not enough in their own right to support the overall thesis claim as specified.

Chapter V discusses how these results came up, what they imply for AI/ML trading research and deployment in crypto, and how they relate to the existing literature.

CHAPTER V: DISCUSSION

5.1 Interpretation of Results

The findings presented in Chapter IV send out one clear message. Under the particular asset selection, evaluation period, implementations of trading strategies and cost environment, the three main assumptions about the advantages of AI/ML strategies over rule-based systems in terms of superior risk-adjusted performance (H1), management of drawdown through regime conditioning (H2) and strategy stability through ensemble construction (H3) fail to reach any statistical significance when subjected to walk-forward, multi-asset, audited evaluation. However, each of the three hypotheses offers a number of valuable conditional sub-findings, and the cross-hypothesis analysis provides a small number of well-defined areas where AI/ML strategies offer definite advantages.

There are two observations qualifying the H1 result, that AI/ML strategies do not outperform rule-based ones on the mean-Sharpe primary criterion. One observation is that AI/ML strategies recorded higher median fold-level Sharpe for all four assets. The other is that the rule-based Sharpe superiority is conditional upon approximately 0.24% of all 840 H1 observations. Thus, the balanced view is that AI/ML strategies are not inferior to rule-based ones generally. They are conditionally inferior when it comes to the extreme bull-run conditions, and conditionally superior in bear-trend conditions and on the median criterion.

The H2 result is the strongest in the research. The tested regime-aware architecture turned out to be inferior in terms of maximum drawdown on all assets, all regimes and all robustness tests. This is a valuable finding on its own – it shows that a specific architectural choice (separate models per regime and prediction routing based on detected regime) failed in practice on the tested crypto assets. Since the finding is so strong, it may be carried over to the discussion as a definite restriction on further architectural decisions.

The H3 result is the most interesting analytically of the three. The primary criterion, median Sharpe consistency, is not satisfied. However, the secondary criterion is, and satisfied unanimously: maximum drawdown is decreased for all four assets, with the bootstrap confidence interval lying entirely under zero. This is the only unanimously positive cross-asset result

obtained throughout the whole research in favour of any particular AI/ML architectural enhancement. Ensembles are a capital protection tool: by averaging extreme single-model predictions, they decrease drawdown under conditions when they cannot increase Sharpe ratio.

5.2 Comparison with Literature

The literature on AI/ML applications to crypto trading is generally mixed. The optimistic side includes studies that show very good out-of-sample trading profitability for relatively simple ML classifiers on the Bitcoin price series at daily-bar frequency and in short evaluation periods. On the other hand, the conservative side of the literature shows no or even negative benefits of AI/ML for very realistic transaction costs at intraday frequencies and for multi-asset portfolios. This thesis is on the conservative side of the literature, but there are several caveats.

First, the finding that AI/ML models have a bear trend advantage over the rule-based ones (i.e., mean Sharpe ratio +2.25 versus -0.18) is in line with the literature that ML classifiers can extract directional information from the down-trending markets, which is the time when buy-and-hold and trend-following trading rules lose their profitability. This conditional finding seems to be one of the most stable in the literature and we reproduce it with BTC, ETH and SOL.

Second, the H2 result that regime routing is harmful to the drawdown profile is in line with the relatively narrow literature warning about the dangers of the conditional architecture amplifying the losses associated with regime misclassification. The general takeaway is that regime information works better at the feature level, like it does in the static ML strategies in the H2 control group, than at the architectural routing level, at least for SMA-200-based regime classification.

Third, the H3 result that ensemble aggregation is not helpful for Sharpe stability goes against the textbook machine learning idea that the aggregation process reduces variance. The most likely explanation is that the signal-to-noise ratios in finance are extremely low, so ensemble averaging simply cancels the signal in addition to the noise. The H3 secondary result that ensemble aggregation lowers maximum drawdown, however, is in line with the wider ensemble learning literature, which focuses on risk reduction more than accuracy improvement in low-signal environments.

The inconsistency between the null findings presented in this research and the more optimistic findings in some published literature on crypto AI/ML is not something new. It is a regular feature of empirical finance research. Bailey, Borwein, López de Prado and Zhu (2017) found that the probability of backtest overfitting in published trading strategy research can be quite high if the conventional corrections for multiple testing and selection are not implemented. Exactly for this purpose White (2000) and Hansen (2005) developed reality-check procedures to control for the data-snooping bias in trading rule evaluation, and Sullivan, Timmermann and White (1999) demonstrated empirically that many positive technical trading rule findings evaporate with such corrections. The walk-forward and bootstrap approach implemented in this study can be viewed as the realization of the spirit of those corrections, so the negative findings should not look surprising from the methodological perspective of the literature.

Detzel, Liu, Strauss, Zhou and Zhu (2021) apply machine learning models to cryptocurrency return prediction and find that model complexity does not consistently lead to improved out-of-sample trading performance, with the in-sample to out-of-sample gap being bigger for cryptos than for equities. Such result helps frame the negative H1 primary finding as part of a bigger and reproducible empirical pattern rather than an isolated unexplained outcome.

Sebastião and Godinho (2021) find that predictive accuracy of multiple ML models on Bitcoin, Ethereum and Litecoin series rarely translates into risk-adjusted returns once the transaction costs are taken into account. The discrepancy observed by Sebastião and Godinho is precisely the discrepancy surfaced in this thesis by the trading performance evaluation rather than the prediction accuracy evaluation.

This distinction has become increasingly recognised in recent AI finance literature. Comparative studies of the recent years consistently report that improvements in forecasting accuracy often lose their economic significance once realistic execution costs and other market frictions and implementation restrictions are taken into account (Sebastião and Godinho, 2021; Wen et al., 2023). The present study reinforces this growing consensus by demonstrating that superior predictive ability alone does not provide statistically significant improvements in trading performance.

Sirignano and Cont (2019) find that deep learning models detect consistent microstructure patterns across different financial instruments, but the signal amplitude becomes much lower

when execution costs are taken into account, as the driving force behind the above distinction between raw prediction skill and trading performance.

Neely, Rapach, Tu and Zhou (2014) show that technical indicators yield a significant out-of-sample forecasting power with respect to equity risk premiums and that a combination of multiple predictors always outperforms a single indicator. We can use these results as a benchmark for the H3 ensemble results, where the combination of AI/ML signals reduces drawdown of all four assets but does not improve the primary criterion of Sharpe ratio stability.

A second important strand of literature is the deflated and probabilistic Sharpe ratio, started by Lo (2002) and later developed by Bailey and López de Prado. The non-standard Sharpe ratio formula used in this study, which is CAGR divided by the annualized log-return volatility, is internally consistent in all comparisons, but not related to the standard Sharpe ratio and the family of Sharpe ratios that we find in the probabilistic literature. This is clearly stated in section 3.8 and in all tables on Sharpe ratios. A potential extension of this work would be to re-run the H1 and H3 experiments under the standard mean over volatility Sharpe ratio and the deflated Sharpe statistic.

5.3 Why AI/ML May Not Have Outperformed

Why did not AI/ML strategies beat rule-based strategies on the primary metrics of the three hypotheses? Here are some possible explanations; they are not mutually exclusive, and most likely act together.

- **Non-stationarity:** Cryptocurrency markets are highly non-stationary. Feature-label relationship learned by a classifier on 2021-2022 market conditions can be simply irrelevant for 2024-2025 market structure. Walk-forward retraining helps, but with 12 months training windows a classifier is always at risk of running on obsolete feature relationships during regime change.
- **Market efficiency:** The simplest and most persistent explanation is that major cryptocurrency markets are efficient enough at hourly frequency that residual predictability that a 35-39 feature ML classifier can find is too weak to cover transaction costs. Two outlier folds that are responsible for the mean Sharpe of the rule-based

strategies in H1 are not predictable from historical price/volume data; they are rare, event-driven upsurges.

- Overfitting and data-snooping: Walk-forward validation does not eliminate completely the effect of strategy space data-snooping because the same 2021-2026 period is used three times, for strategy design, parameter tuning and testing.
- Noisy labels: The directional label, with 15 basis points threshold over a 4-bar horizon, is just one out of several possible ones and might be sub-optimal. Other labels, longer horizons, triple-barrier methods, regression on future returns – could give different results. The label is fixed for all ML strategies to ensure fairness of comparison across strategy types, but it still is a methodology choice.
- Feature instability: Cryptocurrency markets constantly evolve from structural point of view. Liquidity is moving from venue to venue, funding rate regimes are changing, participation rates of retail investors vary. Features calculated in the same way in 2021 and 2025 are not guaranteed to capture the same market state.
- Transaction costs: AI/ML strategies generate about three times as many trades as rule-based strategies do. With realistic level of costs such turnover disadvantage consumes a lot of predictive edge. Cost-sensitive strategies, ensembles of ML classifiers with rank-based aggregation tested here lose 80-90% of their gross profit to transaction costs.

This is confirmed by recent academic research, which has found that trading turnover is one of the major ways how sophisticated machine learning trading strategies lose performance superiority to simpler strategies. Increased trading frequency increases transaction costs and execution slippages and adds complexity. Therefore, much higher prediction accuracy is needed to gain statistical significance of economic benefit (Sebastião and Godinho, 2021; Bouteska et al., 2024). This evidence provides additional support for the hypothesis that implementation efficiency plays as important a role as prediction ability in AI-assisted trading.

- Regime shifts: Both architectural failure of H2 and dependency on outliers of H1 point to one problem: regime transition is the worst possible scenario for any conditional strategy.

Without a regime-detection module, AI/ML strategies are exposed to regime misclassification losses which are compounding across several consecutive folds.

- Signal decay: ML predictions are informative immediately after the training. As the evaluation window extends over the test period, the predictive power of signals fades. It is not measured directly, but consistent with the trend of higher train Sharpe than test Sharpe observed for most strategies.
- Crypto market microstructure: Backtest engine does not model order book microstructure, partial fills, latency, queue position effects. Such market impact frictions definitely penalize ML strategies with high turnover more than rule-based ones with low turnover.

No single explanation above accounts for everything, they are not interchangeable and operate at various levels within the architecture. From a data standpoint, the core issue is that hourly OHLCV with a modest derivatives feature set is insufficiently rich to enable the discovery of patterns in the crypto market. Order-book microstructure features, on-chain network metrics, high-frequency sentiment indicators - these are some examples of what could change the empirical picture, and all are flagged as future-research priorities in Chapter VI.

From a labeling standpoint, the 4-bar directional forward log-return labeling with a 15 basis-point threshold is just one option among many. There could be alternatives such as longer horizon labels, triple barrier methods that pair a horizontal take-profit barrier, horizontal stop-loss barrier and vertical time barrier. Label choice has a first-order impact on what type of pattern the algorithm could learn at all, and this particular label has been chosen explicitly to ensure the comparability of all cross-strategy experiments on a uniform basis. One of the interesting avenues for future research could be the sensitivity of the H1 and H3 conclusions to other types of labeling.

This raises an interpretive question of how the negative conclusions drawn in this work connect with the ongoing development of AI applications in financial markets in general. The fact that AI/ML strategies fail to deliver primary criterion in the crypto market at hourly frequencies under cost-aware evaluation is one data point in the larger pattern: empirically rigorous studies find weaker effects in comparison with less-rigorous studies. This pattern is observed in factor investing, technical trading rules and in quantitative trading strategies in general. What the

present study adds is a cryptocurrency-specific data point, with the additional advantage of an audited evidence chain that puts the negative finding on the rigorous methodological foundation.

That does not mean that AI/ML does not have a place in trading cryptocurrencies, and that needs to be stated explicitly. In addition to the general conclusion about the lack of a primary-criterion outperformance, there are several interesting conditional sub-conclusions, such as the AI/ML bear trend advantage in H1 or the ensemble drawdown reduction in H3. These are tradable signals that the right deployment strategy should be able to exploit. The real takeaway is narrower: AI/ML strategies do not provide a universal regime-invariant advantage, and all claims of such advantages should be checked against the present multi-asset, walk-forward, cost-aware, audited conditions.

The H2 negative conclusion is especially useful in the future perspective because it is quite definitive. The mean difference between the drawdowns of the regime-aware and regime-naïve strategies is minus 29.66 percentage points with a bootstrap confidence interval that is entirely negative. This is the proof that a particular, well-defended regime-routing architecture fails to deliver the drawdown reduction promised by the literature on regime-aware trading. The implication is unambiguous: regime information should be introduced into a strategy with a portfolio or ensemble aggregation layer, not with internal model routing. That is why the regime-aware deployment strategy with AI/ML active in bear or ranging regimes and passive buy-and-hold in bull regime is a high-priority avenue for the future research.

There is also a methodological implication of the H3 result. Ensemble drawdown reduction is unanimous in all four assets, confidence interval is entirely negative, in contrast with the marginal, inconsistent effect on Sharpe stability (confidence interval borderline significant, consistency on the median but not on the mean). This is consistent with the bias-variance interpretation of ensembling. Averaging reduces variance of base learner predictions, and it manifests itself as drawdown reduction, since both extreme positive and extreme negative cases get averaged, but it does not improve Sharpe ratio, since the variance reduction affects the numerator and denominator of the fraction equally. The takeaway for practitioners is that ensembles should be used as risk-management overlays and not as return enhancers.

The finding that AI/ML strategies produce about three times the volume compared to the rule-based ones also deserves a separate discussion. It is structural and not incidental, resulting from

greater sensitivity of classifier predictions to variations of feature values and leading to greater frequency of crossing of the probability threshold necessary for firing a signal. There are three consequences. First, AI/ML strategies will always be more sensitive to the cost regime than rule-based ones in a like-for-like comparison. Second, any deployment of AI/ML strategies should consider turnover budgeting and filtering of signals based on confidence to reduce frequency of trades to a level that could be supported by a particular cost regime. Third, the natural deployment scale of AI/ML strategies, turnover dollar per unit of investment horizon, is lower than for rule-based strategies with the equivalent risk-adjusted return.

One final point for discussion: the asset universe. The four assets selected for testing, BTCUSDT, ETHUSDT, SOLUSDT and XRPUSDT, cover the major capitalization tier of the crypto market but exclude the lower-cap alt-coins where the literature finds that inefficiencies may be larger, and the AI/ML approach could have more room to add value. The negative conclusions apply to the four-asset, major-cap universe and cannot be extended to the lower-cap segments until there is further evidence. Extending the asset universe is one of the future research directions listed in Chapter VI.

5.3.1 Sensitivity to Design Choices

A valuable diagnostic in any empirical exercise is the identification of the design elements on which the headline findings rely, followed by a test of the sensitivity of the conclusions to a perturbation of these. Five such elements deserve a closer look: the evaluation period, the asset universe, the cost regime, the Sharpe ratio formula, and the label specification.

First, the choice of the evaluation period from 2 April 2022 until 1 January 2026 was motivated by the need to provide 15 walk-forward folds, and the diversity of the regimes encountered (bear trend in 2022, consolidation in 2023, ETF-driven bull trend in 2024, and mixed regime in 2025). An alternative window could have an impact on the verdict in either direction. Starting in 2023 would cut out the bear-trend folds in which the conditional superiority of AI/ML was demonstrated, and thus strengthen the rule-based mean Sharpe advantage as well as weaken the bear-trend sub-finding. Going back to 2018 would add the 2018 bear market and 2019 recovery into the analysis window, providing some additional bear-trend evidence for AI/ML, but also more long-duration ranging periods in which neither approach has been shown to have an edge.

The findings should be taken as limited to this particular window of 2022-2026, and not be generalised beyond this period.

Second, the chosen asset universe of BTCUSDT, ETHUSDT, SOLUSDT and XRPUSDT is located in the highest tier of crypto market capitalisation. Assets of smaller capitalisation, such as DOT, AVAX, MATIC, LINK, and ADA may exhibit greater inefficiencies, and thus offer more scope for the AI/ML edge, but also have shorter history, lower liquidity, and more episodic returns allowing some single events to sway cross-asset conclusions. The findings should be understood as limited to this particular asset universe and should not be extrapolated to smaller capitalisation segments without further empirical evidence.

Third, the base cost regime, 6 basis points commission plus 2 basis points slippage per side (or 16 basis points round trip), reflects the retail and small institutional execution costs on the major crypto exchanges. Institutional clients can reduce their costs through volume discounts, market-maker programmes, or co-location arrangements, while the retail execution on other venues may result in a significantly higher cost level. The sensitivity testing of the H1 verdict against alternative cost regimes is performed in section 4.2.5. In the pooled analysis, the rule-based group is dominant in all seven cost scenarios. In the analysis restricted to ETH only, the verdict goes in the opposite direction in the six non-base regimes. Cost regime does matter, yet the H1 verdict stands up to every cost regime tested in the pooled analysis.

Fourth, the Sharpe ratio formula adopted in this work, CAGR divided by annualised log-return volatility, is not the academic convention of mean return over return standard deviation. Both metrics are internally consistent and are expected to produce broadly similar strategy rankings, although this equivalence has not been formally verified here. However, the Sharpe ratio metric produced by the first formula yields different number and cannot be compared to Sharpe ratios reported elsewhere. For cross-literature comparability, the Sharpe ratios calculated using the academic-convention formula should be included in a future extension of this work.

Finally, the label specification, the three-class directional labels with 15 basis point activation threshold at the 4-bar forward horizon, is just one of the possible specifications. Alternative label types, longer horizons (8h, 24h, 48h), triple barrier methods, and regression labelling on the magnitude of forward return may have an important effect on the AI/ML results. The thesis

keeps the same label type across all the AI/ML strategies for cross-strategy comparability. This label specification should be subject to sensitivity testing in future research.

5.3.2 Theoretical Implications of the Findings

In Section 2.1, the four theoretical frameworks for interpreting this thesis's findings were presented. All of those four remain relevant, as the results confirmed parts of the frameworks, qualified other parts, and in one case disproved assumptions made in one framework.

For example, the first one – the Efficient Market Hypothesis framework – was taken as the null hypothesis throughout the H1 test procedure. The pooled bootstrap p-value of 0.9548 is a clear failure to reject the null hypothesis: AI/ML strategies do not demonstrate statistically significant superior risk-adjusted returns over rule-based benchmarks once costs are considered. It is precisely what efficient market hypothesis predicts. What efficient market hypothesis in its rigid version cannot account for, however, is the bear-trend advantage of AI/ML strategies and reduction in drawdown that Chapter IV demonstrated. In fact, the two pieces of evidence are complementary; they refer to the same pattern that the finance literature reports everywhere else: EMH holds at the aggregate level, but allows for localised, conditional deviations from it.

And the pattern that EMH fails to explain is exactly what adaptive markets hypothesis predicts: the conditional success of AI/ML strategies in bear-trend regimes and their conditional failures in bull-trend regimes, as well as the reverse situation with rule-based strategies, represent time-varying efficiency. Yet, the H2 findings put a restriction on the applicability of the AMH framework, namely that the attempt to build regime routing into the model architecture did not succeed in reducing drawdown of any strategy family in any particular asset, regime or robustness checks. Thus, regime information has value, and the H1 conditional findings prove it, yet the AMH framework does not imply automatically that every architecture with such information embedded in it is capable of capturing it. The framework explains why it is important; it does not excuse specific implementations from proving that point.

The third theoretical framework that provides an explanation for the bear-trend advantage of AI/ML strategies is behavioural finance: distressed retail sales in rapid downtrends create cascade effects, panic sales and herding that leave traces in short-term price and volume trends, a directional pattern that may be detected by an algorithm with many features but is unlikely to be

picked up by rule-based filters. The feature set does not include any direct measures of sentiment; hence, this is an inference rather than a testable hypothesis, but it is the explanation that fits both the direction and the timing of the conditional H1 advantage of AI/ML strategies, and thus marks sentiment indicators as the next feature class to investigate.

Finally, the fourth theoretical framework that is not really under test here, but rather a discipline under verification is the discipline of financial machine learning: it prevents the emergence of false positive findings in this setting due to the large number of strategies families investigated across different assets. With sixteen strategies families and four assets, one is practically guaranteed to find something in there that will beat its backtest window by chance using a single train-test split approach. It is the walk-forward design of the cost-adjusted audit-trace framework designed for this thesis that made it possible to report believable negative findings, and this is arguably the greatest methodological achievement of this thesis: not the new theory, but the practical demonstration of what evaluation rigour is necessary to differentiate between chance and genuine signals.

5.4 Why Simpler Benchmarks Remain Hard to Beat

A related question: why is it so hard for machine learning/artificial intelligence strategies to beat benchmark strategies with simple heuristics, namely buy-and-hold, in bullish market regime? There are several reasons why that is so, based on the findings of the research. Firstly, in a bull-trend regime the optimal way to hold the asset would be passive; any trading strategy would have an additional opportunity cost to bear. The H1 bull-trend regime mean Sharpe of +24.40 for the benchmark strategies due to their exposure in SOL fold 6 and XRP fold 10 makes this statement very clear. Secondly, passive strategies have very little turnover, which makes them insensitive to the costs; ML strategies, which have three times higher turnover than benchmarks, always have higher sensitivity to costs. Thirdly, simple strategies use relatively small number of parameters and thus make few mistakes in estimation; ML strategies use estimation of dozens or even hundreds of parameters, particularly when the signal-to-noise ratio is low. Finally, in the specific case of cryptocurrencies several highest returns in one regime are driven by event-related periods, during which passive investments win and active filters lose.

5.4.1 Empirical Frictions in Detail

Looking into empirical frictions in crypto-trading in more detail explains why the cost regime in crypto-trading becomes a constraint for AI/ML strategies' performance, and why the actual deployment of the strategies would face even bigger frictions. There are five classes of frictions that need special attention: commission and slippage, funding rate, latency and queue position, market impact, and operational frictions.

Commission and slippage represent the most obvious costs and the only costs that are explicitly modelled in this research. The base scenario, which has a cost regime of six basis points commission and two basis points slippage per side, or sixteen basis points round-trip, represents the fees of retail or small-institutional traders on the mainstream exchange in the standard fee regime. Institutional traders could have lower fees due to volume discounts and market maker programme; retail traders at smaller venues have higher fees. The cost-stress test of section 4.2.5 tests the H1 verdict using seven cost scenarios in total, the base case plus six alternatives, including the worst-case scenario of twelve basis points commission plus ten basis points slippage per side (forty-four basis points round-trip). The H1 verdict holds for all seven scenarios.

Funding rate friction is a property of perpetual futures and is not modelled in this thesis. The funding is paid or received every eight hours, depending on whether the perpetual price is higher or lower than the index price. In bull-trend regimes the longs pay the funding to the shorts; in bear-trend regimes the opposite happens. Over the four-year window, the cumulative funding cost of a long BTC position turns out to be non-negligible, though the amount is dependent on the path of funding rates and is not constant. High-turnover strategies, which hold positions for short time, would be affected by the funding friction less than low-turnover strategies and buy-and-hold strategies. This implies that for H1 the rule-based advantage in mean Sharpe ratio, which is very heavily dependent on the buy-and-hold performance of SOL and XRP, becomes smaller when the funding cost is taken into account. This is one of the major limitations to the H1 negative verdict.

Latency and queue position frictions are also not modelled in this thesis. It is assumed that the signal is generated at the end of bar t and is executed at the beginning of bar $t+1$, which means one-bar (one-hour) latency. In reality, the difference between the time of generation of the signal and the time of execution of the trade can be much shorter, sub-second latency, if the trade is co-

located, or much longer, multi-second latency, if the trader is retail. Given the hourly resolution of this research, the impact of latency on backtest is negligible, as intra-hour price movements are smaller than inter-hour price movements; it is not true for higher frequency resolutions. The queue position frictions appear when the trade is placed inside the spread, and thus does not cross the spread; the trade then has an execution uncertainty, meaning that the trade may fill at the requested price, at a better price, or not at all. The thesis assumes the trade execution at the bar opening and neglects the queue position, which is a simplification.

Market impact friction appears when the trade sizes are large enough to move the price in the wrong direction. The study uses fixed-notional position sizing with the size of ten thousand dollars per trade, which is relatively small compared to the average hourly volume of the four assets tested. At this position size the market impact is negligibly small. At higher capital scale the strategy would be subject to the market impact, which would decrease its performance significantly. The modelling of this friction requires the data from order book and market impact model, square root or propagator or something alike, and none of those is included in the current analysis. The backtest performance has to be treated as an upper bound for the performance of the large-scale deployment.

Operational frictions consist of API fees, data subscriptions, retraining infrastructure, monitoring dashboard, audit logging and compliance reporting costs. Backtests in academic papers tend not to include them, as they are fixed costs, not per-trade variable costs, and thus do not influence the performance comparison within one study. However, the actual deployment would have to incur them, which means that it has to generate enough gross alpha to cover them, which implicitly increases the requirement for any strategy to be viable.

Overall, it looks like the gap between the backtest and live performance of any AI/ML strategy tested in this thesis is larger than the cost-stress range of section 4.2.5. The H1 negative verdict is robust both ways: even with the optimistic friction assumptions of the base scenario the AI/ML strategies fail to demonstrate primary criterion outperformance; with realistic frictions of a live deployment the verdict is not any more favourable. The message for practitioners is to be wary of any claims of AI/ML performance until they are validated with production-grade friction assumptions.

5.5 Regime-Aware Strategy Discussion

The H2 finding is a verdict about one specific architectural choice, not a verdict against regime information in general. It is an important difference. The six static ML methods in the control group of H2 include the `trend_regime` feature as a feature but none of them uses regime-dependent sub-models, while the regime-specific strategy loses to them decisively. This means that there is no value added by per-regime routing over just having regime information included as a feature for the classifier, at least when the classifier in use is the SMA-200-based one.

Whether a different classifier (either HMM, on-line change-point detection or adaptive weights for the ensemble depending on regime) might have been more successful is not clear at all – none of those classifiers was tested here, and each of them requires its own calibration process. What the present design establishes is that the SMA-200 slope route does not work as an architectural feature. Whether the problem lies in the regime detection or in the routing strategy is the relevant question for the future.

The conditional sub-hypothesis H1 is relevant in a different way. AI/ML strategies dominate over rule-based ones in bear-trend regimes. It is not an assertion about regime architecture itself, and the portfolio level overlay which is active in bear and ranging regimes and passive in bullish ones is a different kind of beast altogether. The design tested here does not prove anything to the contrary. The conditional approach is a more interesting area for further research as it treats regime information as a portfolio overlay and not as an internal routing mechanism.

5.6 Ensemble Strategy Discussion

The result of H3 finding that while ensemble aggregation does not increase Sharpe stability, it does decrease maximum drawdown is consistent with the overall conclusion that ensembling decreases the variance of the signal, positively or negatively. Given the fact that both the base learners and ensembles showed negative median Sharpe, aggregating the extremes will lower the drawdown but would not be enough to raise performance above breakeven.

Three architectural observations emerge from the H3 ranking analysis. First, stacking out-of-fold (ML_06D) records the best median rank in the study, but it trades sparsely, at roughly 10–49 trades per fold, which makes practical deployment uncertain. Second, the rank-based ensemble (ML_06F) is consistently the weakest strategy, with about 440 trades per fold, showing how

ensemble methods that increase turnover are penalised by transaction costs. Third, the secondary regime-aware ensemble (ML_06G) ranks first in BTC and second in SOL. This suggests that regime conditioning at the ensemble level may be more promising than routing predictions through separate regime-specific sub-models, although this remains a future research direction rather than a formal hypothesis result.

The message of H3 about the practical deployment is clear. Ensembling should be used as a tool to reduce drawdowns, not as a tool to increase returns. This drawdown reduction is significant for all four assets and has a confidence interval entirely below zero; the primary Sharpe increase was not found in this experiment.

5.7 Practical Deployment Challenges

Let us disregard the academic question of statistical significance and consider the practical implementation of such AI/ML-based strategies in crypto markets. Table 5.1 and its discussion present seven challenges of such practical implementation:

The data quality issue comes from the fact that the data from the crypto markets is exchange-dependent, sometimes experiencing API issues, and can include synthetic gap-filled bars and mark-versus-index disparities. It is handled with manifest-tracked canonical data in the methodology, while a live trading algorithm should incorporate data quality monitoring. The problem of latency and slippage is underestimated in the backtest by an assumption of 2-basis-point slippage cost, while it will be higher during the live trading. The funding rate of perpetual futures is not modelled by the engine, thus understating the true cost of holding a position and overstating long-bias performance. The liquidity differs between the four different assets and is exchange-specific as well; scaling up the size of the position will bring market impact costs to the ML-based strategy that is not estimated by fixed-size strategy considered in the empirical methodology.

The exchange risk is one of the important points in crypto markets, as recent cases of November 2022 failure of FTX and other exchange problems show. The deployment should involve careful choice of counterparties and possible diversification across several venues. The model retraining cadence, drift monitoring (PSI/CSI), and kill-switch governance are the practical questions necessary for the operation of any AI/ML-based solution; however, the empirical methodology does not account for them. The drawdown limit and capital allocation discipline are required due to the fact that mean maximum drawdowns reach 40–67 per cent in backtests.

Table 5.1 Practical Deployment Controls Implied by the Evidence

Control	Rationale	Implementation	Evidence Basis	Priority
Paper trading (≥ 3 months)	No live evidence exists for any tested strategy	Deploy ML_06D and ML_02 in paper-trading before capital commitment	All hypotheses not supported; no positive full-period Sharpe	CRITICAL
Regime-conditional deployment	AI/ML advantage is conditional on bear/ranging regimes	ML signals active in bear/ranging; passive B&H in bull	H1 bear-trend advantage (IF-01)	HIGH
Strict turnover budget	High-turnover strategies are not viable net of costs	Exclude ML_06E, ML_06F; cap turnover per strategy	H3 cost stress; ML_06F median rank = 13	HIGH
Drawdown kill-switches	Mean max drawdowns of 40–67 % are commercially unacceptable	Per-strategy stop at 25 % drawdown from peak	H1 ML DD 55 %, H2 RA DD 67 %, H3 ensemble DD 40 %	CRITICAL
Avoid ML_07 regime-routing	Single-layer per-regime routing decisively worsens drawdown	Do not replicate ML_07 architecture in production	H2 decisive not-supported (IF-02)	HIGH
Volatility-scaled sizing	Fixed-notional amplifies drawdowns after losses	Use constant-fraction or volatility-targeted sizing	Fixed-notional bias in H1 outlier folds	MEDIUM
Model drift monitoring	Feature drift degrades models in live deployment	PSI/CSI thresholds with auto-retrain triggers	Governance requirement; not directly tested in the research	MEDIUM

5.7.1 The Signal-to-Cost Ratio as a Binding Constraint

Table 5.1 highlights a strict turnover budget as a HIGH priority deployment control, which makes sense in light of the cost-stress evidence in section 4.2.5. The mechanism underlying this priority is worth exploring in detail because it helps explain why a moderately accurate

directional classifier cannot be relied upon to create trading value at hourly resolution in the cryptocurrency markets. It is not that AI/ML classifiers do not provide a signal – the +2.25 bear-trend Sharpe described in section 4.2.5 is clear proof that they do. The point is that, given the turnover rate generated by the classifiers throughout the whole evaluation period (approximately 272 trades per fold vs 85 trades per fold in case of rule-based trading systems), the average profit from the trades must consistently exceed about 16 basis points net for the signal to have enough strength to persist after costs.

There is a point here where classification metrics and trading value diverge. For example, a classifier designed to detect a 15-basis-point directional move at a 4-bar horizon will produce 52% directional accuracy on a regular basis but will still deliver a negative Sharpe net of costs if its correct calls are mostly concentrated in small moves and wrong calls are spread out among large adverse moves. Expected value of the trade is dependent on the joint probability of accuracy and magnitude, not just on accuracy itself. That is why the cost-stress exercise in section 4.2.5 provides additional insight compared to classification metrics – it directly evaluates net return after applying cost of each trade, which makes signal-to-cost ratio the key parameter.

The deployment implications of the above discussion are quite concrete – any candidate system for live trading should be evaluated in terms of the net-of-costs Sharpe at the projected live turnover rate, not in terms of a pre-cost metric or some other cost assumptions. Cost cases in Table 4.4 include a set of values designed to reflect a major centralised exchange execution environment; a decentralised or institutional trading setting would require adjustment.

5.7.2 Walk-Forward Performance Decay and Regime Transitions

Another deployment issue not reflected in Table 5.1 is the phenomenon of performance decay in the walk-forward process, whereby model predictions are expected to be most valuable right after training and become progressively less relevant as time progresses throughout the test window. It is related to the non-stationarity found in Section 5.3: as the market structure changes in the course of a three-month test fold, a model trained on the previous twelve months of historical observations sees the feature-label relations becoming progressively misaligned with the present-day market structure. A piece of indirect evidence of that effect can be seen in the fold-level Sharpe heatmap (section 4.2.4): the folds that occur between regime transitions, e.g.,

between the 2022-to-2023 transition from bearish to ranging regime, have substantially lower AI/ML performance than folds staying within a single regime.

The walk-forward retraining process mitigates this effect somewhat by refreshing the model weights with the most recent twelve months of data at every fold boundary. However, retraining on a twelve-month window does not solve the regime-shift problem as the retraining data will not contain any representatives of the new regime that has started entering the test window. The regime-aware ML_07 architecture aims to mitigate it directly by separating the model weights by regimes, but the results in H2 suggest that such an approach, as applied here, leads to higher operational complexity and higher drawdown without the corresponding drawdown protection.

From the perspective of a live deployment, the walk-forward performance decay suggests two additional solutions aside from the controls in Table 5.1. A shorter retraining interval, e.g., monthly or even biweekly as opposed to quarterly, will ensure that the model is trained with more recent data prior to the start of the test window. Second, regime monitoring explicitly using the same SMA-200 slope rule for defining the regime feature will let the deployment system detect a regime transition underway and increase caution (smaller position sizes or suppression of signals) during the regime transition period. Both of these options have not been researched in the current study and were noted as potential future research directions in section 6.6.

5.7.3 Conditional Deployment as a Practical Resolution

As can be seen from the two issues stated above, the signal-to-cost ratio problem as well as walk-forward regime transition decay, both speak of one and the same solution: deployment conditional on certain circumstances and not always on. The basis for such approach is the bear trend sub- finding (IF-01 from section 4.5.4), which states that AI/ML strategies post mean Sharpe of +2.25 in bear trend folds compared to -0.18 of rule-based strategies on three of the four assets. Bear trend environment is exactly that regime in which (a) ML classifiers have the best calibration based on recent training data, (b) buy-and-hold and trend following rule-based strategies post the lowest results, and (c) cost pressure of increased turnover is partly compensated for by greater magnitude of price movements that result in increased trade profits.

Conditional deployment scheme will use ML signals only in cases when SMA-200 slope rule points to the presence of a bear trend or ranging market regime and otherwise use passive

exposure strategy of simple buy-and-hold of four assets in proportions. Such a switch would have kept the strategy passive through the two bull-run outlier folds (SOL fold 6 and XRP fold 10) and thereby neutralised the outlier-driven rule-based mean Sharpe advantage in the pooled H1 findings. Furthermore, conditional deployment would keep the deployment strategy active in bear trend folds where AI/ML strategy has the greatest advantage. Regime conditional deployment strategy is the highest-priority future research item (Item 1) from section 6.6 and highest priority non-critical deployment control in Table 5.1. This solution is not tested in this study and should be analyzed within walk-forward audit framework.

5.8 Implications for Practitioners and Researchers

5.8.1 Implications for Practitioners

Given the above results, crypto users interested in using AI/ML in trading must view such evidence as establishing an extremely high threshold for proof. In other words: (i) any AI/ML trading system requires robust, multi-asset, walk-forward testing, under realistic transaction costs, before risking any money on its performance; (ii) simple rule-based trading systems must be part of the test portfolio since the burden of proof is always on the AI/ML system to beat them; (iii) cost stress testing is absolutely necessary, particularly where there is high turnover; (iv) regime-specific AI/ML deployment is preferable to universal AI/ML deployment; (v) null results must be viewed as significant rather than failures in the test methodology; and (vi) drawdown management must come first, given that there were 40–67 per cent mean drawdowns for the different AI/ML strategy groups.

5.8.2 Implications for Researchers

For researchers, several priorities emerge. Reproducibility infrastructure, manifest tracking, configuration hashing, audit logs should be considered a standard rather than an option for empirical AI/ML trading research. Negative findings should be reported openly and accepted as scientific contribution. Using strong baselines for comparison, whether rule-based or machine learning models based on static regime assumptions, increases evidence requirements but produces better results. Testing in multiple asset and multiple regime configurations decreases the danger of single-asset/single-period generalization. And statistical inference has to respect

the limited cross-asset sample size: with $n = 4$ assets in pooled analyses, the minimum detectable effect size is about Cliff's $\delta = 0.20$, well above the small-to-negligible effect sizes seen in H1 and H3.

5.9 Hypothesis Decisions

Table 5.2 presents the ultimate decision for every tested hypothesis of the thesis. It should be noted that all decisions are made based on the pooled multi-asset empirical evidence, supplemented with individual asset-level findings and fold-level diagnostics. Empirical results do not confirm the major thesis claim, that is, the superiority of AI- and ML-based trading strategies in terms of risk-adjusted performance in cryptocurrency markets relative to simpler rule-based and benchmarking methods.

For H1, the hypothesis is not supported. The pooled bootstrap test gives a p-value of 0.9548 which does not provide any statistical grounds for making conclusions about the superiority of ML strategies over the rule-based benchmarks in terms of Sharpe performance. The result can be considered with moderate confidence, as the mean rule-based advantage was affected by outliers. The median results were more positive for ML, as it demonstrated a median advantage in all four assets, and there was even some evidence of conditional effectiveness of ML in bear-trend environment. This evidence is important, yet not sufficient to sustain the main hypothesis of the thesis.

For H2, the hypothesis is not supported with a very high level of confidence. The pooled bootstrap p-value was 1.0000 and the estimated drawdown delta is materially negative, being a mean difference of -29.66 percentage points. Confidence interval remained negative as well, which means that regime-aware AI/ML approaches were unable to reduce drawdowns relative to comparator strategies. The absence of the hypothesis support was also confirmed by the whole range of evidence base: there was no asset among four, no regime out of three and only 5 out of 60 folds supporting the hypothesis. Thus, there is a high level of confidence in the absence of the expected regime-aware drawdown benefit.

For H3, the hypothesis is not supported on its primary criterion. The pooled median Sharpe delta was -0.0956 , with a confidence interval of $[-0.193, -0.001]$ which means that the ensemble-based AI/ML approach was not able to produce superior Sharpe performance. The situation regarding drawdown is more optimistic – the drawdown reduction was observed across all four assets, with confidence interval $[-13.67, -4.20]$. The primary performance hypothesis is refuted, yet the drawdown control effect remains as an interesting secondary outcome of the study, provided that it never turned out to be a Sharpe performance driver.

The thesis claim, examined across three independently tested hypotheses, is not supported. In view of the fact that the primary tests of three independently designed hypotheses fail to support the thesis claim, it can be considered as a high confidence statement. However, the study provides a number of secondary findings – bear-trend conditional ML advantages, ensemble-based drawdown reduction, and out-of-fold stacking signals. AI/ML-based methods are capable of delivering value in bear-trend regimes and in drawdown reduction through ensemble aggregation, but not in the comprehensive way assumed by the main thesis claim.

Table 5.2 Final Hypothesis Decision Table

Hypothesis	Verdict	Confidence	Evidence Basis
H1	NOT SUPPORTED	MODERATE (fragile)	Bootstrap $p = 0.9548$; rule-based mean advantage outlier-driven; ML median advantage in 4 of 4 assets; bear-trend conditional advantage
H2	NOT SUPPORTED, strong evidence against	HIGH	Bootstrap $p = 1.0000$; mean DD delta -29.66 pp; CI entirely below zero; 0 of 4 assets, 0 of 3 regimes, 5 of 60 folds support
H3	NOT SUPPORTED on primary criterion	HIGH	Median Sharpe delta -0.0956 ; CI $[-0.193, -0.001]$; secondary drawdown reduction unanimous (4 of 4 assets, CI $[-13.67, -4.20]$)
Overall thesis claim	NOT SUPPORTED	HIGH	Three independent hypothesis tests against the claim; conditional sub-findings (bear-trend, ensemble drawdown reduction, stacking OOF) recorded with caveats

CHAPTER VI: SUMMARY, IMPLICATIONS, AND RECOMMENDATIONS

6.1 Summary of the Research

The thesis attempted to test, in highly rigorous and reproducible empirical conditions, whether AI/ML-based trading strategies have clear advantages over traditional rule-based and simpler benchmark strategies in cryptocurrency markets. It tested three independent and audited hypotheses for four big cryptocurrency perpetual futures contracts on Binance Futures, each 1 hour of resolution over a common out-of-sample period from 2 April 2022 to 1 January 2026, split into fifteen non-overlapping walk-forward folds per asset. To this end, H1 tested AI/ML strategies against rule-based strategies on the Sharpe-ratio basis, H2 tested a regime-aware AI/ML strategy versus a static benchmark pool on the maximum-drawdown basis, and H3 tested ensemble AI/ML strategies over single-model AI/ML strategies on the Sharpe-stability basis. The framework contained strict anti-leakage controls, six-case stress testing, Monte Carlo block bootstrap, regime breakdown, leave-one-asset-out sensitivity analysis, and an independent audit trail.

6.2 Answers to Research Questions

6.2.1 RQ1: Risk-Adjusted Outperformance

RQ1 examined if AI- and ML-based trading algorithms were more efficient than traditional rule-based strategies in the crypto markets in risk-adjusted terms. It can be noted that no evidence is found supporting the hypothesis at the primary mean-Sharpe criterion. The pooled bootstrap test resulted in a p value of 0.9548 with a mean Sharpe Ratio of -0.36 for AI/ML and +1.24 for rule-based strategies. Note that this mean comparison is driven by two significant outlier folds, SOL fold 6 and XRP fold 10, which account for about 0.24 per cent of the 840 primary H1 observations. Removing these outlier folds brings the mean comparison to parity. Interestingly, AI/ML strategies produce a higher median Sharpe Ratio across all the four assets, and a consistently better performance in bear trending markets which is evident from the Sharpe of +2.25 compared to -0.18 for rule-based strategies. The answer to RQ1 is therefore a qualified no:

AI/ML does not outperform on the primary criterion, but it shows conditional advantages on median performance and in bear-trend regimes that merit further research.

6.2.2 RQ2: Regime Performance

RQ2 asked how different classes of AI and ML algorithms perform across varying market regimes, trending, ranging and high-volatility conditions. The specific regime-aware architecture tested in H2, LightGBM with per-regime signal routing, did substantially worse than static benchmark strategies across all three regimes and all four assets. That is the strongest negative finding in the study. Separately, single-model ML strategies show a conditional advantage in bear-trend regimes within H1, but it does not carry over to bull-trend or ranging conditions. The answer to RQ2 is that AI/ML strategies are regime-conditional rather than universally robust: they add value in bear-trend conditions, lose decisively in bull-trend conditions, and do not improve robustness when regime conditioning is wired in through single-layer per-regime model routing.

6.2.3 RQ3: Practical Implementation

RQ3 asked what the practical limitations and implementation challenges of deploying AI-based trading strategies in real-world crypto trading actually are. The evidence flags seven main limitations: transaction cost sensitivity (AI/ML strategies generate three times the trade volume of rule-based ones, with the highest-turnover ensembles losing 80–90 per cent of gross PnL to costs); drawdown risk (mean maximum drawdowns of 40–67 per cent across AI/ML strategy groups exceed what is commercially acceptable); regime fragility (no AI/ML strategy holds a consistent positive Sharpe across all regimes); the backtest-to-live gap (no live or paper trading evidence exists for any tested strategy); the model governance and reproducibility burden (heavy engineering effort to maintain leakage controls, monitoring, retraining and audit); statistical inference limits ($n = 4$ assets in pooled analyses caps statistical power); and outlier sensitivity under fixed-notional sizing (extreme bull-run folds distort mean-based metrics). The answer to RQ3 is that deploying AI/ML strategies in crypto is constrained by costs, model instability, data quality, regime shifts and operational governance, all of which have to be sorted before any live capital goes in.

6.3 Contributions of the Study

The research makes five kinds of contribution. First, an empirical one: rigorous multi-asset, audited, walk-forward evidence on three distinct AI/ML claims in cryptocurrency markets, building a coherent evidence base that is currently rare in the literature. Second, methodological: the integrated walk-forward, bootstrap, Monte Carlo, cost-stress, regime-breakdown and leave-one-asset-out framework is a reusable template for cost-aware AI/ML strategy evaluation. Third, reproducibility: the manifest-tracked, version-controlled, and audit-traced evidence package makes replication and extension possible. Fourth, practical: the deployment controls in Table 5.1, paper trading, regime-conditional deployment, turnover budgets, drawdown kill-switches, volatility-scaled sizing, and drift monitoring, give practitioners a structured starting point for risk management. Fifth, the contribution of null findings: by reporting and contextualising three not supported hypotheses openly, the study responds to the publication and reporting biases that tend to favour positive AI/ML claims in the existing literature.

One methodological contribution worth mentioning is that reporting failed strategies as well as the successful ones. In nearly all published AI/ML trading research, a strategy that generates zero trades or crashes during execution is not mentioned. Keeping ML_04_xgboost and ML_09_lstm in the evidence base as documented failures is an example of that methodological contribution. Future researchers can see the full stack of planned implementations, not just the filtered set that produced numbers, and any practitioner reading this work gets a real sense of what a production AI/ML trading pipeline is actually like, that a meaningful fraction of well-designed strategies will not be able to come to life in real data under typical evaluation conditions.

A further contribution is a result of the systemic and market-structure literature covered in Chapter II: the deployment-controls framework in Table 5.1. None of these controls are novel on their own, in fact each one has already been addressed in the production trading and operational risk literature (see, for example, López de Prado, 2018 on leakage and audit controls and Roncalli, 2020 on drawdown and risk budgeting in live systems). The point here is to add them to a single list of relevant control elements and the empirical evidence of their significance. The structure can be reused by anyone who is starting to install AI/ML trading systems in crypto, regardless of whether such strategies are the ones we are testing.

Another contribution is the visual evidence package produced alongside the written thesis. The eighteen figures in this study are generated directly from the archived experiment outputs and CSV evidence files and are traceable to the audit list as well as in the statistical tables. This tight coupling between tabular evidence and visual presentation is not standard in cryptocurrency trading research, where figures are often generated from summary statistics rather than raw fold-level data. Every figure in the thesis can be reproduced by running the documented generation scripts against the archived experiment outputs, making the visual layer as reproducible as the statistical layer.

A final contribution is the cross-hypothesis integrated finding framework in section 4.5. The three hypotheses H1, H2 and H3 are designed to be independently testable and independently non-supporting, but the integrated sub-findings that emerge, namely IF-01 (AI/ML bear-trend advantage), IF-02 (regime routing worsens drawdown) and IF-03 (ensemble drawdown reduction), paint a more textured picture than the three binary verdicts alone. The introduction of a systematic framework for presenting conditional results in addition to the results regarding the testing of the primary hypotheses is a methodological advancement to the empirical financial ML research field. This framework makes sure that the basis for future research is more solid than just "three not-supported hypotheses": it defines which conditionalities are confirmed and which remain to be tested.

6.4 Study Limitations

The study's limitations are outlined in the methodology and results chapters and summarized in Table 6.1. The main ones are: a four-asset universe; a one-time study period with macroeconomic exposure; a backtest-only evidence base and no live or paper trading validation; fixed-notional position sizing that results in more outlier folds; the failure of two of the intended H3 base learners (XGBoost and LSTM), which reduced the diversity of H3 ensemble components from five to three; a single regime routing architecture used in H2; a non-standard Sharpe formula that is internally consistent but not directly comparable to the literature; the omission of perpetual futures funding rate costs; and a small ($n = 4$) sample of cross-asset models to make the pooled inference.

Table 6.1 Key Study Limitations and Severity

Limitation	Relevance	Severity
Four-asset universe (BTC, ETH, SOL, XRP)	All hypotheses	HIGH
Backtest-only evidence; no live validation	All hypotheses	HIGH
Single study period 2022–2026; shared macro exposure	All hypotheses	HIGH
Fixed-notional sizing creates implicit leverage	H1 strongly, H2/H3 moderately	HIGH
FAILED base learners (ML_04, ML_09)	H3 ensemble diversity	HIGH (H3-specific)
Single H2 regime-routing architecture (ML_07)	H2 generalisability	CRITICAL (H2-specific)
Non-standard Sharpe formula (CAGR / log-vol)	External comparability	MEDIUM
Perpetual futures funding not modelled	Cost accuracy	MEDIUM
Feature asymmetry (BTC 39 vs others 35)	Cross-asset model capacity	LOW
Two-bar execution lag in ML_07 (H2)	Conservative bias only	LOW
ML_06D sparse trading (≈ 10 –49/fold)	H3 best strategy uncertainty	MEDIUM
n = 4 assets for pooled inference	Statistical power	MEDIUM

One limitation to explore is the relationship between the fixed-notional position sizing and Sharpe-ratio measure used as the H1 and H3 primary criterion for the purpose of our analysis. Fixed-notional sizing means that even after a drawdown the strategy continues trading with the same dollar exposure as before. The implicit leverage that creates amplifies returns both positive and negative in the post-drawdown range. Coupled with the CAGR definition in the thesis's Sharpe formula and two extreme bull-run folds in SOL and XRP, this leads to a significant impact on the pooled mean Sharpe. The thesis tells us this openly, including the leave-one-asset-out and leave-one-fold-out sensitivity analysis which quantifies how fragile the rule-based mean advantage is. A volatility-scaled or constant fraction-of-equity position sizer would have had different Sharpe values, but we would have still found the directional verdicts and the ensemble drawdown finding.

Another limitation concerns the audit boundary. The audit framework ensures that the evidence chain from raw data to reported claims is always consistent. What it does not do is validate the correctness of the underlying research code, the reliability of the exchange data feed, or the absence of bugs in the dependency stack. Those software-engineering sides of reproducibility are

handled by version control, dependency pinning and unit testing rather than by audit, but they stay a residual source of risk. A practitioner wanting to deploy a strategy on the basis of these findings would need to re-derive the result from the audited evidence package and re-validate the production code under their own quality-assurance regime.

6.4.1 Implications by Industry Stakeholder

The findings carry implications for several industry stakeholders beyond the academic audience. Quantitative fund managers, retail trading platforms, cryptocurrency exchanges, regulators, and risk-management providers each face distinct decisions that the evidence here speaks to.

Quantitative fund managers deploying AI/ML strategies in crypto should treat the negative primary findings as setting a high evidentiary bar for any positive claim. The minimum viable evidence package for a deployment decision should include multi-asset walk-forward validation under cost-aware evaluation, leave-one-asset-out sensitivity, cost stress testing across realistic ranges, regime breakdown, and an audit trail from raw data to reported claims. Strategies that have passed through the above framework may be considered in allocating capital; those that are only based on single security or in-sample testing should be considered to be at the research stage and not ready for deployment. The deployment controls framework (Table 5.1) provides a systematic way forward from research into operational risk management.

The retail trading platform that sells AI/ML strategies to individuals bears a unique responsibility in ensuring that performance claims have been made under realistic conditions. The evidence shows backtest performance under optimistic assumptions can diverge clearly from performance under realistic friction regimes, and that out-of-sample performance can diverge from in-sample. Retail platform disclosures should make the gap between marketing-ready backtest figures and realistic live-trading performance plain to the end investor, in line with the investor protection principles that Bailey, Borwein, López de Prado and Zhu (2014) identify as essential safeguards against backtest overfit being marketed as live performance.

Cryptocurrency exchanges and infrastructure providers stand to gain from the documentation of the data pipeline and feature engineering process, which can feed exchange-side data-quality monitoring and product development. The systematic documentation of the H3 failed strategies (ML_04_xgboost and ML_09_lstm) is informative for infrastructure design too: deployment

platforms should expect a real fraction of strategies to fail at the signal-generation stage, and should build monitoring that catches zero-trade conditions fast.

Regulators and supervisory authorities worried about systemic risk and market abuse in AI-driven trading should note that the evidence here does not support claims of broad AI/ML outperformance in major crypto markets. The supervisory takeaway is that systemic-risk concerns from AI-driven trading homogenisation are less acute than worst-case scenarios imply, at least insofar as the AI/ML strategies tested here show no universal performance advantage of the kind that would drive concentrated adoption. The findings also back the supervisory interest in evidence-based disclosure regimes for AI-driven investment products, a concern explicitly raised by the Financial Stability Board (FSB, 2017) in its assessment of machine learning risks in financial services.

Risk-management service providers, the platform vendors for AI/ML model governance, drift monitoring, and audit-trail infrastructure, will find a direct demand signal in the deployment controls. The model-approval workflow, PSI and CSI drift monitoring, kill-switch architecture, retraining cadence and audit logging are not available in general-purpose data science platforms: it is mainly the specialised vendors who package these capabilities in the case of crypto trading that are very useful.

One limitation we did not mention in Table 6.1 is the influence of the walk-forward fold structure on the distribution of market regimes in the evaluation period. As the fold timeline in section 3.9 shows, bear-trend folds (F1-F4) are more abundant than bull-trend folds (F8-F11) and the bull-trend folds contain the two extreme outlier observations that dominate the pooled rule-based mean Sharpe. A different evaluation period, one beginning in 2020 and extending through the 2020–2021 bull cycle, would have produced a materially different distribution of regime exposures and might have altered the H1 primary verdict, the H2 per-regime breakdown, and the H3 cross-asset consistency findings. The study is explicitly bounded to the 2022–2026 period for data-availability and fold-count reasons, and readers should treat the regime distribution in the evaluation window as a conditioning factor on all three hypothesis verdicts.

A second limitation concerns the generalisability of the deployment-controls framework in Table 5.1 to institutional rather than retail execution contexts. The backtest engine uses fixed-notional position sizing at \$10,000 per trade, a scale appropriate for retail or small-institutional trading on

major centralised exchanges. Institutional deployers operating at significantly larger notional sizes would face material market-impact frictions not modelled here, which would widen the gap between backtest and live performance beyond what the cost-stress scenarios in section 4.2.5 capture. The deployment-controls framework is therefore most directly applicable to the retail-to-small-institutional scale at which the backtest assumptions are calibrated, and should be re-evaluated under institution-specific execution models before being applied at larger notional sizes.

6.5 Recommendations

6.5.1 For Practitioners

Practitioners should not deploy any of the tested AI/ML strategies in live trading on the thesis evidence alone. The recommended sequence for a live deployment is: (i) replicate the walk-forward backtest under the practitioner's own cost regime, including funding rates and venue-specific fee tiers; (ii) run a minimum three-month paper trading validation period, with live performance tracked against the backtest benchmark; (iii) put per-strategy drawdown kill-switches in place at conservative thresholds (no greater than 25 per cent peak-to-trough); (iv) restrict deployment to controlled-turnover architectures, leaving out the rank-based ensemble and similar high-turnover designs; (v) consider regime-conditional deployment (AI/ML signals in bear/ranging regimes, passive participation in bull regimes); and (vi) keep a governance framework, model approval, retraining cadence, and monitoring dashboards, running before any capital goes in, following the model risk management principles documented in the US Federal Reserve and OCC SR 11-7 guidance (Board of Governors of the Federal Reserve System, 2011).

6.5.2 For Researchers

Researchers should treat this framework as a baseline to build on. Extensions of particular value would include alternative regime-conditioning architectures (HMM, online changepoint detection, regime-adaptive ensemble weights); alternative label designs (longer horizons, triple-barrier methods, regression on forward returns); explainability analysis (SHAP-based feature attribution), for the bear-trend advantage especially; portfolio-level evaluation of the tested strategies as components of a multi-strategy portfolio; and adaptive ensemble weighting

schemes. Researchers should also adopt the manifest-tracked, audit-traced reproducibility infrastructure as a standard for future empirical AI/ML trading studies, and report null findings transparently as informative results.

6.5.3 For Future System Development

Future systems supporting AI/ML trading research and deployment in crypto should build in: per-asset cost models that reflect the differential liquidity and fee structures across BTC, ETH, SOL, XRP and beyond; volatility-scaled position sizing as an alternative to fixed-notional sizing; integrated funding-rate modelling for perpetual futures; on-chain and order-book microstructure features as extra inputs; real-time drift monitoring with auto-retraining and human-review gates (Gama et al., 2014, provide the theoretical basis for distinguishing gradual drift from abrupt regime change in data streams); and a paper-trading harness that runs the same code paths as live trading, with delayed market data and simulated fills.

Beyond the practical recommendations in section 6.5, two structural recommendations deserve emphasis. First, in case of any future research in AI/ML applications to cryptocurrency, reproducibility infrastructure must be manifest-tracked and audit-traced from day one, as opposed to being implemented after the first findings come out. The reproducibility infrastructure developed in this thesis is one that was created on the go, along with the empirical work, and a considerable amount of initial findings about audit came down to configuration drift between the run-time environment and the written-up version of things. Having the right design of the reproducibility infrastructure with claim-level provenance enforced at the moment of writing would have addressed the problem right away.

Second, this thesis proved just how much of the empirical result relies on the pre-strategy choices made. In future research, sensitivity analyses for each of these should be part of the methodology rather than relying on a single configuration. A good template for this would be a multi-panel verdict matrix with hypotheses in the rows and different approaches in the columns.

6.6 Future Research

Table 6.2 sets out a prioritised future research agenda derived directly from the findings. Fourteen items are recorded, grouped by hypothesis relevance and tractability.

Table 6.2 Future Research Agenda by Priority and Hypothesis

Priority	Item	Hypothesis	Tractability
1	Regime-conditional deployment test (ML in bear/ranging; B&H in bull)	H1	HIGH
2	Alternative H2 architectures (HMM, regime features, adaptive ensemble weights)	H2	MEDIUM
3	Paper trading validation for ML_06D and ML_02 (≥ 3 months)	H1, H3	HIGH
4	Repair FAILED base learners (ML_04, ML_09); retest H3 with full pool	H3	HIGH
5	Extended asset universe (AVAX, MATIC, LINK, ADA, BNB)	All	MEDIUM
6	Label horizon sensitivity (8h, 24h labels; regression labels)	H1	HIGH
7	SHAP-based bear-trend feature attribution	H1	HIGH
8	Portfolio-level multi-strategy allocation (ML_06D + ML_02 + passive)	H1, H3	MEDIUM
9	Adaptive ensemble weighting (online learning, decay-weighted)	H3	MEDIUM
10	Model governance framework documentation and operationalisation	Cross	HIGH
11	Crisis-period isolation (LUNA May 2022; FTX November 2022)	H1, H2	MEDIUM
12	Transformer / deep RL model architectures	H1	LOW
13	On-chain and order book microstructure features	H1	LOW
14	Live monitoring infrastructure (PSI/CSI drift, auto-retraining)	Cross	HIGH

6.6.1 Rationale for Selected Future Research Items

Item 1 in the future research agenda, the regime-conditional deployment test, has the strongest motivation behind it. The H1 evidence shows an AI/ML bear-trend advantage of roughly 2.43 Sharpe points over rule-based strategies in bear-trend folds, while the rule-based advantage in bull-trend folds is driven by the SOL and XRP outlier folds. Therefore a regime-conditional deployment, one that switches AI/ML signals on only in bear and ranging regimes and reverts to passive buy-and-hold in bull regimes, is a natural composite strategy. The test would compare

the regime-conditional composite with both AI/ML alone and rule-based alone, under the same walk-forward design used here.

Item 2, alternative H2 architectures, addresses the most decisive negative finding in the study. The H2 verdict is specific to the ML_07_regime_lgbm architecture and does not rule out positive results from other regime-conditioning designs. Hidden Markov Model regime classification, online changepoint detection, and regime-adaptive ensemble weighting are three architecturally distinct alternatives that should be passed through the same empirical framework. Each takes non-trivial implementation work and careful hyperparameter selection.

Item 3, paper trading validation for ML_06D and ML_02, addresses the biggest deployment-readiness gap. The two best-ranked AI/ML strategies in the study (ML_06D stacking OOF and ML_02 random forest) have backtest support but no live or paper-traded evidence. A minimum three-month paper-trading window on out-of-time data the model has not seen would give the first direct read on the live-versus-backtest performance gap.

Item 4, which would fix the broken H3 FAILED base learners, is easy to write and would answer directly the H3 sub-question, that the diversity of the ensemble is limited to a three-model active base learner pool. The next step would be to rebuild ML_04_xgboost with the original probability output and ML_09_lstm with corrected sequence length accuracy, and then re-test H3.

Item 5, the extended asset universe, would address the challenge of the lack of diversity of the data. By adding AVAX, MATIC, LINK, ADA and BNB, the number of assets would be almost twice what it is now, and the evidence from the different assets would be much higher. The cost of this would be about 2.25 times that of the current four-asset study.

Item 6, label horizon sensitivity, addresses the most important methodological decision that is made in every strategy. Testing the H1 verdict under 8-hour, 24-hour, and regression labels would establish whether the negative finding is an artefact of the four-bar threshold-based label or holds up across label specifications.

Item 7, SHAP-based bear-trend feature attribution, would turn the conditional H1 sub-finding into an interpretable explanation using the Shapley-value framework Lundberg and Lee (2017) developed for attributing individual predictions to input features. Compute SHAP values on the

AI/ML predictions in bear-trend folds, compare them against the SHAP values in bull-trend folds, and the analysis would identify which features drive the conditional advantage, and whether the bear-trend signal holds up under feature subset selection.

Item 8, portfolio-level multi-strategy allocation, would test the deployment hypothesis that a portfolio combining ML_06D stacking OOF, ML_02 random forest, and passive participation can beat any single strategy on risk-adjusted performance. The portfolio-level evaluation needs a capital allocation rule that is itself a research question; equal-weight, inverse-volatility-weighted, and Kelly-criterion allocation are three natural candidates.

Item 9, adaptive ensemble weighting, would explore whether online learning of base-learner weights, decay-weighted moving averages of validation performance, or regime-conditional weighting can deliver the Sharpe improvement that static ensembles never manage.

Item 10, model governance framework documentation, is the most operationally important item on the list. Moving from a research-grade backtest framework to a production-grade trading system demands explicit governance over model approval, retraining cadence, drift monitoring, kill-switch architecture, and audit-trail requirements. The reproducibility infrastructure of the thesis is necessary for that but not sufficient on its own. Item 12, transformer and deep reinforcement learning architectures, was deliberately deprioritised until the higher-priority items above are complete. The literature discussed in section 2.7 shows that attention-based architectures such as the Temporal Fusion Transformer (Lim et al., 2021) can outperform traditional models on pure forecasting benchmarks but the effectiveness of this approach in terms of cost-aware trading is still not clear. Testing the transformer or deep-RL approach using the same walk-forward, audited approach as we do would also be a good opportunity to test whether the architectural sophistication gap resolves once we make some realistic assumptions about how the model will work, or if it merely reproduces the accuracy versus trading-value gap documented in section 2.7 and Chapter IV.

6.7 Final Closing Statement

This DBA research shows that AI- and ML-based trading strategies in crypto markets have to be evaluated under rigorous, cost-aware, multi-asset, out-of-sample and audited conditions. Its contribution is to map not just where AI/ML methods may hold promise, in bear-trend regimes,

on median performance, and in drawdown reduction through ensemble aggregation, but also where their performance claims do not survive serious empirical scrutiny: on primary mean-Sharpe outperformance, on single-layer regime routing for drawdown control, and on Sharpe stability through ensemble construction. The conditional sub-findings are scientifically defensible and point to productive directions for future research and for careful, governance-anchored deployment. The negative primary findings put a check on overconfident claims of universal AI/ML superiority in crypto trading, and they reinforce a simple discipline: any AI/ML claim has to be tested against simple, low-turnover, low-parameter benchmarks under realistic cost conditions before anyone treats it as established. In the end, the most important contribution here may be the demonstration that null findings, reported transparently and audited rigorously, are valid and informative academic results.

BIBLIOGRAPHY

Alessandretti, L., ElBahrawy, A., Aiello, L.M. and Baronchelli, A. (2018). Anticipating cryptocurrency prices using machine learning, *Complexity*, 2018, Article 8983590. <https://doi.org/10.1155/2018/8983590>

Ang, A. and Bekaert, G. (2002). International asset allocation with regime shifts, *The Review of Financial Studies*, 15(4), pp. 1137–1187. <https://doi.org/10.1093/rfs/15.4.1137>

Aronson, D.R. (2007). Evidence-Based Technical Analysis: Applying the Scientific Method and Statistical Inference to Trading Signals. Hoboken, NJ: Wiley.

Atsalakis, G.S. and Valavanis, K.P. (2009). Forecasting stock market short-term trends using a neuro-fuzzy based methodology, *Expert Systems with Applications*, 36(7), pp. 10696–10707. <https://doi.org/10.1016/j.eswa.2009.02.043>

Bailey, D.H. and López de Prado, M. (2014). The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality, *Journal of Portfolio Management*, 40(5), pp. 94–107. <https://doi.org/10.3905/jpm.2014.40.5.094>

Bailey, D.H., Borwein, J.M., López de Prado, M. and Zhu, Q.J. (2014). Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance, *Notices of the American Mathematical Society*, 61(5), pp. 458–471.

Bailey, D.H., Borwein, J.M., López de Prado, M. and Zhu, Q.J. (2017). The probability of backtest overfitting, *Journal of Computational Finance*, 20(4), pp. 39–69. <https://doi.org/10.21314/JCF.2016.322>

Barberis, N. and Thaler, R. (2003). A survey of behavioral finance, in Constantinides, G.M., Harris, M. and Stulz, R.M. (eds.), *Handbook of the Economics of Finance, Volume 1B*. Amsterdam: Elsevier, pp. 1053–1128. [https://doi.org/10.1016/S1574-0102\(03\)01027-6](https://doi.org/10.1016/S1574-0102(03)01027-6)

Bariviera, A.F. (2017). The inefficiency of Bitcoin revisited: A dynamic approach, *Economics Letters*, 161, pp. 1–4. <https://doi.org/10.1016/j.econlet.2017.09.013>

Baur, D.G., Hong, K. and Lee, A.D. (2018). Bitcoin: Medium of exchange or speculative assets? *Journal of International Financial Markets, Institutions and Money*, 54, pp. 177–189. <https://doi.org/10.1016/j.intfin.2017.12.004>

Bergmeir, C. and Benítez, J.M. (2012). On the use of cross-validation for time series predictor evaluation, *Information Sciences*, 191, pp. 192–213. <https://doi.org/10.1016/j.ins.2011.12.028>

Board of Governors of the Federal Reserve System (2011). *SR 11-7: Guidance on Model Risk Management*. Washington, DC: Federal Reserve. <https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm>

- Bouteska, A., Abedin, M.Z., Hajek, P. and Yuan, K. (2024). Cryptocurrency price forecasting – A comparative analysis of ensemble learning and deep learning methods, *International Review of Financial Analysis*, 92, Article 103055. <https://doi.org/10.1016/j.irfa.2023.103055>
- Breiman, L. (2001). Random forests, *Machine Learning*, 45(1), pp. 5–32. <https://doi.org/10.1023/A:1010933404324>
- Brock, W., Lakonishok, J. and LeBaron, B. (1992). Simple technical trading rules and the stochastic properties of stock returns, *The Journal of Finance*, 47(5), pp. 1731–1764. <https://doi.org/10.1111/j.1540-6261.1992.tb04681.x>
- Chen, J. (2023). Analysis of Bitcoin price prediction using machine learning, *Journal of Risk and Financial Management*, 16(1), 51. <https://doi.org/10.3390/jrfm16010051>
- Chen, J. and Tsang, E.P.K. (2020). Classification of normal and abnormal regimes in financial markets, in Tsang, E.P.K. and Chen, S.H. (eds.), *Handbook of Financial Econometrics and Statistics*. Boca Raton: CRC Press, pp. 39–58. <https://doi.org/10.1201/9781003087595-4>
- Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, pp. 785–794. <https://doi.org/10.1145/2939672.2939785>
- Corbet, S., Meegan, A., Larkin, C., Lucey, B. and Yarovaya, L. (2018). Exploring the dynamic relationships between cryptocurrencies and other financial assets, *Economics Letters*, 165, pp. 28–34. <https://doi.org/10.1016/j.econlet.2018.01.004>
- De Long, J.B., Shleifer, A., Summers, L.H. and Waldmann, R.J. (1990). Noise trader risk in financial markets, *Journal of Political Economy*, 98(4), pp. 703–738. <https://doi.org/10.1086/261703>
- Detzel, A., Liu, H., Strauss, J., Zhou, G. and Zhu, Y. (2021). Learning and predictability via technical analysis: Evidence from bitcoin and stocks with hard-to-value fundamentals, *Financial Management*, 50(1), pp. 107–137. <https://doi.org/10.1111/fima.12310>
- Dietterich, T.G. (2000). Ensemble methods in machine learning, in *Proceedings of the First International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science*, Vol. 1857. Berlin: Springer, pp. 1–15. https://doi.org/10.1007/3-540-45014-9_1
- Fama, E.F. (1965). The behavior of stock-market prices, *The Journal of Business*, 38(1), pp. 34–105. <https://doi.org/10.1086/294743>
- Fama, E.F. (1970). Efficient capital markets: A review of theory and empirical work, *The Journal of Finance*, 25(2), pp. 383–417. <https://doi.org/10.2307/2325486>
- Fischer, T. and Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions, *European Journal of Operational Research*, 270(2), pp. 654–669. <https://doi.org/10.1016/j.ejor.2017.11.054>

Financial Stability Board (FSB) (2017). *Artificial Intelligence and Machine Learning in Financial Services: Market Developments and Financial Stability Implications*. Basel: FSB. <https://www.fsb.org/2017/11/artificial-intelligence-and-machine-learning-in-financial-service>

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M. and Bouchachia, A. (2014). A survey on concept drift adaptation, *ACM Computing Surveys*, 46(4), Article 44. <https://doi.org/10.1145/2523813>

Gandal, N., Hamrick, J.T., Moore, T. and Oberman, T. (2018). Price manipulation in the Bitcoin ecosystem, *Journal of Monetary Economics*, 95, pp. 86–96. <https://doi.org/10.1016/j.jmoneco.2017.12.004>

Gu, S., Kelly, B. and Xiu, D. (2020). Empirical asset pricing via machine learning, *Review of Financial Studies*, 33(5), pp. 2223–2273. <https://doi.org/10.1093/rfs/hhaa009>

Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle, *Econometrica*, 57(2), pp. 357–384. <https://doi.org/10.2307/1912559>

Hansen, P.R. (2005). A test for superior predictive ability, *Journal of Business & Economic Statistics*, 23(4), pp. 365–380. <https://doi.org/10.1198/073500105000000063>

Harvey, C.R., Liu, Y. and Zhu, H. (2016). ... and the cross-section of expected returns, *The Review of Financial Studies*, 29(1), pp. 5–68. <https://doi.org/10.1093/rfs/hhv059>

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory, *Neural Computation*, 9(8), pp. 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>

Holm, S. (1979). A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics*, 6(2), pp. 65–70.

Hurst, B., Ooi, Y.H. and Pedersen, L.H. (2017). A century of evidence on trend-following investing, *The Journal of Portfolio Management*, 44(1), pp. 15–29. <https://doi.org/10.3905/jpm.2017.44.1.015>

Jaquart, P., Dann, D. and Weinhardt, C. (2021). Short-term bitcoin market prediction via machine learning, *The Journal of Finance and Data Science*, 7, pp. 45–66. <https://doi.org/10.1016/j.jfds.2021.03.001>

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk, *Econometrica*, 47(2), pp. 263–291. <https://doi.org/10.2307/1914185>

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y. (2017). LightGBM: A highly efficient gradient boosting decision tree, in *Advances in Neural Information Processing Systems*, 30, pp. 3146–3154.

Krauss, C., Do, X.A. and Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500, *European Journal of Operational Research*, 259(2), pp. 689–702. <https://doi.org/10.1016/j.ejor.2016.10.031>

- Lim, B., Arık, S.Ö., Loeff, N. and Pfister, T. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting, *International Journal of Forecasting*, 37(4), pp. 1748–1764. <https://doi.org/10.1016/j.ijforecast.2021.03.012>
- Lo, A.W. (2002). The statistics of Sharpe ratios, *Financial Analysts Journal*, 58(4), pp. 36–52. <https://doi.org/10.2469/faj.v58.n4.2453>
- Lo, A.W. (2004). The adaptive markets hypothesis: Market efficiency from an evolutionary perspective, *Journal of Portfolio Management*, 30(5), pp. 15–29. <https://doi.org/10.3905/jpm.2004.442611>
- Lo, A.W. (2005). Reconciling efficient markets with behavioral finance: The adaptive markets hypothesis, *Journal of Investment Consulting*, 7(2), pp. 21–44.
- Lundberg, S.M. and Lee, S.I. (2017). A unified approach to interpreting model predictions, in *Advances in Neural Information Processing Systems 30*, Long Beach, CA, pp. 4765–4774.
- López de Prado, M. (2018). *Advances in Financial Machine Learning*. Hoboken, NJ: Wiley. <https://doi.org/10.1002/9781119482086>
- López de Prado, M. (2020). *Machine Learning for Asset Managers*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781108883658>
- McNally, S., Roche, J. and Caton, S. (2018). Predicting the price of Bitcoin using machine learning, in *Proceedings of the 26th Euromicro International Conference on Parallel, Distributed and Network-Based Processing*, Cambridge, UK, pp. 339–343. <https://doi.org/10.1109/PDP2018.2018.00060>
- Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. *White Paper*. <https://bitcoin.org/bitcoin.pdf>
- Nadarajah, S. and Chu, J. (2017). On the inefficiency of Bitcoin, *Economics Letters*, 150, pp. 6–9. <https://doi.org/10.1016/j.econlet.2017.01.022>
- Neely, C.J., Rapach, D.E., Tu, J. and Zhou, G. (2014). Forecasting the equity risk premium: The role of technical indicators, *Management Science*, 60(7), pp. 1772–1791. <https://doi.org/10.1287/mnsc.2013.1838>
- Nosek, B.A., Ebersole, C.R., DeHaven, A.C. and Mellor, D.T. (2018). The preregistration revolution, *Proceedings of the National Academy of Sciences*, 115(11), pp. 2600–2606. <https://doi.org/10.1073/pnas.1708274114>
- Nystrup, P., Hansen, B.W., Madsen, H. and Lindström, E. (2017). Dynamic portfolio optimization across hidden market regimes, *Quantitative Finance*, 18(1), pp. 83–95. <https://doi.org/10.1080/14697688.2017.1342857>

- Park, C.H. and Irwin, S.H. (2007). What do we know about the profitability of technical analysis? *Journal of Economic Surveys*, 21(4), pp. 786–826. <https://doi.org/10.1111/j.1467-6419.2007.00519.x>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research*, 12, pp. 2825–2830.
- Pring, M.J. (2002). *Technical Analysis Explained*, 4th edn. New York: McGraw-Hill.
- Romano, J.P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping, *Econometrica*, 73(4), pp. 1237–1282. <https://doi.org/10.1111/j.1468-0262.2005.00615.x>
- Roncalli, T. (2020). *Handbook on Financial Risk Management*. Boca Raton, FL: CRC Press.
- Samuelson, P.A. (1965). Proof that properly anticipated prices fluctuate randomly, *Industrial Management Review*, 6(2), pp. 41–49.
- Sebastião, H. and Godinho, P. (2021). Forecasting and trading cryptocurrencies with machine learning under changing market conditions, *Financial Innovation*, 7(1), 3. <https://doi.org/10.1186/s40854-020-00217-x>
- Sharpe, W.F. (1966). Mutual fund performance, *The Journal of Business*, 39(1), pp. 119–138. <https://doi.org/10.1086/294846>
- Shiller, R.J. (2017). Narrative economics, *American Economic Review*, 107(4), pp. 967–1004. <https://doi.org/10.1257/aer.107.4.967>
- Sirignano, J. and Cont, R. (2019). Universal features of price formation in financial markets: Perspectives from deep learning, *Quantitative Finance*, 19(9), pp. 1449–1459. <https://doi.org/10.1080/14697688.2019.1622295>
- Sullivan, R., Timmermann, A. and White, H. (1999). Data-snooping, technical trading rule performance, and the bootstrap, *The Journal of Finance*, 54(5), pp. 1647–1691. <https://doi.org/10.1111/0022-1082.00163>
- Urquhart, A. and Hudson, R. (2013). Efficient or adaptive markets? Evidence from major stock markets using very long run historic data, *International Review of Financial Analysis*, 28, pp. 130–142. <https://doi.org/10.1016/j.irfa.2013.03.005>
- Urquhart, A. (2016). The inefficiency of Bitcoin, *Economics Letters*, 148, pp. 80–82. <https://doi.org/10.1016/j.econlet.2016.09.019>
- Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J. and Sun, L. (2023). Transformers in time series: A survey, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI 2023)*, pp. 6778–6786. <https://doi.org/10.24963/ijcai.2023/759>

White, H. (2000). A reality check for data snooping, *Econometrica*, 68(5), pp. 1097–1126.
<https://doi.org/10.1111/1468-0262.00152>

Wilder, J.W. (1978). *New Concepts in Technical Trading Systems*. Greensboro, NC: Trend Research.

Wolpert, D.H. (1992). Stacked generalization, *Neural Networks*, 5(2), pp. 241–259.
[https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1)

APPENDIX A: FULL STRATEGY INVENTORY

A complete inventory of the strategy implementations evaluated in this research must be read by hypothesis namespace rather than as a single flat ID list. The rule-based group contains eight families (RB_01 buy-and-hold; RB_02 moving-average crossover with five configurations; RB_03 RSI mean-reversion with eighteen configurations; RB_04 Bollinger Band mean-reversion with four configurations; RB_05 Donchian breakout with three configurations; RB_06 MACD trend with fixed parameters; RB_07 volatility breakout with six configurations; RB_08 EMA trend-filter with two configurations). H1/H2 active AI/ML implementations include ML_01_logistic_regression, ML_02_random_forest, ML_03_gradient_boosting, ML_04_svm, ML_05_gru, ML_06_ensemble, and ML_07_regime_lgbm. H3 active single-model comparators include ML_01_logistic_regression, ML_02_random_forest, ML_03_gradient_boosting, ML_05_lightgbm, ML_07_svm, and ML_08_mlp. H3 ensembles include ML_06A_majority_vote, ML_06B_weighted_vote, ML_06C_soft_vote, ML_06D_stacking_oof, ML_06E_strategy_signal_ensemble, ML_06F_rank_based_ensemble, and ML_06G_regime_aware_ensemble, with ML_06G reported as secondary evidence only. H3 failed strategies are ML_04_xgboost and ML_09_lstm.

Because strategy IDs get reused across experiment runners, the inventory should be read through the full strategy identifier and hypothesis role, not the numeric prefix alone. For instance, ML_04_svm is an active H1/H2 implementation, while ML_04_xgboost is a failed H3 implementation; ML_07_regime_lgbm is the H1/H2 regime-aware LightGBM implementation and H2 treatment, while ML_07_svm is an H3 single-model comparator. The detailed hypothesis-asset-strategy mapping is provided in Appendix M.

Each strategy implementation is uniquely identified by a strategy identifier, an experiment identifier and a configuration hash. The configuration hashes mean any future reproduction can be matched against the exact parameter set used here. The audit and evidence-management infrastructure works on strategy identifiers and source files rather than shortened display names; the names attached to the strategy tables are presentation labels, not complete identifiers.

Source: Crypto_Research/H1/H1_MULTI_ASSET_AI_ML_vs_RuleBased_Sharpe_v1/,
Crypto_Research/H2/H2_MULTI_ASSET_RegimeAware_AI_ML_vs_Static_Strategies_v1/,
Crypto_Research/H3/H3_MULTI_ASSET_EnsembleML_vs_SingleModel_Stability_v1/.

APPENDIX B: DATA PIPELINE AND SCHEMA

The data pipeline is fully documented in section 3.3 (Data Preparation and Canonical Schema) and Table 3.2 of the main text. Briefly: raw exchange feeds for all four perpetual futures are normalised into a common schema before any feature or model code touches them. Bars with gaps are synthetically filled, flagged, and excluded from model training to prevent imputed prices from entering the learning process. The four-feature difference between BTC (39 features) and the other three assets (35 features) traces to open-interest data availability on Binance Futures at the time of collection: BTC had a clean, complete series; ETH, SOL and XRP did not.

Source: `Crypto_Research/Methodology/METHODOLOGY_END_TO_END_RESEARCH.md`,
`Crypto_Research/Methodology/METHODOLOGY_DATA_LOADING.md`.

APPENDIX C: FEATURE ENGINEERING DETAILS

The full feature set is catalogued in Table 3.3 and the accompanying text in section 3.4. All features are computed as backward-looking transformations of the canonical OHLCV-plus-derivatives data, with the point-in-time invariants described in that section enforced at the code level. The categories covered are: multi-horizon log returns, moving-average levels and slopes, momentum oscillators (RSI, MACD, rate of change), volatility measures (ATR, Bollinger width, rolling realised volatility), candle body and wick structure, volume z-scores and ratios, open-interest derivatives for BTC only, and three deterministic regime classification features shared by all assets.

The three regime features (`trend_regime`, `volatility_regime`, `drawdown_regime`) are discussed at length in section 3.4 of the main text, including the design rationale for the threshold values and window lengths. Their construction as deterministic rules rather than learned outputs is a deliberate choice documented there: it keeps the H2 regime-routing architecture interpretable and avoids adding a second layer of overfitting risk on top of the primary ML models.

The anti-leakage enforcement that governs the feature pipeline is explained in section 3.4 and illustrated in Figure 3.1. Assertion checks at the end of each pipeline run confirm that no column could have required future data, providing a programmatic guarantee rather than a documentation convention.

Source: `Crypto_Research/Methodology/METHODOLOGY_FEATURE_ENGINEERING.md`.

APPENDIX D: BACKTEST ENGINE ASSUMPTIONS

The backtest engine design and assumptions are set out in section 3.7 and Table 3.5 of the main text. Key parameters for traceability: starting capital \$10,000 per fold, fixed-notional sizing at \$10,000 per trade, base round-trip cost of 16 bps (6 bps commission plus 2 bps slippage per side), one-bar execution lag, and no stop-loss or take-profit rules. Perpetual futures funding costs are not modelled, which is documented as a material limitation in section 3.13. The six cost-stress scenarios used in robustness testing span from the base to a high-friction scenario of 12 bps commission plus 10 bps slippage per side.

Source: `Crypto_Research/Methodology/METHODOLOGY_BACKTEST_FRAMEWORK.md`.

APPENDIX E: H1 DETAILED RESULT TABLES

Detailed H1 result tables include: per-asset fold-by-fold Sharpe ratios; per-asset best-of-strategy ranking; cost-stress sensitivity per asset and pooled; regime breakdown per asset and pooled; Monte Carlo block bootstrap distributions; leave-one-asset-out sensitivity outputs.

The H1 detailed result tables, in the audited evidence package at `Crypto_Research/H1/H1_MULTI_ASSET_AI_ML_vs_RuleBased_Sharpe_v1/results/`, contain 900 total H1 fold-level observations: 15 H1 strategy implementations across four assets and 15 folds per asset. The primary H1 statistical test uses 840 observations after excluding `ML_06_ensemble` from the AI/ML primary group, 480 rule-based observations and 360 AI/ML primary observations. The pooled multi-asset analysis runs the full battery of statistical tests on this primary evidence pool.

Cost-stress sensitivity for H1 is reported across seven scenarios (base plus six alternatives) per asset and pooled. The pooled result is dominated by the SOL fold 6 and XRP fold 10 outliers in all seven scenarios, but the ETH-only result tells a different story, with AI/ML strategies outperforming rule-based strategies in the six non-base scenarios. That ETH-only observation fits the broader thesis reading: AI/ML advantage is conditional, not universal.

Source: `Crypto_Research/H1/H1_MULTI_ASSET_AI_ML_vs_RuleBased_Sharpe_v1/results/` and

`Crypto_Research/H1/H1_{BTC,ETH,SOL,XRP}_AI_ML_vs_RuleBased_Sharpe_v1_strict_audit/results/`.

APPENDIX F: H2 DETAILED RESULT TABLES

Detailed H2 result tables include: per-asset fold-by-fold maximum drawdown; per-asset regime-aware-versus-static deltas; cost-stress per asset and pooled; per-regime drawdown breakdowns; Monte Carlo Sharpe p5/p25/p50/p75/p95 percentiles; leave-one-asset-out sensitivity outputs.

The H2 detailed result tables, at [Crypto_Research/H2/H2_MULTI_ASSET_RegimeAware_AI_ML_vs_Static_Strategies_v1/results/](#), report per-asset fold-by-fold maximum drawdown values for the regime-aware ML_07_regime_lgbm strategy and the static benchmark pool. That benchmark pool deliberately includes regime-feature-informed AI strategies on top of the rule-based ones, which raises the bar for H2 support and means any positive H2 finding would be especially defensible. The decisive negative finding (mean drawdown delta minus 29.66 percentage points; bootstrap confidence interval [minus 35.83, minus 23.48]) holds in every alternative cost scenario, every market regime, and every leave-one-asset-out perturbation.

Source:

[Crypto_Research/H2/H2_MULTI_ASSET_RegimeAware_AI_ML_vs_Static_Strategies_v1/results/](#) and

[Crypto_Research/H2/H2_{BTC,ETH,SOL,XRP}_RegimeAware_AI_ML_vs_Static_Strategies_v1/results/](#).

APPENDIX G: H3 DETAILED RESULT TABLES

Detailed H3 result tables include: ensemble-versus-single-model deltas per asset; ranking of all 13 active strategies; drawdown reduction per asset and pooled; ensemble-construction architecture comparisons; sparse-trading sensitivity for ML_06D stacking OOF; regime-aware-ensemble (ML_06G) secondary results.

The H3 detailed result tables, at [Crypto_Research/H3/H3_MULTI_ASSET_EnsembleML_vs_SingleModel_Stability_v1/results/](#), include ensemble-versus-single-model deltas per asset and pooled across the four. The ranking of all thirteen ranked H3 strategies reports six primary ensemble strategies, one secondary regime-aware ensemble, and six active single-model strategies. Take just the H3 primary set and the comparison is six primary ensembles versus six active single-model strategies; ML_06G is left out of the primary group because it combines regime conditioning with ensemble aggregation. The median-rank measure shrugs off outlier rank fluctuations in any single asset and gives a stable ordering of strategies that feeds the deployment recommendations in Chapter VI.

Sparse-trading sensitivity for ML_06D stacking OOF deserves a note. ML_06D produces between ten and forty-nine trades per fold, far fewer than the median ML strategy (about 272 trades per fold). That sparse profile comes from the stacking architecture's high probability threshold for ensemble agreement, the meta-learner only fires a trading signal when multiple base learners concur. As a result, ML_06D's per-fold Sharpe and drawdown metrics carry wider confidence intervals than denser-trading strategies, and its ranking advantage should be read with that caution in mind. A practical deployment built on ML_06D would need explicit handling of sparse-trading periods, for example capital allocation rules that scale exposure to the number of recent signals.

Source:

[Crypto_Research/H3/H3_MULTI_ASSET_EnsembleML_vs_SingleModel_Stability_v1/results/](#)
and

[Crypto_Research/H3/H3_{BTC,ETH,SOL,XRP}_EnsembleML_vs_SingleModel_Stability_v1/results/](#).

APPENDIX H: STATISTICAL TEST OUTPUTS

Section 3.10 (Statistical Testing) and Table 3.7 document the full test battery applied to each hypothesis. For traceability: primary bootstrap inference used 1,000 resamples at the asset level and 5,000 for pooled analyses, all with random seed 42. Secondary tests were Welch t-test and Mann-Whitney U for mean and rank comparisons respectively. Effect sizes are Cohen's d (parametric) and Cliff's delta (non-parametric). Multiple-testing correction within each hypothesis used the Holm-Sidak stepdown procedure; no cross-hypothesis correction was applied on the grounds that H1, H2 and H3 are pre-specified, independently motivated tests.

Source: `Crypto_Research/H1_H2_H3/final_synthesis/V1/cross_hypothesis_tables_for_thesis.csv` and per-hypothesis statistical output files.

APPENDIX I: AUDIT REPORTS AND CLAIM GOVERNANCE

The audit hierarchy has three layers: per-asset audit reports, pooled multi-asset audit reports, and a cross-hypothesis audit report. The cross-hypothesis synthesis identifier is SYN_H1H2H3_20260515_001 (PASS after remediation of all five issues). The H1 pooled audit identifier is POOL_20260513_162408_63acb78 (PASS, twelve issues closed). The H2 pooled audit returns ACCEPTED_FOR_THESIS_USE. The H3 pooled audit identifier is AUD_POOL_20260515_001 (PASS after remediation).

Source:

Crypto_Research/H1_H2_H3/final_synthesis/V1/final_evidence_chapter/audit/dissertation_chapter_strict_audit_report.md and the per-hypothesis audit_report.md files under each pooled experiment directory.

APPENDIX J: EXPERIMENT IDs, CONFIGS, AND MANIFESTS

Each experiment is identified by a unique manifest covering the experiment configuration hash, git commit hash, software versions, strategy identifiers, evaluation period, random seeds and all reported metric values. Manifest files sit alongside each experiment's results directory.

Source: `Crypto_Research/H{1,2,3}/H{1,2,3}_*/**/*manifest.json` files in each results directory

APPENDIX K: LIMITATIONS AND METHODOLOGY DISCLOSURES

The limitations governing interpretation of the results are consolidated in section 6.4 and Table 6.1, which should be the primary reference. For quick traceability, the disclosures that recur across all result chapters are: the Sharpe ratio is defined as CAGR over annualised log-return volatility rather than the standard academic mean-over-standard-deviation form, so numerical values cannot be directly compared with externally reported Sharpe ratios; perpetual futures funding costs are excluded from the backtest engine, which biases long-bias strategies favourably; and fixed-notional position sizing at \$10,000 per trade creates implicit leverage whenever cumulative drawdown reduces the available equity below that notional. H3-specific: ML_04_xgboost and ML_09_lstm produced zero trades throughout; these are H3 failures only and have no bearing on the active H1/H2 implementations of ML_04_svm and ML_05_gru. The regime-aware ML_07_regime_lgbm also carries a two-bar execution lag from a signal-pipeline defect that was identified and disclosed in the audit report; this makes its reported drawdown a conservative estimate rather than a best-case one.

Source: `Crypto_Research/H1_H2_H3/final_synthesis/V1/cross_hypothesis_synthesis_report.md`
and

`Crypto_Research/H1_H2_H3/final_synthesis/V1/cross_hypothesis_executive_conclusion.md`.

APPENDIX L: GLOSSARY OF KEY TERMS

This glossary defines the principal technical terms used throughout the thesis. The definitions are scoped to how each term is used in the thesis; for broader treatments, the standard finance and machine learning textbooks should be consulted.

Term	Definition
Audit	An independent review of an experiment's data, methods, statistical inference, and claim traceability. In the present research, every reported result is supported by an audit report with a binding finding.
Bootstrap	A resampling procedure that generates an empirical sampling distribution by drawing repeatedly from the observed data. Confidence intervals for Sharpe-ratio differences in this thesis are built from 1,000 to 5,000 bootstrap draws at the strategy-fold level, so the within-fold return dependence is preserved.
Cost-stress test	An evaluation of strategy performance under alternative cost regimes that span plausible execution scenarios. In the present work, seven cost scenarios are evaluated, from a base of 6 basis points commission plus 2 basis points slippage to a stress of 12 basis points commission plus 10 basis points slippage.
Drawdown (maximum)	How far an equity curve falls from its running high before recovering, expressed as a percentage. Maximum drawdown is the worst such decline over a given evaluation window; across the fifteen walk-forward folds, the per-fold maximum drawdown is averaged to produce the H2 primary metric.
Embargo	A gap, measured in bars, between the training window end and the validation window start, used to prevent label leakage when

	the label depends on a forward-return horizon. In the thesis experiments, the embargo is twenty-four bars.
Ensemble	A combination of multiple base learners whose predictions are aggregated through voting, weighted averaging, stacking, or rank-based combination. Seven ensemble variants are evaluated in the H3 experiment (ML_06A through ML_06G).
Fold	One of the fifteen non-overlapping three-month test windows into which the 2 April 2022 to 1 January 2026 evaluation period is divided. Each fold has its own training and validation history, its own fitted model weights, and its own out-of-sample performance record.
Funding rate	A periodic payment exchanged between long and short positions in a perpetual futures contract to align the perpetual price to the underlying index price. Funding rates are not modelled in this study's backtest engine.
Holm-Sidak correction	A sequential correction that adjusts p-value thresholds across a family of tests to keep the probability of any false rejection below the target level, without the blanket conservatism of a Bonferroni adjustment. Within each hypothesis, it is applied to the four per-asset bootstrap tests to account for the fact that the same hypothesis is being probed four times.
Look-ahead bias	The contamination of a backtest when data from the future creeps into the training or prediction process, typically through a feature calculation that references a bar that would not yet have existed at the time of the signal. The feature pipeline here prevents this at the code level through assertion checks that verify no column value requires data beyond the bar being processed.

Manifest	A versioned record file generated automatically when an experiment runs, capturing the configuration hash, git commit reference, software environment, random seeds used, and every reported metric value. Manifests are the mechanism by which any reported number in the thesis can be traced back to the exact code and parameters that produced it.
Monte Carlo block bootstrap	A resampling procedure that draws contiguous blocks of observations rather than individual points, so the short-run autocorrelation structure of the return series survives the reshuffling. Five-day blocks (120 hourly bars) are used because autocorrelation in hourly crypto returns decays substantially by that horizon, making the block large enough to preserve dependence without being so large that meaningful path diversity is lost.
Out-of-sample (OOS)	Data that the model has never seen during training or hyperparameter selection. The common out-of-sample window runs from 2 April 2022 to 1 January 2026 across all four assets; within that window, fold-level test periods are further protected by embargo bars that prevent label leakage at the fold boundary.
Regime-aware strategy	A strategy whose signals depend on a detected market regime, typically through per-regime sub-models or regime-conditional feature weighting. The H2 treatment strategy ML_07_regime_lgbm uses per-regime LightGBM sub-models.
Sharpe ratio (in this study)	Compound annual growth rate divided by the annualised standard deviation of log returns. Differs from the academic convention of mean return over return standard deviation but is internally consistent across all comparisons in this study.
Stacking out-of-fold	An ensemble approach where a second-stage classifier learns to

	combine base learners by training on predictions those base learners made on data they were not trained on. Generating predictions out-of-fold rather than in-sample prevents the second stage from simply memorising whatever worked best in training. In H3, ML_06D implements this with logistic regression as the second-stage model.
Walk-forward validation	An out-of-sample evaluation discipline that keeps the chronological order of training and testing intact: the model trains on historical data up to some cutoff, is evaluated on the period immediately after, and the whole process repeats as the cutoff advances. Here, fifteen such cycles are run per asset, each using a twelve-month training window and a three-month test window separated by a twenty-four bar embargo.

APPENDIX M: HYPOTHESIS-ASSET-STRATEGY MAPPING

This appendix provides the detailed mapping between the formal hypotheses, the asset universe, the strategy groups, and the evidence used in the thesis. It is included in the main dissertation rather than kept as a separate file so that the reader can trace every hypothesis decision back to the strategies that produced the reported metrics. The mapping should be read by hypothesis namespace. Some strategy numbers are reused across runners, but the full strategy identifier and role remove that ambiguity.

The appendix is intentionally practical. It does not repeat every fold-level value from the experiment outputs. Instead, it states which strategy or strategy group entered each hypothesis test, what role it played, and which metrics were used to judge the claim. The detailed numerical outputs remain in the audited result files, while this appendix acts as the bridge between those files and the written findings.

A few conventions are worth noting before reading the mapping. First, RB identifiers refer to rule-based strategy families, some with several validation-selected parameter configurations. Second, ML identifiers must be read inside their hypothesis namespace. Third, H3 contains two failed candidate strategies, ML_04_xgboost and ML_09_lstm. They are shown because their failure is part of the empirical record, but they do not enter the active single-model comparison. Fourth, ML_06G is secondary in H3 because it combines ensemble aggregation with regime conditioning.

The mapping table below is unnumbered and is not part of the formal List of Tables. Its purpose is traceability rather than presentation. It records the strategy role at the level needed to understand the evidence chain without forcing the main chapters to carry a long technical table.

Hypothesis	Asset	Strategy or group	Role in test	Primary metrics	Evidence use
H1	BTCUSDT	RB_01 buy-and-hold	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; especially relevant in bull-trend folds
H1	BTCUSDT	RB_02 moving-average crossover	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; captures trend-following behaviour
H1	BTCUSDT	RB_03 RSI mean-reversion	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; tests mean-reversion behaviour
H1	BTCUSDT	RB_04 Bollinger Band mean-reversion	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; volatility-adjusted mean reversion
H1	BTCUSDT	RB_05 Donchian breakout	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; breakout and momentum behaviour

Hypothesis	Asset	Strategy or group	Role in test	Primary metrics	Evidence use
H1	BTCUSDT	RB_06 MACD trend	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; common trend and momentum benchmark
H1	BTCUSDT	RB_07 volatility breakout	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; high-turnover breakout family
H1	BTCUSDT	RB_08 EMA trend filter	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; trend-filtered rule-based family
H1	BTCUSDT	ML_01 logistic regression	AI/ML primary strategy	Fold-level Sharpe, median Sharpe, drawdown	Low-variance linear ML baseline
H1	BTCUSDT	ML_02 random forest	AI/ML primary strategy	Fold-level Sharpe, median Sharpe, drawdown	Tree-based non-linear classifier
H1	BTCUSDT	ML_03 gradient boosting	AI/ML primary strategy	Fold-level Sharpe, median Sharpe, drawdown	Boosted-tree classifier
H1	BTCUSDT	ML_04_svm	AI/ML primary strategy	Fold-level Sharpe, median Sharpe, drawdown	Support vector machine classifier
H1	BTCUSDT	ML_05_gru	AI/ML primary strategy	Fold-level Sharpe, median Sharpe, drawdown	Recurrent model used in the H1/H2 namespace
H1	BTCUSDT	ML_07_regime_lgbm	AI/ML primary strategy	Fold-level Sharpe, median Sharpe, drawdown	Regime-aware LightGBM included in H1 primary set
H1	BTCUSDT	ML_06_ensemble	Supplemental AI/ML evidence	Sharpe and robustness metrics	Reported but excluded from primary H1 tests to avoid overlap with H3
H1	ETHUSDT	RB_01 buy-and-hold	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; especially relevant in bull-trend folds
H1	ETHUSDT	RB_02 moving-average crossover	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; captures trend-following behaviour
H1	ETHUSDT	RB_03 RSI mean-reversion	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; tests mean-reversion behaviour
H1	ETHUSDT	RB_04 Bollinger Band mean-reversion	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; volatility-adjusted mean reversion
H1	ETHUSDT	RB_05 Donchian breakout	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; breakout and momentum behaviour
H1	ETHUSDT	RB_06 MACD trend	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; common trend and momentum benchmark
H1	ETHUSDT	RB_07 volatility breakout	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; high-turnover breakout family
H1	ETHUSDT	RB_08 EMA trend filter	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; trend-filtered rule-based family
H1	ETHUSDT	ML_01 logistic regression	AI/ML primary strategy	Fold-level Sharpe, median Sharpe, drawdown	Low-variance linear ML baseline
H1	ETHUSDT	ML_02 random forest	AI/ML primary strategy	Fold-level Sharpe, median Sharpe, drawdown	Tree-based non-linear classifier
H1	ETHUSDT	ML_03 gradient boosting	AI/ML primary strategy	Fold-level Sharpe, median Sharpe, drawdown	Boosted-tree classifier
H1	ETHUSDT	ML_04_svm	AI/ML primary	Fold-level Sharpe,	Support vector

Hypothesis	Asset	Strategy or group	Role in test	Primary metrics	Evidence use
			strategy	median Sharpe, drawdown	machine classifier
H1	ETHUSDT	ML_05_gru	AI/ML primary strategy	Fold-level Sharpe, median Sharpe, drawdown	Recurrent model used in the H1/H2 namespace
H1	ETHUSDT	ML_07_regime_lgbm	AI/ML primary strategy	Fold-level Sharpe, median Sharpe, drawdown	Regime-aware LightGBM included in H1 primary set
H1	ETHUSDT	ML_06_ensemble	Supplemental AI/ML evidence	Sharpe and robustness metrics	Reported but excluded from primary H1 tests to avoid overlap with H3
H1	SOLUSDT	RB_01 buy-and-hold	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; especially relevant in bull-trend folds
H1	SOLUSDT	RB_02 moving-average crossover	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; captures trend-following behaviour
H1	SOLUSDT	RB_03 RSI mean-reversion	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; tests mean-reversion behaviour
H1	SOLUSDT	RB_04 Bollinger Band mean-reversion	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; volatility-adjusted mean reversion
H1	SOLUSDT	RB_05 Donchian breakout	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; breakout and momentum behaviour
H1	SOLUSDT	RB_06 MACD trend	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; common trend and momentum benchmark
H1	SOLUSDT	RB_07 volatility breakout	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; high-turnover breakout family
H1	SOLUSDT	RB_08 EMA trend filter	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; trend-filtered rule-based family
H1	SOLUSDT	ML_01 logistic regression	AI/ML primary strategy	Fold-level Sharpe, median Sharpe, drawdown	Low-variance linear ML baseline
H1	SOLUSDT	ML_02 random forest	AI/ML primary strategy	Fold-level Sharpe, median Sharpe, drawdown	Tree-based non-linear classifier
H1	SOLUSDT	ML_03 gradient boosting	AI/ML primary strategy	Fold-level Sharpe, median Sharpe, drawdown	Boosted-tree classifier
H1	SOLUSDT	ML_04_svm	AI/ML primary strategy	Fold-level Sharpe, median Sharpe, drawdown	Support vector machine classifier
H1	SOLUSDT	ML_05_gru	AI/ML primary strategy	Fold-level Sharpe, median Sharpe, drawdown	Recurrent model used in the H1/H2 namespace
H1	SOLUSDT	ML_07_regime_lgbm	AI/ML primary strategy	Fold-level Sharpe, median Sharpe, drawdown	Regime-aware LightGBM included in H1 primary set
H1	SOLUSDT	ML_06_ensemble	Supplemental AI/ML evidence	Sharpe and robustness metrics	Reported but excluded from primary H1 tests to avoid overlap with H3
H1	XRPUSDT	RB_01 buy-and-hold	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; especially relevant in bull-trend folds
H1	XRPUSDT	RB_02 moving-average crossover	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; captures trend-following behaviour
H1	XRPUSDT	RB_03 RSI mean-	Rule-based	Sharpe, drawdown,	Primary comparator;

Hypothesis	Asset	Strategy or group	Role in test	Primary metrics	Evidence use
		reversion	benchmark	trade count	tests mean-reversion behaviour
H1	XRPUSD	RB_04 Bollinger Band mean-reversion	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; volatility-adjusted mean reversion
H1	XRPUSD	RB_05 Donchian breakout	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; breakout and momentum behaviour
H1	XRPUSD	RB_06 MACD trend	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; common trend and momentum benchmark
H1	XRPUSD	RB_07 volatility breakout	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; high-turnover breakout family
H1	XRPUSD	RB_08 EMA trend filter	Rule-based benchmark	Sharpe, drawdown, trade count	Primary comparator; trend-filtered rule-based family
H1	XRPUSD	ML_01 logistic regression	AI/ML primary strategy	Fold-level Sharpe, median Sharpe, drawdown	Low-variance linear ML baseline
H1	XRPUSD	ML_02 random forest	AI/ML primary strategy	Fold-level Sharpe, median Sharpe, drawdown	Tree-based non-linear classifier
H1	XRPUSD	ML_03 gradient boosting	AI/ML primary strategy	Fold-level Sharpe, median Sharpe, drawdown	Boosted-tree classifier
H1	XRPUSD	ML_04_svm	AI/ML primary strategy	Fold-level Sharpe, median Sharpe, drawdown	Support vector machine classifier
H1	XRPUSD	ML_05_gru	AI/ML primary strategy	Fold-level Sharpe, median Sharpe, drawdown	Recurrent model used in the H1/H2 namespace
H1	XRPUSD	ML_07_regime_lgbm	AI/ML primary strategy	Fold-level Sharpe, median Sharpe, drawdown	Regime-aware LightGBM included in H1 primary set
H1	XRPUSD	ML_06_ensemble	Supplemental AI/ML evidence	Sharpe and robustness metrics	Reported but excluded from primary H1 tests to avoid overlap with H3
H2	BTCUSD	RB_01 to RB_08 rule-based pool	Static benchmark component	Mean and median maximum drawdown	Part of the static comparator group
H2	BTCUSD	ML_01 logistic regression	Static ML benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed ML comparator with regime feature available
H2	BTCUSD	ML_02 random forest	Static ML benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed tree-based comparator
H2	BTCUSD	ML_03 gradient boosting	Static ML benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed boosted-tree comparator
H2	BTCUSD	ML_04_svm	Static ML benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed SVM comparator
H2	BTCUSD	ML_05_gru	Static ML benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed recurrent comparator
H2	BTCUSD	ML_06_ensemble	Static ensemble benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed ensemble comparator
H2	BTCUSD	ML_07_regime_lgbm	Regime-aware treatment	Maximum drawdown, cost stress, Monte Carlo	H2 treatment strategy with per-regime routing
H2	BTCUSD	Bear, ranging and bull segments	Regime diagnostic	Regime-segmented drawdown proxy	Supporting evidence only, not primary

Hypothesis	Asset	Strategy or group	Role in test	Primary metrics	Evidence use
					full-fold drawdown
H2	ETHUSDT	RB_01 to RB_08 rule-based pool	Static benchmark component	Mean and median maximum drawdown	Part of the static comparator group
H2	ETHUSDT	ML_01 logistic regression	Static ML benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed ML comparator with regime feature available
H2	ETHUSDT	ML_02 random forest	Static ML benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed tree-based comparator
H2	ETHUSDT	ML_03 gradient boosting	Static ML benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed boosted-tree comparator
H2	ETHUSDT	ML_04_svm	Static ML benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed SVM comparator
H2	ETHUSDT	ML_05_gru	Static ML benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed recurrent comparator
H2	ETHUSDT	ML_06_ensemble	Static ensemble benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed ensemble comparator
H2	ETHUSDT	ML_07_regime_lgbm	Regime-aware treatment	Maximum drawdown, cost stress, Monte Carlo	H2 treatment strategy with per-regime routing
H2	ETHUSDT	Bear, ranging and bull segments	Regime diagnostic	Regime-segmented drawdown proxy	Supporting evidence only, not primary full-fold drawdown
H2	SOLUSDT	RB_01 to RB_08 rule-based pool	Static benchmark component	Mean and median maximum drawdown	Part of the static comparator group
H2	SOLUSDT	ML_01 logistic regression	Static ML benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed ML comparator with regime feature available
H2	SOLUSDT	ML_02 random forest	Static ML benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed tree-based comparator
H2	SOLUSDT	ML_03 gradient boosting	Static ML benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed boosted-tree comparator
H2	SOLUSDT	ML_04_svm	Static ML benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed SVM comparator
H2	SOLUSDT	ML_05_gru	Static ML benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed recurrent comparator
H2	SOLUSDT	ML_06_ensemble	Static ensemble benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed ensemble comparator
H2	SOLUSDT	ML_07_regime_lgbm	Regime-aware treatment	Maximum drawdown, cost stress, Monte Carlo	H2 treatment strategy with per-regime routing
H2	SOLUSDT	Bear, ranging and bull segments	Regime diagnostic	Regime-segmented drawdown proxy	Supporting evidence only, not primary full-fold drawdown
H2	XRPUSDT	RB_01 to RB_08 rule-based pool	Static benchmark component	Mean and median maximum drawdown	Part of the static comparator group
H2	XRPUSDT	ML_01 logistic regression	Static ML benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed ML comparator with regime feature available
H2	XRPUSDT	ML_02 random forest	Static ML benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed tree-based comparator

Hypothesis	Asset	Strategy or group	Role in test	Primary metrics	Evidence use
H2	XRPUSD	ML_03 gradient boosting	Static ML benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed boosted-tree comparator
H2	XRPUSD	ML_04_svm	Static ML benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed SVM comparator
H2	XRPUSD	ML_05_gru	Static ML benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed recurrent comparator
H2	XRPUSD	ML_06_ensemble	Static ensemble benchmark	Maximum drawdown, Calmar, Sortino	Static, non-routed ensemble comparator
H2	XRPUSD	ML_07_regime_lgbm	Regime-aware treatment	Maximum drawdown, cost stress, Monte Carlo	H2 treatment strategy with per-regime routing
H2	XRPUSD	Bear, ranging and bull segments	Regime diagnostic	Regime-segmented drawdown proxy	Supporting evidence only, not primary full-fold drawdown
H3	BTCUSD	ML_01 logistic regression	Active single-model comparator	Median Sharpe, median max drawdown	Primary single-model group
H3	BTCUSD	ML_02 random forest	Active single-model comparator	Median Sharpe, median max drawdown	Primary single-model group; strongest single-model rank
H3	BTCUSD	ML_03 gradient boosting	Active single-model comparator	Median Sharpe, median max drawdown	Primary single-model group
H3	BTCUSD	ML_05_lightgbm	Active single-model comparator	Median Sharpe, median max drawdown	Primary single-model group
H3	BTCUSD	ML_07_svm	Active single-model comparator	Median Sharpe, median max drawdown	Primary single-model group; distinct from ML_07_regime_lgbm
H3	BTCUSD	ML_08_mlp	Active single-model comparator	Median Sharpe, median max drawdown	Primary single-model group
H3	BTCUSD	ML_06A majority vote	Primary ensemble	Median Sharpe, median max drawdown	Primary H3 ensemble group
H3	BTCUSD	ML_06B weighted vote	Primary ensemble	Median Sharpe, median max drawdown	Primary H3 ensemble group
H3	BTCUSD	ML_06C soft vote	Primary ensemble	Median Sharpe, median max drawdown	Primary H3 ensemble group
H3	BTCUSD	ML_06D stacking OOF	Primary ensemble	Median Sharpe, median max drawdown, rank	Primary H3 ensemble group; best median rank with sparse-trading caveat
H3	BTCUSD	ML_06E strategy-signal ensemble	Primary ensemble	Median Sharpe, median max drawdown	Primary H3 ensemble group; cost-sensitive
H3	BTCUSD	ML_06F rank-based ensemble	Primary ensemble	Median Sharpe, median max drawdown	Primary H3 ensemble group; weakest ranked strategy
H3	BTCUSD	ML_06G regime-aware ensemble	Secondary ensemble evidence	Median Sharpe, median max drawdown, rank	Secondary only because it fuses regime conditioning with ensembling
H3	BTCUSD	ML_04_xgboost	Failed H3 candidate	Trade count and failure status	Reported as failed; excluded from active single-model comparison
H3	BTCUSD	ML_09_lstm	Failed H3 candidate	Trade count and failure status	Reported as failed; excluded from active single-model comparison
H3	ETHUSD	ML_01 logistic	Active single-model	Median Sharpe,	Primary single-model

Hypothesis	Asset	Strategy or group	Role in test	Primary metrics	Evidence use
		regression	comparator	median max drawdown	group
H3	ETHUSDT	ML_02 random forest	Active single-model comparator	Median Sharpe, median max drawdown	Primary single-model group; strongest single-model rank
H3	ETHUSDT	ML_03 gradient boosting	Active single-model comparator	Median Sharpe, median max drawdown	Primary single-model group
H3	ETHUSDT	ML_05_lightgbm	Active single-model comparator	Median Sharpe, median max drawdown	Primary single-model group
H3	ETHUSDT	ML_07_svm	Active single-model comparator	Median Sharpe, median max drawdown	Primary single-model group; distinct from ML_07_regime_lgbm
H3	ETHUSDT	ML_08_mlp	Active single-model comparator	Median Sharpe, median max drawdown	Primary single-model group
H3	ETHUSDT	ML_06A majority vote	Primary ensemble	Median Sharpe, median max drawdown	Primary H3 ensemble group
H3	ETHUSDT	ML_06B weighted vote	Primary ensemble	Median Sharpe, median max drawdown	Primary H3 ensemble group
H3	ETHUSDT	ML_06C soft vote	Primary ensemble	Median Sharpe, median max drawdown	Primary H3 ensemble group
H3	ETHUSDT	ML_06D stacking OOF	Primary ensemble	Median Sharpe, median max drawdown, rank	Primary H3 ensemble group; best median rank with sparse-trading caveat
H3	ETHUSDT	ML_06E strategy-signal ensemble	Primary ensemble	Median Sharpe, median max drawdown	Primary H3 ensemble group; cost-sensitive
H3	ETHUSDT	ML_06F rank-based ensemble	Primary ensemble	Median Sharpe, median max drawdown	Primary H3 ensemble group; weakest ranked strategy
H3	ETHUSDT	ML_06G regime-aware ensemble	Secondary ensemble evidence	Median Sharpe, median max drawdown, rank	Secondary only because it fuses regime conditioning with ensembling
H3	ETHUSDT	ML_04_xgboost	Failed H3 candidate	Trade count and failure status	Reported as failed; excluded from active single-model comparison
H3	ETHUSDT	ML_09_lstm	Failed H3 candidate	Trade count and failure status	Reported as failed; excluded from active single-model comparison
H3	SOLUSDT	ML_01 logistic regression	Active single-model comparator	Median Sharpe, median max drawdown	Primary single-model group
H3	SOLUSDT	ML_02 random forest	Active single-model comparator	Median Sharpe, median max drawdown	Primary single-model group; strongest single-model rank
H3	SOLUSDT	ML_03 gradient boosting	Active single-model comparator	Median Sharpe, median max drawdown	Primary single-model group
H3	SOLUSDT	ML_05_lightgbm	Active single-model comparator	Median Sharpe, median max drawdown	Primary single-model group
H3	SOLUSDT	ML_07_svm	Active single-model comparator	Median Sharpe, median max drawdown	Primary single-model group; distinct from ML_07_regime_lgbm
H3	SOLUSDT	ML_08_mlp	Active single-model comparator	Median Sharpe, median max drawdown	Primary single-model group
H3	SOLUSDT	ML_06A majority vote	Primary ensemble	Median Sharpe, median max	Primary H3 ensemble group

Hypothesis	Asset	Strategy or group	Role in test	Primary metrics	Evidence use
				drawdown	
H3	SOLUSDT	ML_06B weighted vote	Primary ensemble	Median Sharpe, median max drawdown	Primary H3 ensemble group
H3	SOLUSDT	ML_06C soft vote	Primary ensemble	Median Sharpe, median max drawdown	Primary H3 ensemble group
H3	SOLUSDT	ML_06D stacking OOF	Primary ensemble	Median Sharpe, median max drawdown, rank	Primary H3 ensemble group; best median rank with sparse-trading caveat
H3	SOLUSDT	ML_06E strategy-signal ensemble	Primary ensemble	Median Sharpe, median max drawdown	Primary H3 ensemble group; cost-sensitive
H3	SOLUSDT	ML_06F rank-based ensemble	Primary ensemble	Median Sharpe, median max drawdown	Primary H3 ensemble group; weakest ranked strategy
H3	SOLUSDT	ML_06G regime-aware ensemble	Secondary ensemble evidence	Median Sharpe, median max drawdown, rank	Secondary only because it fuses regime conditioning with ensembling
H3	SOLUSDT	ML_04_xgboost	Failed H3 candidate	Trade count and failure status	Reported as failed; excluded from active single-model comparison
H3	SOLUSDT	ML_09_lstm	Failed H3 candidate	Trade count and failure status	Reported as failed; excluded from active single-model comparison
H3	XRPUSDT	ML_01 logistic regression	Active single-model comparator	Median Sharpe, median max drawdown	Primary single-model group
H3	XRPUSDT	ML_02 random forest	Active single-model comparator	Median Sharpe, median max drawdown	Primary single-model group; strongest single-model rank
H3	XRPUSDT	ML_03 gradient boosting	Active single-model comparator	Median Sharpe, median max drawdown	Primary single-model group
H3	XRPUSDT	ML_05_lightgbm	Active single-model comparator	Median Sharpe, median max drawdown	Primary single-model group
H3	XRPUSDT	ML_07_svm	Active single-model comparator	Median Sharpe, median max drawdown	Primary single-model group; distinct from ML_07_regime_lgbm
H3	XRPUSDT	ML_08_mlp	Active single-model comparator	Median Sharpe, median max drawdown	Primary single-model group
H3	XRPUSDT	ML_06A majority vote	Primary ensemble	Median Sharpe, median max drawdown	Primary H3 ensemble group
H3	XRPUSDT	ML_06B weighted vote	Primary ensemble	Median Sharpe, median max drawdown	Primary H3 ensemble group
H3	XRPUSDT	ML_06C soft vote	Primary ensemble	Median Sharpe, median max drawdown	Primary H3 ensemble group
H3	XRPUSDT	ML_06D stacking OOF	Primary ensemble	Median Sharpe, median max drawdown, rank	Primary H3 ensemble group; best median rank with sparse-trading caveat
H3	XRPUSDT	ML_06E strategy-signal ensemble	Primary ensemble	Median Sharpe, median max drawdown	Primary H3 ensemble group; cost-sensitive
H3	XRPUSDT	ML_06F rank-based ensemble	Primary ensemble	Median Sharpe, median max drawdown	Primary H3 ensemble group; weakest ranked strategy
H3	XRPUSDT	ML_06G regime-aware ensemble	Secondary ensemble evidence	Median Sharpe, median max	Secondary only because it fuses

Hypothesis	Asset	Strategy or group	Role in test	Primary metrics	Evidence use
				drawdown, rank	regime conditioning with ensembling
H3	XRPUSD	ML_04_xgboost	Failed H3 candidate	Trade count and failure status	Reported as failed; excluded from active single-model comparison
H3	XRPUSD	ML_09_lstm	Failed H3 candidate	Trade count and failure status	Reported as failed; excluded from active single-model comparison

The table shows why the thesis avoids treating the strategy IDs as a single flat numbering system. In H1 and H2, ML_04_svm and ML_05_gru are active implementations; in H3, ML_04_xgboost and ML_09_lstm are failed candidates. Likewise, ML_07_regime_lgbm is the H2 treatment strategy, while ML_07_svm is an H3 single-model comparator. The written findings therefore use full names when a numeric identifier could mislead.

The mapping also shows how the same asset universe supports different empirical claims. H1 asks whether AI/ML improves risk-adjusted return against rule-based strategies. H2 asks whether regime routing reduces drawdown against a demanding static benchmark. H3 asks whether ensembles improve median Sharpe stability and whether they also reduce drawdown. The evidence is comparable because the assets, walk-forward folds, execution assumptions and audit rules are common, even though the hypotheses differ.

For submission purposes, Appendix M should be read as a reproducibility aid. It does not add a new empirical claim beyond Chapters IV, V and VI. Instead, it documents the evidence structure in a way that makes the thesis easier to audit and easier to extend. A future researcher can replace one strategy family, add one asset, or repair a failed model and immediately see where the revised result would enter the hypothesis framework.